Natural Language Processing

David C. Wyld,
Dhinaharan Nagamalai (Eds)

Computer Science & Information Technology

- 8th International Conference on Natural Language Processing (NATP 2022), January 22 ~ 23, 2022, Zurich, Switzerland
- 8th International Conference on Advances in Computer Science and Information Technology (ACSTY 2022)
- 3rd International Conference on Cloud Computing and IOT (CCCIOT 2022)
- 3rd International Conference on Machine Learning and Soft Computing (MLSC 2022)
- 8th International Conference on Information Technology Convergence and Services (ITCSS 2022)

Published By

AIRCC Publishing Corporation

Volume Editors

David C. Wyld,
Southeastern Louisiana University, USA
E-mail: David.Wyld@selu.edu

Dhinaharan Nagamalai (Eds),
Wireilla Net Solutions, Australia
E-mail: dhinthia@yahoo.com

ISSN: 2231 - 5403
ISBN: 978-1-925953-59-6
DOI: 10.5121/csit.2022.120101- 10.5121/csit.2022.120110

This work is subject to copyright. All rights are reserved, whether whole or part of the material is concerned, specifically the rights of translation, reprinting, re-use of illustrations, recitation, broadcasting, reproduction on microfilms or in any other way, and storage in data banks. Duplication of this publication or parts thereof is permitted only under the provisions of the International Copyright Law and permission for use must always be obtained from Academy & Industry Research Collaboration Center. Violations are liable to prosecution under the International Copyright Law.

Typesetting: Camera-ready by author, data conversion by NnN Net Solutions Private Ltd., Chennai, India

Preface

8th International Conference on Natural Language Processing (NATP 2022), January 22 ~ 23, 2022, Zurich, Switzerland, 8th International Conference on Advances in Computer Science and Information Technology (ACSTY 2022), 3rd International Conference on Cloud Computing and IOT (CCCIOT 2022), 3rd International Conference on Machine Learning and Soft Computing (MLSC 2022) and 8th International Conference on Information Technology Convergence and Services (ITCSS 2022) was collocated with 8th International Conference on Natural Language Processing (NATP 2022). The conferences attracted many local and international delegates, presenting a balanced mixture of intellect from the East and from the West.

The goal of this conference series is to bring together researchers and practitioners from academia and industry to focus on understanding computer science and information technology and to establish new collaborations in these areas. Authors are invited to contribute to the conference by submitting articles that illustrate research results, projects, survey work and industrial experiences describing significant advances in all areas of computer science and information technology.

The NATP 2022, ACSTY 2022, CCCIOT 2022, MLSC 2022 and ITCSS 2022 Committees rigorously invited submissions for many months from researchers, scientists, engineers, students and practitioners related to the relevant themes and tracks of the workshop. This effort guaranteed submissions from an unparalleled number of internationally recognized top-level researchers. All the submissions underwent a strenuous peer review process which comprised expert reviewers. These reviewers were selected from a talented pool of Technical Committee members and external reviewers on the basis of their expertise. The papers were then reviewed based on their contributions, technical content, originality and clarity. The entire process, which includes the submission, review and acceptance processes, was done electronically.

In closing, NATP 2022, ACSTY 2022, CCCIOT 2022, MLSC 2022 and ITCSS 2022 brought together researchers, scientists, engineers, students and practitioners to exchange and share their experiences, new ideas and research results in all aspects of the main workshop themes and tracks, and to discuss the practical challenges encountered and the solutions adopted. The book is organized as a collection of papers from the NATP 2022, ACSTY 2022, CCCIOT 2022, MLSC 2022 and ITCSS 2022.

We would like to thank the General and Program Chairs, organization staff, the members of the Technical Program Committees and external reviewers for their excellent and tireless work. We sincerely wish that all attendees benefited scientifically from the conference and wish them every success in their research. It is the humble wish of the conference organizers that the professional dialogue among the researchers, scientists, engineers, students and educators continues beyond the event and that the friendships and collaborations forged will linger and prosper for many years to come.

David C. Wyld,
Dhinaharan Nagamalai (Eds)

General Chair

Organization

David C. Wyld, Southeastern Louisiana University, USA
Dhinaharan Nagamalai (Eds) Wireilla Net Solutions, Australia

Program Committee Members

Abdalhossein Rezai, University of Science and Culture, Iran
Abdelhadi Assir, Hassan 1st University, Morocco
Abderrahim Siam, University of Khenchela, Algeria
Abtoy Anouar, Abdelmalek Essaadi University, Morocco
Afaq Ahmad, Sultan Qaboos University,Oman
Ahmed Farouk AbdelGawad, Zagazig University, Egypt
Ajit Singh, Patna University, India
Ali A. Amer, Taiz University, Yemen
Amal Azeroual, Mohammed V University, Morocco
Amin Bazzazi, Islamic Azad University, Iran
Amizah Malip, University of Malaya, Malaysia
Anita Yadav, Harcourt Butler Technical University, India
Aridj Mohamed, Hassiba Benbouali University Chlef, Algeria
Aridj Mohamed, University Chlef algeria, Algeria
Ashutosh Kumar Dubey, Chitkara University, India
Atanu Nag, IFTM University, India
Atheer Yousif Oudah, Thi-Qar University, Iraq
Azeddine Chikh, University of Tlemcen, Algeria
Behrouz Gordan, Islamic Azad University, Iran
Beshair Alsiddiq, Riyad Bank, Saudi Arabia
Boukari nassim, Skikda University, Algeria
Brahim Lejdel, University of El-Oued,Algeria
Christian Mancas, Ovidius University, Romania
Dário Ferreira, University of Beira Interior, Portugal
Der-Chyuan Lou, Chang Gung University, Taiwan
Dorra Driss, University of Sfax, Tunisia
Ekbal Rashid, RTC Institute of Technology, India
El Murabet Amina, Abdelmalek Essaadi University, Morocco
Emad Awada, Applied Science University, Jordan
Ez-zahout abderrahmane, Mohammed V University, Morocco
F. Abbasi, Islamic Azad University, Iran
Fatemeh Deregeh, Shahid Bahonar University of Kerman, Iran
Felix J. Garcia Clemente, University of Murcia, Spain
Fezile Ozdamli, Near East University, Cyprus
Francesco Zirilli, Sapienza Universita Roma, Italy
Gajendra Sharma, Kathmandu University, Nepal
Gitesh K. Raikundalia, Victoria University, Australia
Goswami, Indian Institute of Technology, Kharagpur, India
Grigorios N. Beligiannis, University of Patras,Greece
Grzegorz Sierpiński, Silesian University of Technology, Poland
Gueddah Hicham, Mohammed V University, Morocco
Guezouli Larbi, University of batna 2, Algeria
Hamzeh Khalili, CTTC, Spain

Hatem Yazbek,	Broadcom Inc., Israel
Hayder Dibs,	AL-Qasim Green University, Iraq
Hongzhi,	Harbin Institute of Technology, China
Hossein Jadidoleslamy,	The University of Zabol, Iran
Ilham Huseyinov,	Istanbul Aydin University, Turkey
Irina Perfilieva,	University of Ostrava, Czech Republic
Jabbar,	Vardhaman College of Engg, India
Jafar Mansouri,	Ferdowsi University of Mashhad, Iran
Jagadeesh HS,	APS College of Engineering (VTU), India
Jalel Akaichi,	University of Tunis, Tunisia
Jawad K. Ali,	University of Technology, Iraq
Jin-Whan Kim,	Youngsan University, South Korea
Jose Alfredo F. Costa,	Federal University, Brazil
Julie M David,	MES College Marampally, India
Jun Zhang,	South China University of Technology, China
K. Senthamarai Kannan,	MS University, India
Kais Haddar,	University of Sfax, Tunisia
Karan Veer,	NIT Jalandhar, India
Karim El Moutaouakil,	FPT/USMBA, Morocco
Karim Mansour,	Salah Boubenider University, Algeria
Kazuyuki Matsumoto,	Tokushima University, Japan
Khaled Osama Elzoghaly,	Alexandria University, Egypt
Kirtikumar Patel,	Chemic Engineers, USA
Labed Said,	University of Constantine, Algeria
Ljubomir Lazic,	Belgrade UNION University, Serbia
Luisa Maria Arvide Cambra,	University of Almeria, Spain
M V Ramana Murthy,	Osmania University, India
M.A. Jabbar,	Vardhaman College of Engineering, India
Mabroukah Amarif,	Sebha University, Libya
Mahdi Sabri,	Islamic Azad University Urmia Branch, Iran
Malleswara Talla,	Concordia University, Canada
Marcin Paprzycki,	Polish Academy of Sciences, Poland
Marco Javier Suarez Baron,	University in Tunja, Colombia
Maryam hajakbari,	Islamic Azad University, Iran
Masoud Asghari,	Urmia University, Iran
Md. Sadique Shaikh,	AIMSR, Maharashtra, India
Mehmet Ali Erturk,	Istanbul University, Turkey
Michail Kalogiannakis,	University of Crete, Greece
Mirsaeid Hosseini Shirvani,	Islamic Azad University, Iran
Mohamedmaher Benismail,	King saud University, Saudi Arabia
Mohammed A.M. Salem,	German University, Egypt
Mounir Zrigui,	University of Monastir, Tunisia
Mueen Uddin,	Universiti Brunei Darussalam, Brunei
Muhammad Sarfraz,	Kuwait University, Kuwait
Mu-Song Chen,	Da-Yeh University, Taiwan
Mussab Alaa,	University of Malaya, Malaysia
N P G Bhavani,	Saveetha School of Engineering, India
Nadia Abd-Alsabour,	Cairo University, Egypt
Nasim Sadat,	University of Minho, Portugal
Nicolae Tudoroiu,	John Abbott College, Canada
Nikola Ivkovic,	University of Zagreb, Croatia
Noura Taleb,	Badji Mokhtar University, Algeria

Omar Boussaid,	University of Lyon, France
Otilia Manta,	Romanian-American University, Romania
Patrick Fiati,	Cape Coast Technical University, Ghana
Pavel Loskot,	ZJU-UIUC Institute, China
Prudhvi Parne,	FVP, Software Development Manager, USA
Przemyslaw Falkowski-Gilsk,	Gdansk University of Technology, Poland
Rabhat Mahanti,	University of New Brunswick, Canada
Rahul Kosarwal,	OAARs CORP, United Kingdom
Rahul Saha,	Lovely Professional University, India
Ramadan Elaiess,	University of Benghazi, Libya
Ramgopal Kashyap,	Amity University Chhattisgarh, India
Rami Raba,	Al Azhar University, Palestine
Reyhane Attarian,	Shiraz University, Iran
Richa Purohit,	DY Patil International University, India
Ruiying Geng,	Alibaba Group, China
Saad Aljanabi,	Alhikma College University, Iraq
Saad Al-Janabi,	Al-Hikma College University, Iraq
Saeed Rouhani,	University of Tehran, Iran
Sahil Verma,	IAENG, India
Saïd Nouh,	Hassan II University of Casablanca, Morocco
Saikumar Tara,	CMR Technical Campus Hyderabad, India
Sajadin Sembiring,	Universitas Sumatera Utara, Indonesia
Sandro Sessarego,	University of Texas at Austin, USA
Sebastian Floerecke,	University of Passau, Germany
Selçuk Helel,	Akdeniz University, Turkey
Seppo Sirkemaa,	University of Turku, Finland
Shahram Babaie,	Islamic Azad Unicersity, Iran
Shariq Aziz Butt,	University of Lahore, Pakistan
Shashikant Patil,	SVKMs NMIMS Mumbai, India
Shaveta Malik,	University of Mumbai, India
Sikandar Ali,	China University of Petroleum, China
Siuly Siuly,	Victoria University, Australia
Sourav Banerjee,	Kalyani Government Engineering College, India
Suhad Faisal Behadili,	University of Baghdad, Iraq
Sultan Ahmad,	Prince Sattam Bin Abdulaziz University, Saudi Arabia
Tanzila Saba,	Prince Sultan University, Saudi Arabia
Terumasa Aoki,	Tohoku University, Japan
Tomasz Wojciechowski,	Poznań University of Life Sciences, Poland
Umesh Kumar Singh,	Istitute of Computer Science, India
V. Dinesh Reddy,	SRM University, India
Venkata Duvvuri,	Purdue University, USA
Vilem Novak,	University of Ostrava, Czech Republic
Vinita Verma,	University of Delhi, India
Virupakshappa,	Sharnbasva University Kalaburagi, India
Wei Cai,	American multinational corporation, USA
Wei-Chiang Hong,	Jiangsu Normal University, China
Wu Hao,	Beijing Institute of Technology, China
Xiao-Zhi Gao,	University of Eastern Finland, Finland
Yew Kee WONG,	HuangHuai University, China
Yuan Tian,	King Saud University, Saudi Arabia
Yu-Lin (Eugene) Song,	Asia University, Taiwan
Zoran Bojkovic,	University of Belgrade, Serbia

Technically Sponsored by

Computer Science & Information Technology Community (CSITC)

Artificial Intelligence Community (AIC)

Soft Computing Community (SCC)

Digital Signal & Image Processing Community (DSIPC)

8th International Conference on Natural Language Processing (NATP 2022)

8th International Conference on Advances in Computer Science and Information Technology (ACSTY 2022)

3rd International Conference on Cloud Computing and IOT (CCCIOT 2022)

3rd International Conference on Machine Learning and Soft Computing (MLSC 2022)

8th International Conference on Information Technology Convergence and Services (ITCSS 2022)

IDENTIFICATION OF KEY NODES IN EQUIPMENT SYSTEM NETWORK BASED ON FUNCTION CHAIN

Cheng Huang, Yong Gang Li and Ying Wang

School of Communication and Information Engineering, Chongqing University of Posts and Telecommunications, Chong Qing, China

ABSTRACT

With the rapid development of modern military technology, the combat mode has been upgraded from traditional platform combat to system-level confrontation. In traditional combat network, node function is single and which is no proper assignment of tasks. The equipment system network studied in this paper contains many different functional nodes, which constitute a huge heterogeneous complex network. Most of the key node identification methods are analyzed from the network topology structure, such as degree, betweenness, K-shell, PageRank, etc. However, with the change of network topology, the identification effect of these methods will be biased. In this paper, we construct a nodal attack sequence, Consider the change of the number of effective OODA chains in the equipment system network after the nodes in the sequence are attacked. And combined with the improved Gray Wolf optimization algorithm, this paper proposes a key node evaluation model of equipment system network based on function chain —— IABFI. Experimental results show that the proposed method is more effective, accurate, and applicable to different network topologies than other key node identification methods.

KEYWORDS

Equipment system network, node sequence attack, effective OODA chain, improved Grey Wolf optimization algorithm.

1. INTRODUCTION

Modern military technology is evolving rapidly, and the mode of operation has shifted to system-level combat versus a single weapon platform [1, 2]. It is not a single soldier or a stand-alone operation as we have in mind. It emphasizes the communication between the parts and the division of labor. The equipment system network is a complex heterogeneous network, which is a higher-level whole body composed of various weapons and equipment systems that are connected and interact with each other in terms of function. The network contains a variety of nodes with different functions [3]. In this complex network of equipment system, the importance of different nodes is very different. Once a very important node is attacked and fails, the whole network of equipment system will be greatly affected [4]. Therefore, in this era of rapid development of military technology. Accurately and quickly find out the key nodes in the network, protect our important nodes, hit the enemy's key nodes, So asto win the battle, has epoch-making significance.

Project supported by the National Defense PreResearch Quick Support Foundation of China (no.80911010 302)

David C. Wyld et al. (Eds): NATP, ACSTY, CCCIOT, MLSC, ITCSS - 2022
pp. 01-16, 2022. CS & IT - CSCP 2022 DOI: 10.5121/csit.2022.120101

At present, key node identification in complex networks is mostly based on isomorphic network model, that is, in current key node identification, only the same type of nodes and edges exist in the studied network. These studies do not take into account the heterogeneity of the equipment system network, which can't adapt to the large-scale military system confrontation, so the algorithm proposed by them can't be well applied to the combat research of heterogeneous network. At present, the identification of key nodes mainly considers the network topology structure, such as degree centrality, betweenness centrality, K-shell, PageRank and so on [5]. Or consider the combination of weights, comprehensively consider the weights of evaluation indicators from both subjective and objective dimensions, and then multiply the indicators by the corresponding weights, sum up, to get the sequence of important nodes [6]. These traditional indicators and methods can accurately identify key nodes in some specific scenarios, but the structure and scale of the network will not remain unchanged [7, 8]. Nowadays, with the increasingly large structure and scale of equipment system network, these traditional indexes and methods are obviously unable to meet the needs of large-scale military operations.

In this paper, considering the complex combat tasks of the equipment system, according to the requirements of combat tasks, the network nodes of the equipment system are divided into sensor node S, command node D and strike node I [9]. S nodes, D nodes and I nodes are combined to form a complete chain of reconnaissance, decision making and strike to complete combat missions. According to the above, this paper proposes a key node identification and evaluation model of equipment system network based on function chain——IABFI. By combining the function chain with the improved grey Wolf optimization algorithm, the overall identification method can achieve a more accurate identification effect on the equipment system network of different sizes and structure.

2. ROBUSTNESS MEASUREMENT OF EQUIPMENT SYSTEM NETWORK BASED ON FUNCTION CHAIN

Equipment system network robustness refers to the remaining operational capability of the entire equipment system network after some nodes or edges are attacked.Due to the large difference between the topological structure of equipment system network and the traditional complex network, the traditional methods to measure the robustness of homogeneous network, such as the maximum connected component and network efficiency, are not suitable for heterogeneous equipment system network. In this paper, effective OODA chain is introduced to measure the robustness of equipment system network. The operational effectiveness of modern military operations is no longer a solitary struggle. Considering the performance index of a single machine, it is more important to consider the information interaction between each equipment entity. The communication of control and command information among individual machines has an important influence on the overall confrontation of the system. In the equipment system network, the equipment entity is abstracted as node, and the information transfer is abstracted as edge. Before introducing OODA chains, let's first introduce OODA rings, as shown in figure 1.

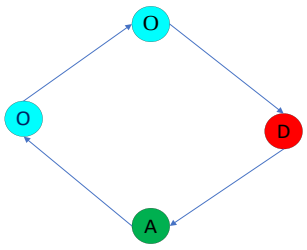

Figure 1. OODA ring model

A complete military operation has four steps: Observe (O) $\rightarrow$ Orient (O) $\rightarrow$Decide (D) $\rightarrow$Act (A). In the network of equipment systems, The OODA ring can be abstracted as an SDI chain model.

As shown in figure 2, in the function chain, combat information starts from node S, passes through node D, and finally reaches node I, which is directional.

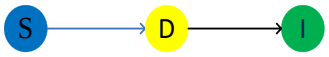

Figure 2. SDI chain model

However, there may be many $S \rightarrow S$ and $D \rightarrow D$ in the SDI chain, which leads to the concept of a generalized SDI chain, as shown in figure 3.

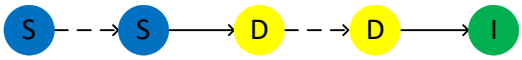

Figure 3. Generalized SDI chain model

For the equipment system network, we put forward the concept of effective OODA chain. The effective OODA chain can be understood as the same function chain, even if the connection mode is different, as long as the node set involved in the same function chain is consistent. as shown in figure 4, in the equipment architecture network model, $S_4 \rightarrow S_5 \rightarrow D_4 \rightarrow I_5$ and $S_5 \rightarrow S_4 \rightarrow D_4 \rightarrow I_5$ can be viewed as the same chain of functions. A valid OODA chain is calculated as shown in formula (1).

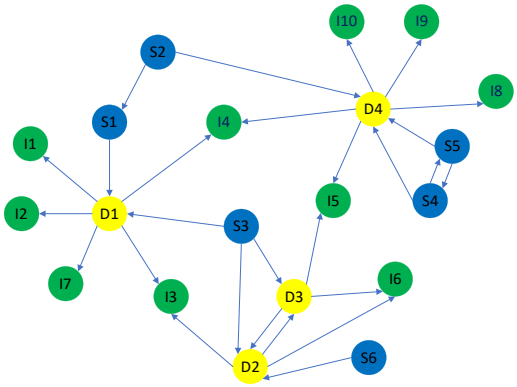

Figure 4. Network model of equipment system based on function chain

$$V_{\text{num}} = \prod(\sum_{i}^{n} D^{i}_{in\_degree\_S} \times D^{i}_{adjoin\_in\_S})(\sum_{i}^{n} D^{i}_{out\_degree\_I} \times D^{i}_{adjoin\_out\_I}) \quad (1)$$

In formula (1), V_{num} represents the number of valid OODA chains, $D^{i}_{in\_degree\_S}$ is the number of S nodes connected to the $i-th$ D node, $D^{i}_{adjoin\_in\_S}$ is the number of adjacent S nodes of adjacent S nodes connected to the $i-th$ D node.

Network performance is represented by the change in the number of effective OODA chains after i nodes are removed from the equipment system network. Obviously, a simple $S \rightarrow D$ chain, such a link does not attack node i, there is no way to inflict a blow on the enemy. Or $D \rightarrow I$ chain, no reconnaissance node S, clueless blind attack, such a combat link is not good. Therefore, in an equipment system network, the greater the number of OODA chains, the higher the network performance and the stronger the network robustness. Assuming that network G has n nodes with labels ranging from 1 to N, a node sequence $k = \{k_1, k_2, k_3, k_4, \cdots, k_n\}$ can be specified. Attack and remove nodes in sequence according to the node sequence. With the removal of nodes, the cumulative robustness of the equipment system network under sequence K is defined as:

$$CR(G,K) = \sum_{i=1}^{n} V_{num}(i) \qquad (2)$$

$V_{num}(i)$ represents the number of valid OODA chains remaining in the network after the i-th node in the equipment system network G is attacked. With different node sequences K, the cumulative robustness CR of the network is different. The smaller CR is, the more destructive the removal of nodes according to the corresponding node sequence will be to the network of the equipment system, and the sequence of nodes arranged in the node sequence will be more accurate [10].

3. IDENTIFICATION OF KEY NODES BASED ON IMPROVED GRAY WOLF OPTIMIZATION ALGORITHM

3.1. Basic Gray Wolf Optimization Algorithm

For the equipment system network G, we rank nodes according to their importance degree from high to low, and the network robustness obtained in this order will also reach the minimum. If we find a corresponding equipment system network G, and the sequence robustness CR reaches the minimum, then this sequence is the best sequence we require. Based on the above ideas, we can get out of the limitation of analyzing node importance from network topology. Whether it is degree centrality, betweenness centrality or K-shell sorting algorithm, the accuracy of these sorting algorithms varies greatly in different network topologies. The attack sequence of nodes is constructed based on the improved Gray Wolf optimization algorithm, and the identification of key nodes is transformed into a function optimization problem, which has high accuracy and is suitable for various network structures. The goal of this paper is to construct a node sequence K that minimizes the cumulative robustness CR(G,K) of the equipment system network after nodes are attacked and invalid. The construction function is shown in formula (3):

$$\begin{cases} \min & \text{CR(G,K)} \\ s.t. & \text{K} \in Set\_k \end{cases} \qquad (3)$$

The above Set_k is a set, which contains the attack sequences of all nodes against the target equipment system network G.

The node size of the equipment system network is very large, and it is normal to have thousands of nodes. And when the studied equipment system network G contains n nodes, the number of elements in Set_k is $n!$. Faced with such a huge solution space scale problem, we consider using intelligent search algorithm. In view of the excellent convergence stability and strong global search ability of the Gray wolf optimization algorithm, this paper adopts the Gray wolf optimization algorithm to realize the sequence optimization and identify the optimized node sequence[11]. Gray wolf optimization algorithm is a heuristic intelligent algorithm proposed by Mirjalili et al in 2014, it is a relatively new optimization technique that simulates the leadership hierarchy of pack hunting in a grey wolf pack, which takes on a golden tower shape[12], as shown in figure 5.

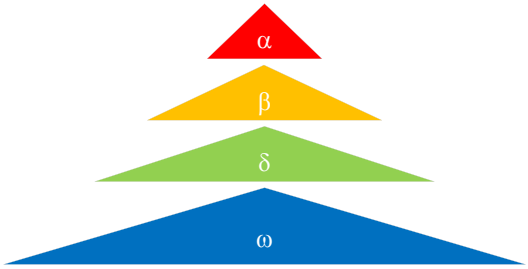

Figure 5. Leadership hierarchy of Gray Wolf pack [13]

As we all know, the Gray wolf is a social carnivore. The Gray wolf family is divided into four classes : α、 β、 δ、 ω . As can be seen from fig.5, wolf α is the top wolf in the Gray wolf world, serving as the leader, followed by the deputy leader, wolf β . wolf ω is the bottom wolf. wolf δ obeys the rule of wolf α and wolf β , but it can manage wolf ω [14].

Search process: when Gray Wolf is searching for prey, a very important judgment standard that prompts Gray Wolf to take hunt is according to the distance between oneself and prey. Let's say we're in the t-th iteration of the search, $X(t)$ is the position of the Gray Wolf, $X_p(t)$ is the position of the prey, then the distance between the Gray Wolf and the prey is shown in formula (4):

$$\begin{cases} D = \left| C \bullet X_p(t) - X(t) \right| \\ C = 2r_2 \\ r_2 = rand(0,1) \end{cases} \quad (4)$$

Encircle process: In the process of encircle prey, Gray wolf establishes the relationship model between Gray wolf and prey according to the distance between them, so as to realize the process of encircle prey. The relation model is shown in formula (5):

$$\begin{cases} X_i^d(t+1) = X_p^d(t) - A_i^d \bullet D_i^d \\ D_i^d = \left| C_i^d \bullet X_p^d(t) - X_i^d(t) \right| \\ A_i^d = 2ar_1 - a \\ C_i^d = 2ar_2 \\ a = 2 - t/t_{\max} \\ r_1, r_2 = rand(0,1) \end{cases} \quad (5)$$

In Formula 5, $A_i^d \bullet D_i^d$ represents the envelop step size in the process of envelop prey, and the maximum number of iterations is represented by $t_{\max}$ in the formula, t is the current number of iterations. For $a = 2 - t / t_{\max}$, we can see that the parameter, a, decreases linearly from 2 to 0. The random initialization of A_i^d and C_i^d ensures that the Gray Wolf will not easily fall into the local optimal position in the process of the search, and can easily reach the global optimal position.

Position update (attack): We can accurately and quickly judge the position of target prey through the position information updated by wolf α, wolf β and wolf δ.

$$\begin{cases} X_1 = X_\alpha - A_1 \bullet D_\alpha \\ X_2 = X_\beta - A_2 \bullet D_\beta \\ X_3 = X_\delta - A_3 \bullet D_\delta \\ X = (X_1 + X_2 + X_3)/3 \end{cases} \quad (6)$$

In formula 6, the positions of wolf α, wolf β and wolf δ are represented by X_1, X_2 and X_3 respectively. A_1, A_2 and A_3 are three random numbers. $A_1 \bullet D_\alpha$、 $A_2 \bullet D_\beta$ and $A_3 \bullet D_\delta$ represent the encircling steps of wolf α, wolf β and wolf δ, X is the final position, where the wolf attacked its prey.

The Gray wolf has strong global ability through the variation of speed, the change of search radius at any time, the update of position and other strategies, and it is easier for the Gray wolf to get the optimal solution and suboptimal solution in the global scope [15].

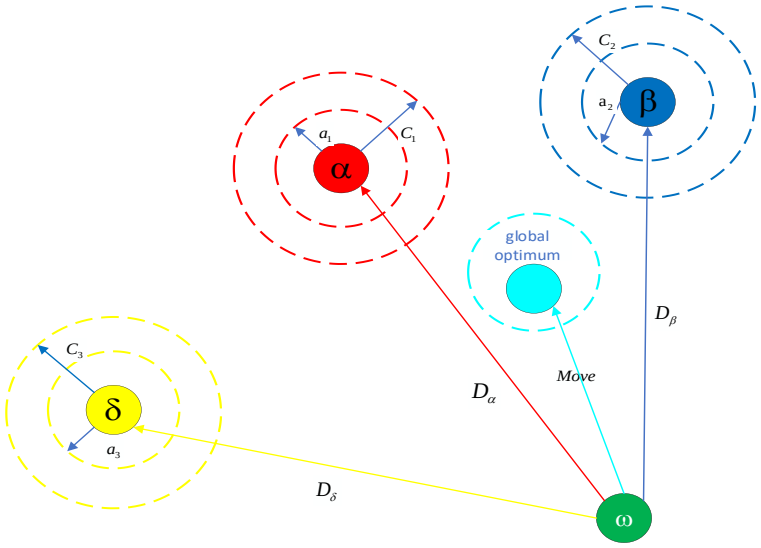

Figure 6. Position update of Gray Wolf in Gray Wolf optimization algorithm [15]

3.2. Improvement of Gray Wolf Optimization Algorithm

The grey wolf optimization algorithm has a fast convergence rate in the initial stage, and the wolves' position changes greatly, so it has strong global search performance. However, as the number of iterations increases in the later period, and changes in position become less volatile, which will easily lead to the algorithm falling into local optimal. To make up for this shortcoming, let's make some improvements to the Gray Wolf algorithm.

(1) Construction of the objective function: The robustness of the equipment system network based on function chain is studied in this paper. In the equipment system network, nodes are attacked in sequence according to the sequence of nodes, and the cumulative robustness of the equipment system network is calculated as the nodes are removed.

To solve the problem of the cumulative robustness of the equipment system network under the node sequence K, we introduce the Gray Wolf algorithm: The population size of Gray Wolf is N, and the number of nodes in the equipment system is D. In the D-dimensional node search space, the position of the i-thGray wolf is X (that is, the node sequence of the equipment system network), As defined in formula (7):

$$X_i = (X_i^1, X_i^2, X_i^3, \cdots X_i^{D-1}, X_i^D) \tag{7}$$

According to the Gray wolf algorithm, when N Gray wolves search for prey in the D-dimensional space, space domain P can be defined as a matrix, as follows:

$$P=\begin{bmatrix} X_1^1 & X_1^2 & \cdots & X_1^j & \cdots & X_1^{D-1} & X_1^D \\ X_2^1 & X_2^2 & \cdots & X_2^j & \cdots & X_2^{D-1} & X_2^D \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ X_i^1 & X_i^2 & \cdots & X_i^j & \cdots & X_i^{D-1} & X_i^D \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ X_N^1 & X_N^2 & \cdots & X_N^j & \cdots & X_N^{D-1} & X_N^D \end{bmatrix} \quad (8)$$

In formula (8), $X_i^j(i \le N, j \le D)$ represents the position of the i-th Gray Wolf in j-th dimension in the spatial domain P. Each line represents a sequence of positions of a Gray Wolf in the search space, that is, a sequence of nodes in the equipment system network. What is required is to find a sequence among these node sequences that minimizes the cumulative robust CR(G,K) of the equipment system network, The value of the objective function f is calculated by $f = \min \ \mathrm{CR(G,K)}$.

To sum up, when the number of nodes to be solved is D, what we need to take is to find an optimal position K of Gray Wolf population, so that the objective function value f is the minimum.

$$\begin{cases} f = \min \quad \mathrm{CR(G,K)} \\ s.t. \quad \mathrm{K} \in Set\_k \end{cases} \quad (9)$$

G is the equipment system network topology. K is a Gray Wolf position.

(2) Construction of initial solution: in the process of Gray wolf algorithm optimization, a good initial solution is particularly important for iterative optimization, which helps to reduce the algorithm complexity and optimization time. For finding the key nodes in the network of the equipment system, the relative quality of the initial solution generated in a random way is very low. The damage of attacking core nodes is much greater than that of attacking marginal nodes, which can result in the sharp reduction of effective OODA chain in the equipment system network.

According to the above idea, the weak centrality $VR(i)$ of node $i(1 \le i \le n)$ is defined as:

$$VR(i) = V_{num} - V_{num}(i) \quad (10)$$

V_{num} represents the number of effective OODA chains in the initial equipment system network. $V_{num}(i)$ indicates the number of remaining valid OODA chains in the network after the i-th node in the installation system network G is attacked. According to formula 10, the larger $VR(i)$ is, the more important node i is. According to the calculation, we take the node i with the highest $VR(i)$ value as the first node in the sequence, then the values are sorted from large to small until a complete sequence of node attacks is constructed.

(3) 2-opt optimization algorithm:2-opt optimization is also known as pairwise swapping, which is a local search algorithm[16]. When we update the position of the Gray wolf, we use this method

of exchanging the two datas, which effectively avoids the blind disorder when the Gray wolf changes the position.

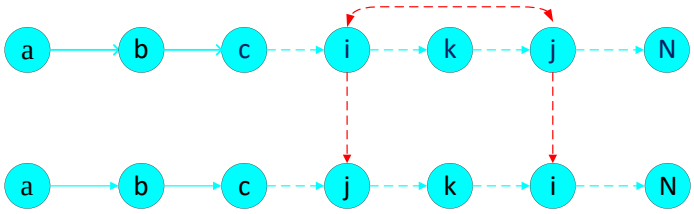

Figure 7. 2—opt schematic diagram[14]

As shown in figure 7, assuming that the second node sequence makes the value of the objective function smaller, the system network node i and j are exchanged, and the position number of the Gray wolf is updated.

The node solution vector of the equipment system network can be expressed as:

$$X_i = (X_i^1, X_i^2, \cdots, X_i^{D-1}, X_i^D) \quad (11)$$

In formula 11, $i(i = 1,2,3 \cdots N)$ is the i-th Gray Wolf in the Gray Wolf population, D is the serial number of the network node of the equipment system traversed by the Gray Wolf. According to the size of the distance between the solution vectors, we choose the two nodes whose distance between the solution vectors is shorter and then exchange the two nodes.

The core of 2-0PT optimization algorithm: The cumulative robustness of the equipment system network after switching nodes is calculated. If the value becomes smaller, it indicates that it is indeed optimized, and we retain the current solution vector as the optimal solution of this Gray Wolf; otherwise, the solution vector remains unchanged. Continue to make the above judgment for the next line of Gray wolves until all the solution vectors of Gray wolves have been optimized, solution vectors of grey Wolf in all rows are traversed, and the solution with the smallest objective function value and its solution vector are retained. At this point, we have completed a loop.

(4) Elite selection system: In the process of searching solution vector, Gray Wolf algorithm is always accompanied with random and blindness. In order to deal with this problem, we adopt the elite selection system to improve the node sorting path. In the large-scale iterative optimization process, we retain some better node sorting, which is called elite. We retain these elites as the starting sequence of nodes for the next cycle, which can greatly improve the speed of calculation and reduce the complexity of the algorithm to a large extent.

3.3. Key node identification method—IABFI

With the above improved strategy, the critical node identification process with cumulative robust CR as the objective function is shown in figure 8.

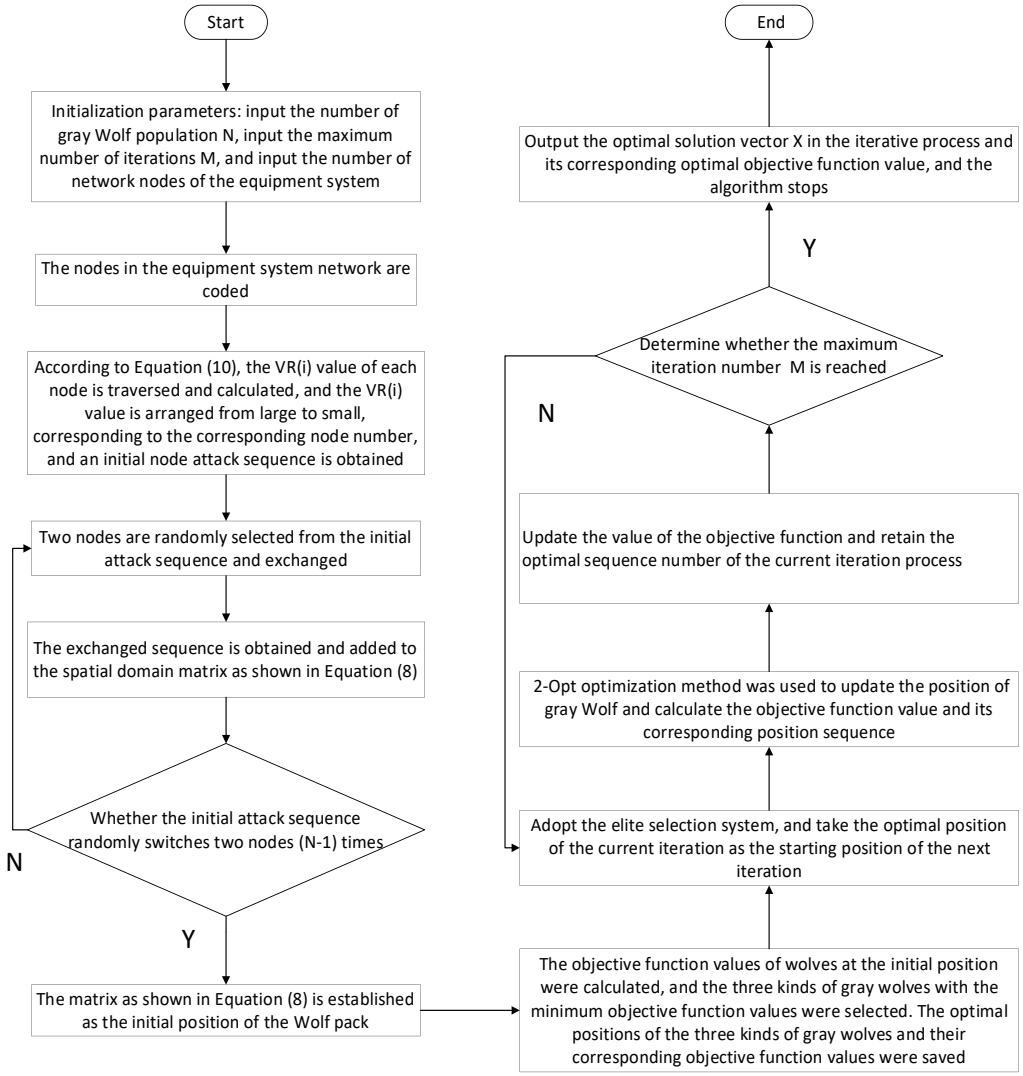

Figure 8. IABFI—key node identification flowchart

4. EXPERIMENTAL SIMULATION AND ANALYSIS

In order to verify the superiority of the proposed IABFI - node identification algorithm in searching key nodes in the equipment system network. In this section, network models of equipment system based on random networks (ER), small world networks (WS) and scale-free networks (BA) are established respectively, They correspond to Figures 9, 10 and 11 below, and verified under different network scales. The key node identification algorithm proposed in this paper is compared with the traditional algorithms of betweenness centrality, degree centrality, improved K-shell algorithm, PageRank algorithm, eigenvector centrality, closeness centrality, etc. According to the order of nodes arranged by each algorithm, the corresponding nodes are attacked in turn, and the simulation is carried out in the above three devices system network.

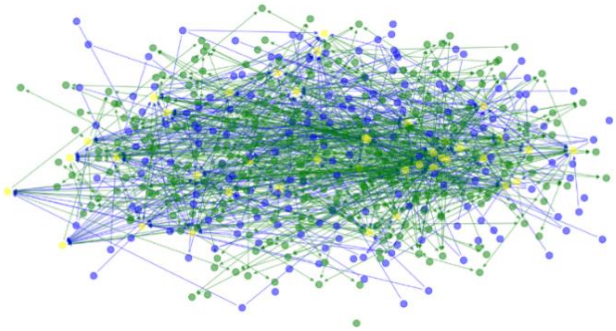

Figure 9. Network model of equipment system based on random network

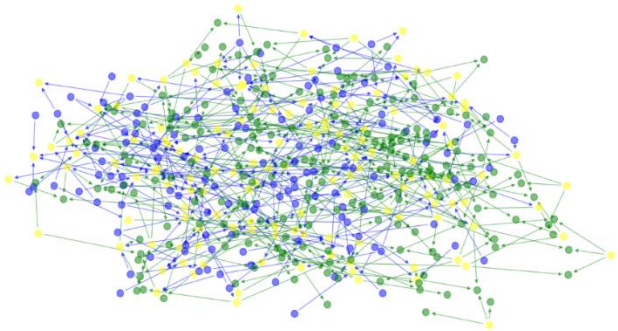

Figure 10. Network model of equipment system based on small world network

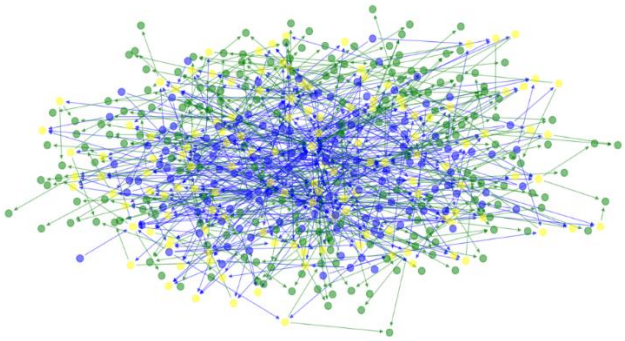

Figure 11. Network model of equipment system based on scale-free network

Combined with corresponding military applications, the scale of the above three equipment system network models is 450 nodes, in which S node is blue, D node is yellow, and I node is green.

After the importance of nodes is sorted according to the corresponding algorithm index, the nodes in the equipment system network are attacked in sequence according to the order of node importance. As nodes are attacked, they become ineffective, and the number of effective OODA chains in the equipment system network decreases, which is accompanied by the decrease of the operational performance of the equipment system network.

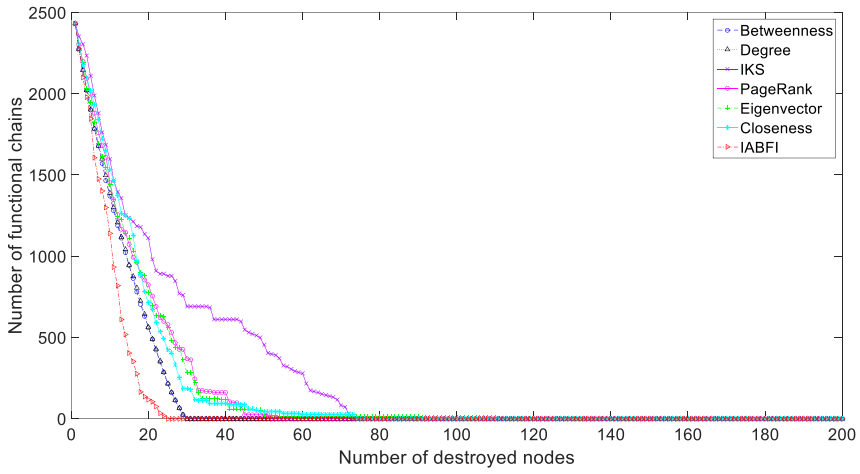

Figure 12. Network performance variation of equipment system based on random network (200 nodes)

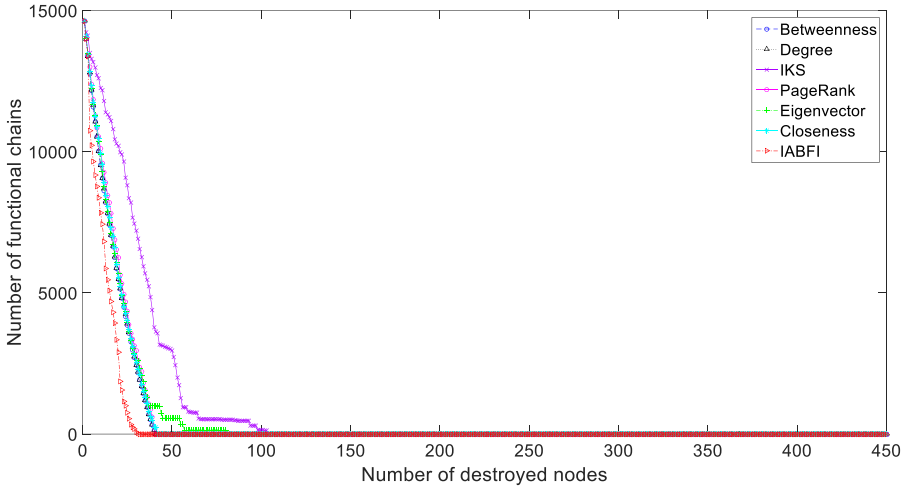

Figure 13. Network performance variation of equipment system based on random network (450 nodes)

Figure 12 and Figure 13 show the network model of equipment system based on random network with node size of 200 and 450 respectively. In these two networks, the nodes are sorted according to the above indexes, and the nodes are hit in sequence according to the sorted sequence. The effective OODA chain is reduced in the equipment system network of the above two scales. In fig.12, there are 2432 original OODA chains at the initial time. A node sequence is obtained based on IABFI algorithm. After the first 24 nodes in the sequence are attacked, there is no function chain in the whole network. In other sorting algorithms, the effect is relatively good is the intermediate centrality and degree centrality. These two algorithms sort the sequence, the number of OODA chains in the entire equipment system network is zero until the first 29 nodes in the sequence are attacked. Others such as PageRank algorithm, feature vector centrality, proximity, improved K-shell algorithm, the sequence obtained by the arrangement of these algorithms. According to sequence, after the first 51, 110, 73 and 72 nodes are attacked, the function chain completely disappears. In addition, it can be seen from the figure above that, compared with other algorithms, the function chain in the equipment system network declines particularly fast after the nodes are attacked according to the sequence sorted by IABFI algorithm, the number of function chains decreases rapidly at the beginning and then slows down,

which fully demonstrates the effectiveness of the algorithm for node sorting. According to the sequence sorted based on the IABFI algorithm, the nodes are attacked sequentially. The nodes initially attacked are generally the hub nodes connected with a large number of OODA chains. After being attacked, the number of OODA chains in the equipment system network is greatly reduced. As shown in Figure 13, the original number of OODA chains is 14,621. After the attack, the downward trend is basically consistent with the model diagram of 200 nodes. In fig.13, the number of function chains in the equipment system network is large, Under the ranking attack of algorithms such as closeness centrality, eigenvector centrality and PageRank algorithm, the difference between algorithms in fig. 13 is less obvious than that in fig. 12. However, it is still clear that compared with other traditional algorithms, the effect of IABFI algorithm is still more obvious.

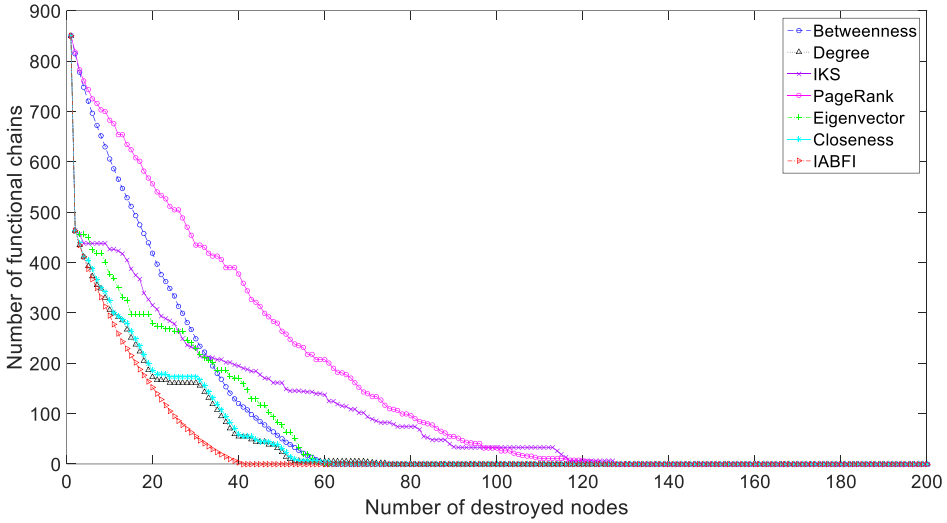

Figure 14. Network performance variation of equipment system based on small world network(200 nodes)

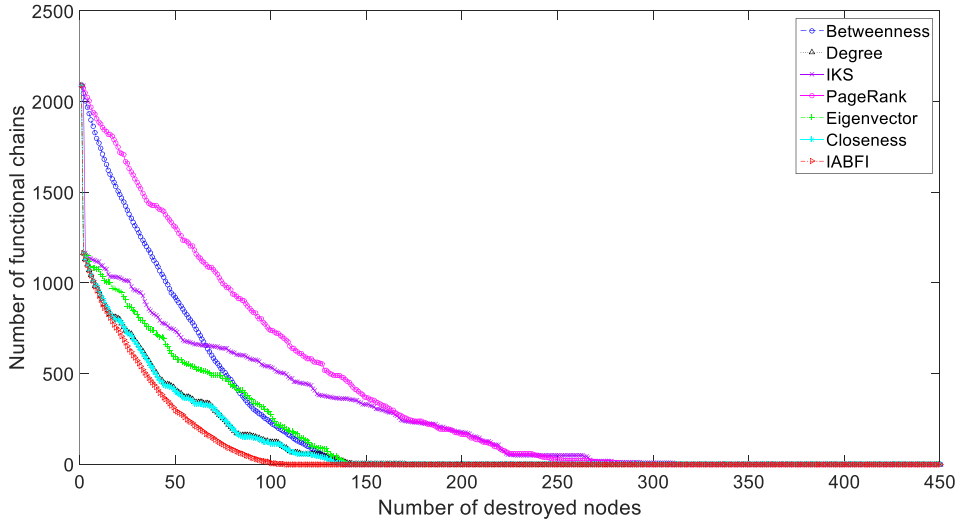

Figure 15. Network performance variation of equipment system based on small world network (450 nodes)

In the above small world network based on equipment system model, there is an important node at the beginning. Betweenness centrality and PageRank algorithm did not find this node in time. As a result, its recognition effect is not very good compared with that in the random network of equipment system network model. Compared with the random network model, the relative error of the PageRank algorithm is more huge. In Figure 14, after the first 125 nodes in the equipment system network are attacked, its function chain is 0, and its cumulative robustness is as high as 32,244, the cumulative robustness of IABFI algorithm in the same period is 7467. It can still be clearly seen from the above two figures that compared with other algorithms, IABFI algorithm still has the most obvious effect. After nodes in the equipment system network are attacked, the decline trend is the most dramatic. In Figure 14, sorted by IABFI algorithm, after attacking the 40 nodes in the sequence, the function chain in the equipment system network is 0, and the effect is the best. In Figure 15, the effect trend is similar to that in Figure 14. According to the simulation diagram of attack effect of WS network nodes of different scales, it can be seen that IABFI algorithm is applicable to the equipment system network of different scales mentioned above.

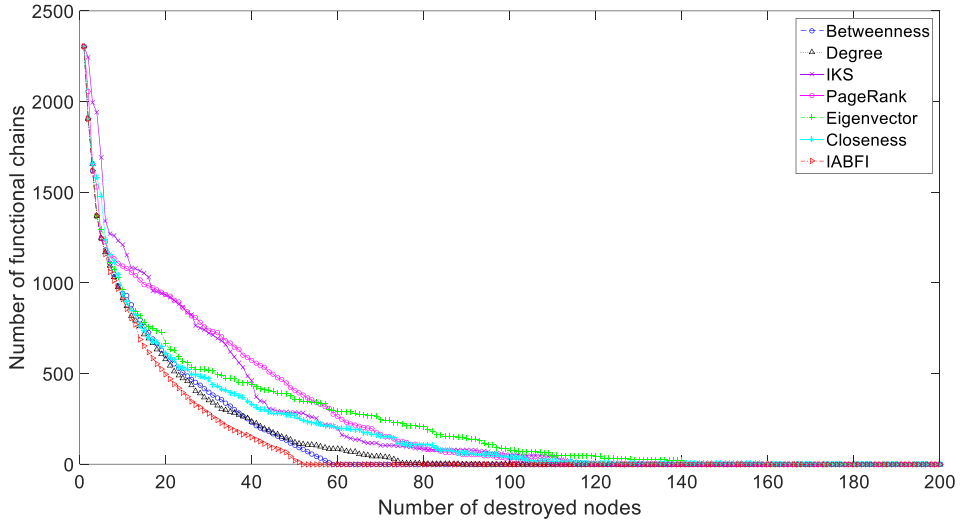

Figure 16. Network performance variation of equipment system based on scale-free network (200 nodes)

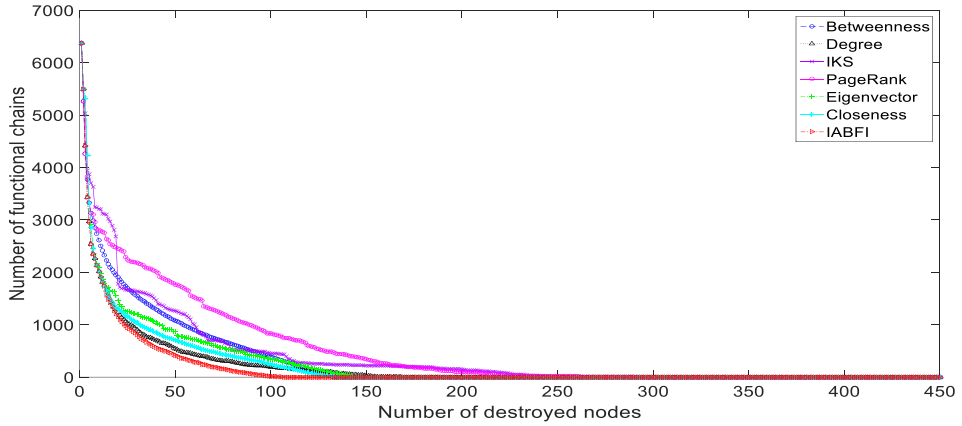

Figure 17. Network performance variation of equipment system based on scale-free network (450 nodes)

We conduct further simulation under scale-free network. As shown in FIG. 16 and 17, the optimization effect of IABFI algorithm is as efficient and accurate as always in scale-free network. Compared with the traditional algorithm which is affected by the network topology structure, the performance of the algorithm differs greatly under different network models. It can be seen that the performance of IABFI algorithm has been stable, and compared with the traditional algorithm, the performance is also the most superior. As can be seen from the figure above, according to the IABFI algorithm, after the nodes in the sequence are hit, the number of OODA chains in the equipment system network still declines the fastest.

According to the above six simulation graphs, the variation trend of the number of function chains in the equipment system network after the nodes in the equipment system network are attacked is compared under different topologies and node sizes. Compared with other algorithms, IABFI algorithm can identify key nodes accurately and quickly, and it is suitable for different network models, unlike the traditional node identification algorithm in different equipment system network model performance difference, it shows that the IABFI algorithm is effective and stable.

5. CONCLUSIONS

The application of optimization algorithm to node optimization has also appeared in previous researches on isomorphic networks. Single point optimization based on tabu search algorithm or discrete firefly algorithm, complex network robustness measure considering network efficiency. They have their own advantages and disadvantages under different measurement indexes. In this paper, effective OODA chain is considered, and the improved Gray Wolf algorithm has more prominent advantages in convergence speed and global search than other algorithms. According to the analysis of the change of the number of effective OODA chains after the network nodes of the equipment system are attacked in sequence with the sequence K, it is shown that compared with other traditional node sorting algorithms, IABFI can be applied to network models with different network topologies and different node sizes, and its performance is more stable. It can find the key nodes in the network most quickly. The IABFI algorithm can be used to quickly find out the key nodes in the equipment system network, and in the military action against the enemy, it can find out the key nodes, take decapitation action, and destroy the enemy's military operation network in the shortest time. At the same time, IABFI algorithm can also be used to find out the important key nodes in our equipment system network, backup and take priority protection strategies for the important nodes. It can ensure that we can maintain the original function of our equipment system network and win the battle in the face of external attack in the fierce military confrontation. However, in the process of our study, we found that it is often difficult to simulate when drawing large-scale node graph. When the node scale is large, the speed of simulation results is also slow. In the following research process, we will do targeted research on this issue.

ACKNOWLEDGEMENTS

Project supported by the National Defense PreResearch Quick Support Foundation of China (no.80911010 302)

REFERENCES

[1] Z.Ni, (2004) “Modeling and simulation of weapon system confrontation”, Military Operations Research and Systems Engineering, Vol. 18, No. 1, pp2-6.

[2] K. Li, W. Wu, (2016) “Research Status of Weapon Equipment System-of-Systems Based on Complex Network”, Journal of Academy of Armored Force Engineering, Vol. 30, No. 4, pp7-13.

[3] J. Wang, M. Wang. & W. Ding, (2016) “A Value-Focused Decision Making Framework for System of Systems Architecture”, Computer & Digital Engineering, Vol. 44, No. 10, pp1948-1951+1962.
[4] R. Li, H. Zhang. & Y. Yin, (2011) “The ideal understanding of several basic problems of weapon equipment architecture optimization”, Military Operations Research and Systems Engineering, Vol. 25, No. 2, pp5-10.
[5] Deng Y, Wu J & Tan YJ, (2016) “Optimal attack strategy of complex networks based on tabusearch”, Physica A Statistical Mechanics, Vol. 442, No. 1, pp74-81.
[6] X. Liu, G. Xu & P. Yang, (2019) “Node importance evaluating of network based on combination weighting VIKOR method”, Application Research of Computers, Vol. 36, No. 8, pp2368-2371+2377.
[7] T. Wang, W. Dai & P. Jiao, (2016) “Identifying influential nodes in dynamic social networks based on degree-corrected stochastic block model”, International Journal of Modern Physics B, Vol. 30, No. 16.
[8] Z. Shao, S. Liu &Y. Zhao, (2019) “Identifying influential nodes in complex networks based on neighbours and edges”, Peer-to-Peer Networking and Applications, Vol. 12, No. 6, pp1528-1537.
[9] H. Li, L. Zhou & W. Xin, (2017) “Optimization of networked Combat Equipment Architecture based on optimal tree”, Military Operations Research and Systems Engineering, Vol. 31, No. 4, pp47-53.
[10] X. Feng, C.Hu. & C. Xu, (2019) “Key node recognition method based on optimal network efficiency”, Computer Engineering And Design, Vol. 40, No. 2, pp328-335.
[11] MIRJALILI S, MIRJALILI S M& LEWIS A, (2014) “Grey wolf optimizer”, Advances in Engineering Software, Vol. 69, No. 7, pp46-61.
[12] S. Gao, L. Meng, (2019) “Greedy randomized adaptive grey wolf optimization algorithm for solving TSP difficulty”, Modern Electronics Technique, Vol. 42, No. 14, pp46-50+54.
[13] X. Zhang, Y. Zhang & Z. MING, (2021) “Improved dynamic grey wolf optimizer”, Frontiers of Information Technology & Electronic Engineering, Vol. 22, No. 6, pp877-891.
[14] R.XU, M. Cao. &M. Huang, (2018) “Research on TSP-like problem based on improved Gray Wolf optimization algorithm -- taking tourism as an example”, Geography and Geo-Information Science, Vol. 34, No. 2, pp14-21.
[15] SAREMI S, MIRJALILI S Z& MIRJALILI S M, (2015) “Evolutionary population dynamics and grey wolf optimizer”, Neural Computing and Applications, Vol. 26, No. 5, pp1257-1263.
[16] CROES G A, (1958) “A method for solving traveling-salesman problems”, Operations Research, Vol. 6, No. 6, pp791-812.

AUTHORS

Cheng Huang graduated from The Department of Electronics and Information Engineering of Chifeng University in 2019 with a bachelor's degree. He is currently pursuing a master's degree in school of communication and information engineering at Chongqing University of Posts and Telecommunications in Chongqing, China. His research interest is equipment system network.

Yong Gang Li graduated from Shanghai Jiao Tong University in 2007 with a ph. D. degree in communication and information systems. From 2012 to 2013, he was a visiting scholar in the Department of Electronic Engineering, University of Wisconsin-Madison, engaged in large-scale network signal processing research. His research interest covers network complexity, tactical communication network simulation, network security, and network visualization.

Ying Wang graduated from Cangzhou Normal University with a bachelor's degree in Communication Engineering in 2019. She is currently pursuing a master's degree in school of communication and information engineering at Chongqing University of Posts and Telecommunications in Chongqing, China. His research interest is equipment system network.

© 2022 By AIRCC Publishing Corporation. This article is published under the Creative Commons Attribution (CC BY) license.

PREPARING LEGAL DOCUMENTS FOR NLP ANALYSIS: IMPROVING THE CLASSIFICATION OF TEXT ELEMENTS BY USING PAGE FEATURES

Frieda Josi[1], Christian Wartena[1] and Ulrich Heid[2]

[1]University of Applied Sciences and Arts Hanover, Expo Plaza 12, 30559 Hannover, Germany
[2]University of Hildesheim, Universitätsplatz 1, 31141 Hildesheim, Germany

ABSTRACT

Legal documents often have a complex layout with many different headings, headers and footers, side notes, etc. For the further processing, it is important to extract these individual components correctly from a legally binding document, for example a signed PDF. A common approach to do so is to classify each (text) region of a page using its geometric and textual features. This approach works well, when the training and test data have a similar structure and when the documents of a collection to be analyzed have a rather uniform layout. We show that the use of global page properties can improve the accuracy of text element classification: we first classify each page into one of three layout types. After that, we can train a classifier for each of the three page types and thereby improve the accuracy on a manually annotated collection of 70 legal documents consisting of 20,938 text elements. When we split by page type, we achieve an improvement from 0.95 to 0.98 for single-column pages with left marginalia and from 0.95 to 0.96 for double-column pages. We developed our own feature-based method for page layout detection, which we benchmark against a standard implementation of a CNN image classifier.

The approach presented here is based on corpus of freely available German contracts and general terms and conditions. Both the corpus and all manual annotations are made freely available. The method is language agnostic.

KEYWORDS

PDF Document Analysis, Legal Documents, Layout Detection, Feature and Text Extraction, Classification, Machine Learning, Deep Convolutional Networks, Image Recognition.

1. INTRODUCTION

Many documents are only available as PDF. This is especially the case for legal documents where one exact copy including layout and signatures is distributed and archived. Extracting the text from a legal document is often challenging since e.g. contracts often have a complex structure with lists, footnotes, side notes, multiple columns, headers and footers and so on. Moreover, contracts often consist of several parts, like address page, signature page, project description, terms of service etc. which each may have a completely different layout.

In order to extract texts from a PDF we first identify characters, then regions of closely neighbouring characters (words) and finally regions with dense text. Once we have defined these

David C. Wyld et al. (Eds): NATP, ACSTY, CCCIOT, MLSC, ITCSS - 2022
pp. 17-29, 2022. CS & IT - CSCP 2022 DOI: 10.5121/csit.2022.120102

regions, we classify them into several types before we extract the text. An example of a PDF page and the extracted text regions is given in Figure 1. In this figure, the class of each text region is given.

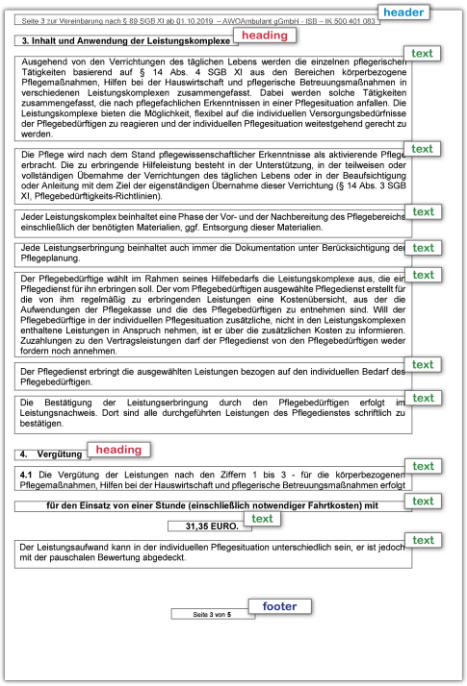

Figure 1. Example of a page from a contract with marking of text regions and their classes.

Our approach can also be used for documents where the number of columns on each page is not known and can be different on each page. By improving text extraction, large collections of documents can be processed more efficiently, as some sources of error are reduced.

E.g. texts located in the header and footer can be filtered out instead of being merged in the sentence covering the page break; texts from different columns can be brought in correct order and by the identification of headings the text can be split up in chapters and sections.

2. RELATED WORK

Extracting text from scanned PDF documents is still a challenge when it is not known how many columns the individual pages in the documents have, when the task is to ignore header and footer text, and when a large number of PDF documents are to be processed automatically.

The documents used in the present work were only available in PDF, not originating from a text processing tool but scanned from printed and signed paper documents. This is a very common situation for legal documents. Therefore, the analysis of legal documents often has to begin with extensive pre-processing followed by comprehensive cleaning of the texts.

Different approaches were tested for extracting the texts from PDF files and for recognizing the structure of the documents. [1] extract text elements from PDF files to analyze the structure of Chinese books that were available in PDF format. After extraction, the content was assigned both a physical and a logical structure. However, since the data came from books, it was possible to assume that all pages have the same layout, i.e. the number of columns and the positioning of headings is consistent across all pages of the book. This allows the definition of global typography classes. The authors divided the logical structure on page level and on document

level. The page level contains the hierarchical arrangement of text elements, such as headers, figures, tables, and footnotes. The document level included the chapter structure, author metadata, and book title. For page-level logical structure extraction, the text and individual letters were extracted from these text blocks to obtain additional features such as boldface for headings. The authors obtained very good results with this process. All classes such as *header and footer*, *heading font* etc. are identified with an accuracy above 90% and the lowest hit rate has *heading font* with 87.94%. The authors used 1,000 books for their method.

[2] use layout analysis to improve the delimitation of sentences boundaries in financial reports. They use layout analysis to filter out tables, among other things, as these are not helpful for sentence boundary detection. They try to separate the content of the document from other information.

Contiguous text sections on a document page are not necessarily extracted as a single unit, so there are also some works that deal with merging contiguous text areas from extracted PDF files, for example[3]. The authors have developed a three-stage procedure for this purpose. In the first stage, contiguous text blocks are to be identified on a layout basis. In the next stage, a rule-based classification of the text blocks is performed using categories, and in the third stage, these classified text blocks are to be summarized in the correct order. At the end, the text can be extracted from the summarized text blocks.

In a simple extraction of text from PDF files, text is also extracted from the header and the footer, such as the page number, or the name of the contract creator in the header. However, these components of the document are a hindrance for analyzing the content of the contract and also for comparing document versions. For the detection of headers and footers, [4] used a layout-based approach. This approach is based on the use of geometric coordinates. In addition, the authors use the occurrence of digits as an indicator of a text element in the header or footer and the length of the text as a supplementary feature. Using the coordinates of the text blocks from the PDF files, a structural sorting per page is possible.

Methods for analyzing the visual, physical, and logical structure of PDF files are often developed for scientific papers, for example by [5], [6], [7] and for Newspapers e.g. by [9]. In [9], also technical documentation is transferred into an XML structure starting from a PDF document. Legal documents have however not yet been extensively used in extraction and analysis work. Some research projects in this area are described by [10], which identifies argumentation structures in court proceedings, and by [11], which classifies and automatically summarizes legal texts. For the removal of texts which originate from headers and footers, [12] use features that are defined based on the neighboring pages. As soon as the text line candidates exceed a threshold of equal characters and correspond to a minimum occurrence, these texts can be filtered out as headers and footers.

The work of [13] is similar to our objective - classifying text elements - but they use a large variety of document types e.g. email files, power point files, as well as of text formats, e.g. word files, PDF, and others. Therefore, they divide for example the class "heading" into headings in tables, of emails and of other document types. [13] use a conditional random field model for the prediction and achieve an average accuracy of 0.83 for all 13 classes which they distinguish.

Scientific works, in which Convolutional Neural Networks (CCNs) are used for the classification task of image-based document pages, are reported by [14], [15], and [16], among others. [14] proposed a method (document domain randomization (DDR)) that does not need manually annotated document pages, but works with generated pseudo-pages. However, the extraction of the texts is not the goal, since the pages are randomly composed from components of scientific

papers. Thus, the authors do not need to use manually labelled data. The goal is to separate textual areas from figures and tables. In [15], a CNN is used to classify document types. The classes for the image-based document pages are e.g. “email”, “news article”, “file folder”, “letters”, “memo”, and so on. [16] performs object detection on image-based document pages using CNN. The objects they want to identify are “stamps”, “logos”, “signatures”, “tables” and “text blocks”.

3. DATA AND METHOD

In this section we will present our method for the classification of text elements and the data used for training and evaluating our methods.

3.1. Document Corpus

For this work, a corpus of publicly available German contracts was compiled, primarily contracts from the city administrations of Hamburg and Bremen, dating from 2014 to 2019, that were released on the internet in the context of the cities open government policy. In addition to these two sources, general terms and conditions, found on the internet, were added to the corpus.

The sources used for the corpus can be found in appendix A.

For a part of the corpus all pages were classified as being in one-column layout, two column layout or one-column layout with marginals. For another small part of this corpus all text regions detected were classified manually. The annotation was done by a student assistant who corrected the classification of a first version of the classifier trained on a very small set of data, not part of the current corpus. Over ten thousand text regions were classified in this way. Exact numbers are given in Table 1. The documents we used for the classification of the text boxes are the same documents that were used as test documents for the page layout recognition. The documents for the training set were separated and no longer used for the recognition of the text classes. For the classification of the text classes, we used x-fold cross-validation with x = 10.

Table 1. Size of the corpus.

	Layout prediction			**Text class prediction**
	Test set	Training set	Total	Total (cross valid. used)
Number of documents	64	249	313	70
Number of pages	417	1,859	2,276	276
Number of text elements	-	-	-	20,938

3.2. Method

The goal of our work is to classify text fields on a page in order to improve the extraction from a scanned PDF-document. An obvious way to do this is to train a classifier using various features of a text region based on the size, position and content of the text region. Many of the geometric features, however, are context dependent, i. e. the meaning of a feature depends on the type of page layout. Thus, we first classify each page, since in legal documents it is often found that different layout types are used within a single document. Subsequently, we use the layout information to classify the text regions. Here we either can use the probabilities for each layout class as features for the second classifier, or alternatively we train different classifiers for each layout class.

The pipeline for using the global layout features to improve the classification of text classes is given in Figure 2.

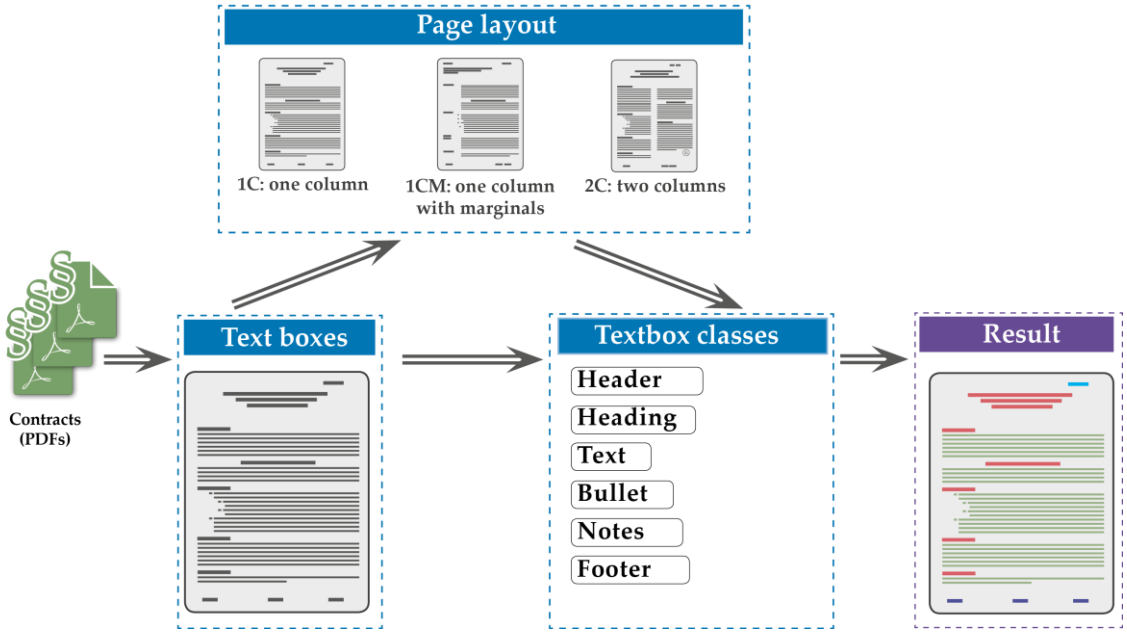

Figure 2. Pipeline for using global page features to classify layout elements.

In the first step, we have compiled a collection of legal documents. Then we extract all text elements from a PDF page and get the exact coordinates and other features such as font size, etc. for each element. After that, the layout of each page is classified using the marker method, see Section 3.3.

We use these layout classes, e.g. two-column page, as a feature to improve the classification of texts classes. The result of this process is a document collection where for each text of an extracted text area, it is known to which class it belongs. Thus, classes that are not needed can be filtered out and the remaining text can be used for text analysis or for other processes.

3.3. Layout Recognition

As mentioned before, in legal documents we find various types of page layout that often change within the document or even on one page. E.g. the main text of a contract can be in a one-column layout while the general conditions, that are part of the contract is in two column layout, followed by a page of general remarks and signatures again in a kind of one-column style. We identified three main layout types in various collections of contracts: one-column layout, two-column layout and one-column with marginals. In the current collection of contracts, presented above, the last type of layout usually is a kind of table, where we have short definitions on the left-hand side and a longer text at the right-hand side. In another collection of contracts that we have used, but that cannot be made publicly available, we also found many examples where the section headings were written in the left margin.

In order to predict which layout each individual PDF page will have, we use 100 vertical markers for each page. For each of these 100 markers, the heights of the text areas that this marker intersects with are accumulated and normalised (with the highest marker value per page). To avoid that text elements from the header or footer distort our calculations, we have not taken into

account the top and bottom height of a page in the calculation. This is shown schematically in Figure 3. The dark blue highlighted line in Figure 3 intersects with three text boxes (see orange ellipses in the figure). The values of all 100 markers are calculated in this way. The text box height is marked in the illustration as a text area with an orange bold border, the other text areas on this page are highlighted by dark grey rectangles.

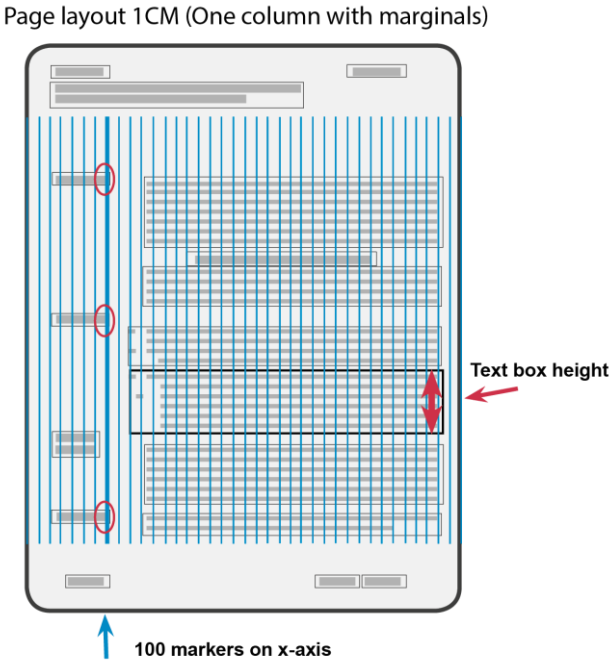

Figure 3. Schematic representation of the calculation of the marker values.

Figure 4 shows prototypical examples for each layout and the associated marker values. The marker values for all three layout types are easily distinguishable for clear prototypical page examples. For the page with one column, the markers value decrease to the right (less intersections, as not all lines are completely filled with text). For the one-column layout with the left marginalia, the markers remain lower over the area of the left headings and increase significantly at the start of the text column. Towards the right margin the values then decrease again. For the two-column page, the low values between the two columns are clearly visible. Here in the example, it is a prototypical page that is well filled with text, but if the two-column page contains only partial text, an incorrect classification may occurs.

3.3.1. Our feature-based Approach

For page layout classification, we trained a Random Forest classifier and a Support Vector classifier. The hyperparameters for Support Vector Classifier are optimized for 100 markers. We use Stratified KFold with random state 42 and a fold number of 10 for the Random Forest Classifier and the Support Vector Classifier. The features we use for training the classifiers are the 100 values of markers truncated at the top and bottom of the page and the area of all text boxes per page. The results are given in Table 3.

3.3.2. Image Classification with Deep Convolutional Networks

For the image-based classification of the document pages with convolutional networks, we use an implementation of PyTorch (https://github.com/aleksandraklofat/image_classifier). The pre-trained model vgg16 [17] (https://pytorch.org/vision/stable/models.html) were trained with the own document pages (417 document pages were converted to png images with pdf2image). We used the default settings. The document pages from the test set are also the pages extracted by our feature-based procedure. The results are given in Table 5.

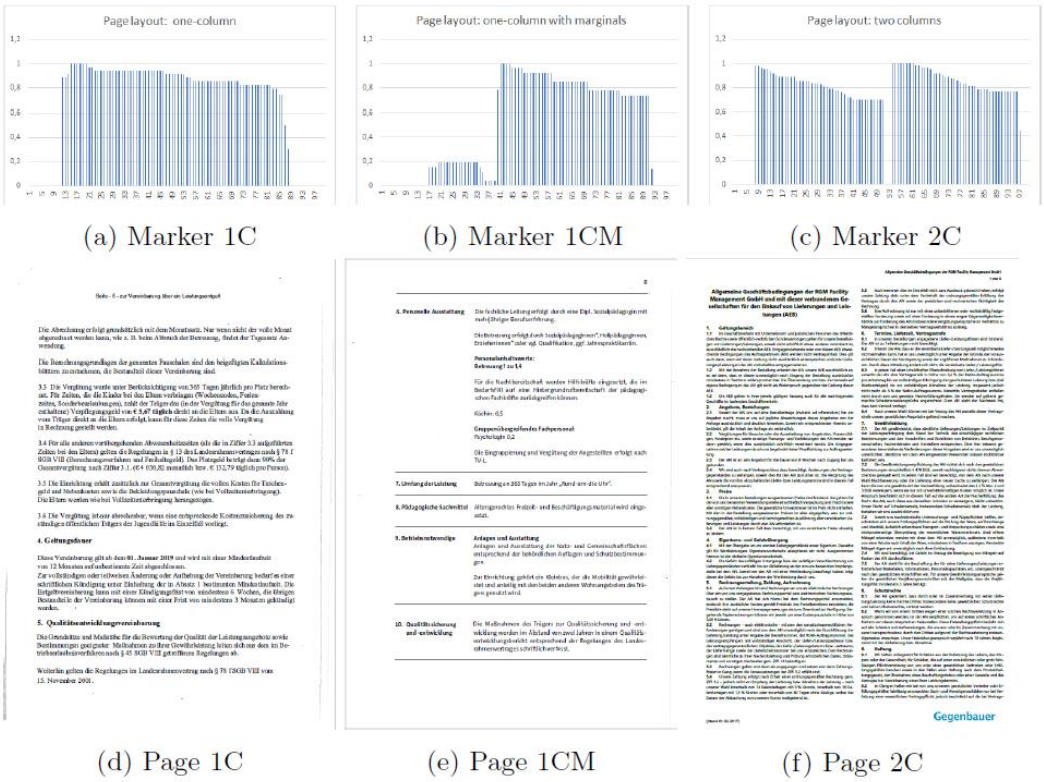

Figure 4. Prototypical page examples and the corresponding marker values.

3.4. Classification of Text Elements

For each page we first extract all dense and cohesive text areas. For the extraction of these areas we use PDFMiner [18]. The exact areas that are extracted of course depend on the parameters used for extraction and we have to be sure that the same values have been used for training and application of the classifier.

A text area can have various functions on a page. We identified six frequently occurring types of text regions that can be distinguished in our collection of documents. Other types, like image captions, might frequently occur in other documents but are not present in our corpus of contracts and general terms and conditions of business. The six types are:

- **Header:** Line of text on the top of the page, with one to three text elements not belonging to the main text flow.
- **Heading:** Headings and subheadings.
- **Text:** Main text, including numbered paragraphs, text from listings etc.

- **Bullet:** Bullet points.
- **Notes:** Hand written notes and stamps or text from a small table.
- **Footer:** Includes everything below the last text line, e.g. page numbers, footnotes.

The features we use for the text structure elements are: the geometry of the text box, e.g. its coordinates; the neighbourhood of the text box, the presence of adjacent text boxes i.e. its distance between the text boxes; the text box area; the height and width of the text box; the features for recognising headings, such as bold, capital letters, a colon is the last character of a text box; the number of special characters and whether the text of a text box element is a bullet and others (detailed list in Table 2).

Table 2. Features for each text box for classification.

Feature(s)	**Data type**
Geometry of the box	
Coordinates (4 values)	float
Height of text box in pixels	float
Width of the text box in pixels	float
Area of the text box	float
Neighbourhood	
Distance to the top | bottom | left | right of the page (4 values)	float
Adjacent text box to the top | bottom | left | right (4 values)	boolean
Font	
Font is bold	boolean
Font is italic	boolean
Font size	float
Font is capitalised	boolean
Text	
Starts with a paragraph mark	boolean
Text ends with a colon in the text box	boolean
Text box consists of bullets	boolean
Number of characters of an extracted text box	integer
Number of special characters	integer

From the page layout classification, four values representing the page layout are added as features for each text element. This is to improve the prediction for the text classes. The prediction of the text structure classes is performed using a *Support Vector Classifier* (SVC) as well as a *Random Forest Classifier*, both from the SciKitLearn-Library (https://scikit-learn.org/stable/index.html). Finally, we train a classifier on a part of the manually annotated data. For evaluation we use the test set for layout prediction and a 10-fold cross validation scheme for text classes. For the cross validation we use stratified sampling and for each partition we balance not only the fractions of the text box classes that have to be predicted but also balance the layout types that the boxes were taken from. Thereby we ensure that all types appear in both training and test sets. Alternatively, we do not use the probabilities for each layout class but simply assume that the most probable one is the correct class and train three different classifiers for each class.

4. RESULTS AND EVALUATION

4.1. Page Layout

The results for the classification of the page layout summarized in Table 3. A Random Forest and an SVC model was trained for the classification. With the Random Forest model we can achieve an accuracy of 0.94 and with the SVC model even 0.95. In the confusion matrix (Table 4), errors in the prediction of the layout classes are shown, from SVC. According to this, between the layout classes 1C and 1CM there are few document pages that are in the intermediate range. 17 pages are incorrectly predicted as a single-column page, but only one one-column page is predicted as two-column page. Table 5 shows the results for image classification with a CNN. The results do not reach the values with the feature-based approach from Table 3. The accuracy is 89%. Detailed values are shown in Table 5 and in Table 6 the confusion matrix gives an overview of the misclassified document pages.

Table 3. Results for layout prediction with SVC and Random Forest.
Accuracy: SVC: 0.95; Random Forest: 0.94

101 Features					
	Class	**Precision**	**Recall**	**F1-score**	**Number of pages**
Random Forest	1C	0.92	1.00	0.96	291
	1CM	0.97	1.00	0.99	37
	2C	1.00	0.73	0.84	89
	wgt. avg:	0.97	0.91	0.94	
				total:	417
SVC	1C	0.94	0.99	0.97	291
	1CM	0.93	1.00	0.96	37
	2C	0.99	0.80	0.88	89
	wgt. avg:	0.95	0.93	0.95	
				total:	417
Heights of the intersected text boxes per marker (100) and area of all text boxes per page (1).					

Table 4. Confusion matrix from SVC for page layout classes

Predicted **Real**	**1C**	**1CM**	**2C**
1C	288	2	1
1CM	0	37	0
2C	17	1	71

Table 5. Results by CNN implementation with PyTorch for page layout classification as image classification. With an accuracy value of 0.89.

	Class	**Precision**	**Recall**	**F1-score**	**Number of pages**
CNN	1C	0.97	0.80	0.88	291
	1CM	0.32	0.84	0.47	37
	2C	0.88	0.79	0.83	89
	wgt. avg:	0.72	0.81	0.73	
				image-based document pages, total:	417

Table 6. Confusion matrix from CNN results by implementation with PyTorch for page layout classification as image classification

Predicted: Real:	**1C**	**1CM**	**2C**
1C	234	48	9
1CM	5	31	1
2C	2	17	70

4.2. Text Classes

4.2.1. Results of Classification

For the prediction of the text classes, we compare two methods. As a baseline, we classify the text classes without features that contain information about the layout. The results for both classifiers are presented in summarized form in Table 7. With the SVC model we can achieve an accuracy of 0.89 and with the Random Forest model 0.95.

Table 7. Results for text class prediction. Accuracy: SCV: 0.89; Random Forest: 0.95

	class	**precision**	**recall**	**F1-score**	**Number of text elements**
SVC	Header	0.84	0.76	0.80	394
	Heading	0.85	0.83	0.84	2,069
	Text	0.96	0.98	0.97	17,241
	Bullet	0.84	0.77	0.80	295
	Notes	0.69	0.45	0.54	437
	Footer	0.95	0.86	0.90	502
	wgt. avg.	0.85	0.78	0.88	
				total:	20,938
Random Forest	Header	0.92	0.84	0.87	394
	Heading	0.58	0.51	0.54	2,069
	Text	0.92	0.95	0.94	17,241
	Bullet	0.72	0.56	0.63	295
	Notes	0.63	0.44	0.52	437
	Footer	0.97	0.93	0.95	502
	wgt. avg.	0.79	0.70	0.94	
				total:	20,938

4.2.2. Results of separate Classification for each predicted Layout Class

We execute a classification in which the text classes are classified separately according to the page layout. This means that only text elements that are on a one-column page are classified together. We do the same with the other two page layout classes. The distribution of the text elements, across the classes, is shown in Table 8. The results for this separate prediction are shown in Table 9, again for the two classifiers SVC and Random Forest. The prediction of the classes for the texts from the two-column pages and from the page type 1CM benefit from this procedure. With the SVC model, we achieve an accuracy of 0.87 for 1C page type, 0.90 for 1CM, and 0.84 for 2C for the text elements. Using the Random Forest model, we obtain 0.94 for 1C, 0.98 for 1CM, and 0.96 for 2C. More precisely, the improvement for 1CM and 2C can be seen, for example, for the class “heading”: this text type can be used as an important anchor in further text processing. We get 0.95 for the text type “heading” for 1CM and 0.90 for 2C. Single-column

pages (1C) do not benefit from detection and separation by page layout types, but do also not represent a major challenge in the extraction process of contract documents.

Table 8. Distribution of the text elements.

Number of text elements			
Class	1C	1CM	2C
Header	260	66	68
Heading	1,015	326	726
Text	9,349	1,722	6,170
Bullet	183	44	68
Notes	339	15	83
Footer	308	82	112
total:	11,454	2,255	7,227

Table 9. Results for text class prediction separated for each layout class.

	Class	**Precision**				**Recall**				**F1-score**			
		1C	1CM	2C	wgt. avg.	1C	1CM	2C	wgt. avg.	1C	1CM	2C	wgt. avg.
SVC	Header	.77	.96	.91	.83	.41	.68	.43	.46	.54	.80	.58	.59
	Heading	.54	.95	.34	.53	.42	.64	.35	.44	.47	.77	.34	.47
	Text	.90	.89	.90	.90	.96	.99	.92	.95	.93	.94	.91	.92
	Bullet	.73	1.0	.76	.78	.39	.02	.46	.35	.51	.04	.57	.45
	Notes	.70	.83	.07	.58	.47	.33	.04	.38	.56	.48	.05	.46
	Footer	.94	.96	1.0	.96	.72	.79	.80	.75	.82	.87	.89	.84
wgt. avg.		.86	.91	.83	.86	.87	.90	.84	.86	.86	.89	.84	.86
Random Forest	Header	.82	.90	.90	.85	.76	.92	.65	.77	.79	.91	.75	.80
	Heading	.80	.96	.88	.90	.83	.94	.92	.93	.81	**.95**	**.90**	.92
	Text	.96	.99	.98	.97	.97	.99	.98	.98	.97	.99	.98	.98
	Bullet	.80	.91	.87	.83	.68	.89	.91	.76	.74	.90	.89	.80
	Notes	.73	.67	.46	.68	.56	.53	.37	.52	.64	.59	.37	.59
	Footer	.94	.96	.96	.95	.87	.98	.95	.91	.91	.97	.95	.93
wgt. avg.		.94	.98	.96	.95	.94	.98	.96	.95	.94	.98	.96	.95
Legend: 1C = one-column, 1CM = one-column with marginals, 2C = two column													

Table 10 compares all values for accuracies. The values for the classification of the text elements, which were all trained together and the classification of the text elements divided by page layout. In order to compare the results directly, the accuracy values for the text elements divided by page layout were added as a weighted average. By splitting by page layout, the classification with the SVC becomes worse, from 0.89 to 0.87. The accuracy of the classification with the Random Forest improves from 0.95 to 0.96. The classification with the Random Forest and the text elements split by page layout is thus our best prediction for the text classes.

Table 10. Comparison: Accuracy and weighted average values for text class prediction separated for each layout class (Table 9) and for text class prediction without layout class information (Table 7).

		All text classes (Table 7)	**Text classes separated for each layout class (Table 9)**			
			1C	1CM	2C	wgt. avg.
SVC	wgt. avg.	0.88	0,86	0,89	0,84	0.86
	accuracy	0.89	0,87	0,9	0,84	0.87
Random Forest	wgt. avg.	0.94	0,94	0,98	0,96	**0.95**
	accuracy	0.95	0,94	0,98	0,96	**0.96**
Legend: 1C = one-column, 1CM = one-column with marginals, 2C = two column						

5. CONCLUSIONS

With the use of global page information, we can improve the mapping of text elements to a text class for two-column pages and single-column pages with margins. By using the global layout information, texts in legal documents can be extracted more correctly and are thus available for further processing, possibly cleaned of unwanted text classes such as headers and footers.

Our approach is well suited for preprocessing corpora from the legal domain, also if they include documents that have an imbalance in the number of double-column and single-column pages. The size of the corpus does not need to be as large as when using CCNs, but our method still achieves good results. Especially legal contract documents often contain single-column and double-column pages and the number of columns must be identified to ensure an accurate extraction of the text and to maintain the reading flow.

REFERENCES

[1] Gao, L., Tang, Z., Lin, X., Liu, Y., Qiu, R., Wang, Y. (2011) "Structure Extraction from PDF-based Book Documents" In: Proceedings of the 11th Annual International ACM/IEEE Joint Conference on Digital Libraries. pp. 11-20. JCDL '11, ACM, New York, NY, USA. https://doi.org/10.1145/1998076.1998079

[2] Giguet, E. &Lejeune, G. (2021) "Daniel at the FinSBD-2 task: Extracting list and sentence boundaries from PDF documents, a model-driven approach to PDF document analysis" In: Proceedings of the Second Workshop on Financial Technology and Natural Language Processing. pp. 67-74. - , https://www.aclweb.org/anthology/2020.finnlp-1.11

[3] Ramakrishnan, C., Patnia, A., Hovy, E., Burns, G.A. (2012) "Layout-aware text extraction from full-text PDF of scientific articles" Source Code for Biology and Medicine 7(1), 7 https://doi.org/10.1186/1751-0473-7-7

[4] Dejean, H. & Meunier, J.L. (2006) "A System for Converting PDF Documents into Structured XML Format" In: Document Analysis Systems VII. pp. 129, 140. Lecture Notes in Computer Science, Springer, Berlin, Heidelberg. https://doi.org/10.1007/11669487 12

[5] Klamp, S., Granitzer, M., Jack, K., Kern, R. (2014) "Unsupervised document structure analysis of digital scientific articles" International Journal on Digital Libraries 14 (3- 4), 83-99 https://doi.org/10.1007/s00799-014-0115-1

[6] Klamp, S. & Kern, R. (2016) "Reconstructing the Logical Structure of a Scientific Publication Using Machine Learning" In: Semantic Web Challenges. pp. 255-268. Communications in Computer and Information Science, Springer, Cham; https://doi.org/10.1007/978-3-319-46565-4

[7] Harmata, S., Hofer-Schmitz, K., Nguyen, P.H., Quix, C., Bakiu, B. (2017) "Layout-Aware Semi-automatic Information Extraction for Pharmaceutical Documents" In: Data Integration in the Life Sciences. pp. 71-85. Lecture Notes in Computer Science, Springer, Cham https://doi.org/10.1007/978-3-319-69751-2 8

[8] Namboodiri, A.M. & Jain, A.K. (2007) "Document structure and layout analysis" In: Chaud- huri, B.B. (ed.) Digital Document Processing, pp. 29-48. Springer London. https://doi.org/10.1007/978-1-84628-726-8 2, series Title: Advances in Pattern Recognition

[9] Nojoumian, M. & Lethbridge, T.C. (2011) "Reengineering PDF-based Documents Targeting Complex Software Specifications" Int. J. Knowl. Web Intell. 2(4), 292-319https://doi.org/10.1504/IJKWI.2011.045165

[10] Wyner, A., Mochales-Palau, R., Moens, M.F., Milward, D. (2010) "Approaches to Text Mining Arguments from Legal Cases" In: Semantic Processing of Legal Texts, pp. 60-79. Lecture Notes in Computer Science, Springer, Berlin, Heidelberg (2010). https://doi.org/10.1007/978-3-642-12837-0 4

[11] Chieze, E., Farzindar, A., Lapalme, G. (2010) "An Automatic System for Summarization and Information Extraction of Legal Information" In: Semantic Processing of Legal Texts, p. 216-234. Lecture Notes in Computer Science, Springer, Berlin, Heidelberg https://doi.org/10.1007/978-3-642-12837-0 12

[12] Lin, X. (2003) “Header and Footer Extraction by Page-Association” In: Document Recognition and Retrieval X. vol. 5010, pp. 164-172. International Society for Optics and Photonics https://doi.org/10.1117/12.472833
[13] Enendu, S., Scholtes, J., Smeets, J., Hiemstra, D., Theune, M. (2019) “Predicting semantic labels of text regions in heterogeneous document images” In: 15th Conference on Natural Language Processing, KONVENS 2019: Bridging the gap between NLP and human under- standing
[14] Meng Ling, Jian Chen, Torsten Moller, P. Isenberg, T. Isenberg, M. Sedlmair, R. Laramee, Han-Wei Shen, Jian Wu, and C. Lee Giles. (2021) “Document domain randomization for deep learning document layout extraction” ArXiv,abs/2105.14931.
[15] Adam W. Harley, Alex Ufkes, and K. Derpanis. (2015) “Evaluation of deep convolutional nets for document image classification and retrieval” In: 13th International Conference on Document Analysis and Recognition (ICDAR), pages 991–995.
[16] Forczmański P., Smoliński A., Nowosielski A., Małecki K. (2020) “Segmentation of Scanned Documents Using Deep-Learning Approach” In: Burduk R., Kurzynski M., Wozniak M. (eds) Progress in Computer Recognition Systems. Advances in Intelligent Systems and Computing, vol 977, pp 141-152. Springer, Cham
[17] Karen Simonyan& Andrew Zisserman. (2015) “Very deep convolutional networks for large-scale image recognition” In: 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings
[18] PDFminer/pdfminer.six (2021), https://github.com/pdfminer/pdfminer.six, original-date: 2014-08-29T14:04:53Z

APPENDIX

A.1 Sources for documents in the corpus

1. City Administration Hamburg:
http://suche.transparenz.hamburg.de/dataset?q=vertrag&esq_title=&check_all_
2. City Administration Bremen: https://www.transparenz.bremen.de, Keyword: Vertrag
3. General terms and conditions: Individually researched on websites. All individual links of the sources were saved.

A.2. Sources for data sets of used corpora

The compiled and used corpora, are available on our project page, under: http://textmining.wp.hs-hannover.de/juver.html

AUTHORS

Frieda Josi M.A. Research assistant at the University of Applied Sciences and Arts Hanover and doctoral candidate at the University of Hildesheim, Faculty of Linguistics and Information Science.

Prof. Dr. Christian Wartena Hanover University of Applied Sciences and Arts, Language and Knowledge Processing at the department of Information and Communication.

Prof. Dr. Ulrich Heid University of Hildesheim, Computational Linguistics and Language Technology, Faculty of Linguistics and Information Science**.**

© 2022 By AIRCC Publishing Corporation. This article is published under the Creative Commons Attribution (CC BY) license.

RESEARCH ON ANTI-INTERFERENCE OF DOUBLE HOP WIRELESS POWERED COMMUNICATION NETWORKS BASED ON TIME REVERSAL

Wei Liu, Fang Wei Li, Jun Zhou Xiong and Ming Yue Wang

School of Communication and Information Engineering, Chongqing University of Posts and Telecommunications, Chongqing, China

ABSTRACT

To solve the problems of dual near and far in wireless powered communication network(WPCN) and the interference in the process of information transmission, a resource allocation method based on time inversal (TR) for WPCN is proposed. An optimization problem to maximize the minimum network throughput is constructed by jointly optimizing the transmission time of each phase of the network, the transmission power of the hybrid access point (HAP) and the transmission power of the relay. Since the constructed problem is non-convex, this paper converts the non-convex problem into an equivalent convex problem by introducing relaxation variables and auxiliary variables, and further divides the convex problem into two sub-problems to obtain the solution of the original problem. Finally, the simulation results show that the proposed resource allocation scheme can alleviate the dual distance and interference effectively, so as to obtain a higher total system throughput.

KEYWORDS

Wireless Powered Communication Network, Ttime Reversal, Jointly Optimizing

1. INTRODUCTION

In the Internet of Things era, different terminals need to be deployed due to different service types. As the number of terminals increases exponentially, massive sensor devices need to be deployed in the network to realize intelligent interconnection of everything. However, deploying a large number of sensor devices would face energy constraints, and regularly manually replacing or recharging batteries would be inconvenient, dangerous, or even impossible and costly[1]. The energy limitation problem has been alleviated by the development of energy harvesting technology, which uses the collected energy to charge network nodes [2], It is theoretically possible to provide equipment with permanent life Among many energy collection methods, RF signals are more suitable for small size and low power wireless node devices due to their predictable, controllable and stable nature [3], Therefore, RF energy collection is a feasible energy collection method. One of the most attractive research categories in the field of RF energy collection is the emerging Wireless Powered Communication Networks (WPCN), The network device can charge the battery by collecting energy from the surrounding environment or RF signals by dedicated energy transmitters [4], then uses the collected energy to fulfill its communication needs. Some scholars believe that WPCN is an integral part of the Internet of Things (IoT) [5], In this network, WPCN provide power to energy-constrained devices and are a

David C. Wyld et al. (Eds): NATP, ACSTY, CCCIOT, MLSC, ITCSS - 2022
pp. 31-49, 2022. CS & IT - CSCP 2022 DOI: 10.5121/csit.2022.120103

new opportunity for the IoT. Therefore, WPCN system has attracted more and more attention [6].

During the downlink Wireless Enengy Trealation (WET) of WPCN, users who are farther away from Hybrid Access Point (HAP) get less energy than those who are closer, But in its uplink Wireless Inforemation Trealation (WIT) it must be transmitted with higher power, Call this phenomenon the dual near and far problem. In view of the dual near and far problem, some scholars have studied WET or WIT in the relay assisted WPCN, which is an effective way to solve the dual near and far problem and improve network performance. In [7], a multi-user relay WPCN is considered, which is based on the relay protocol of " charge than forward ", in which the single-antenna and half-duplex hybrid relay node first supplies power to the source node, and then the source and the relay can use the energy obtained from the hybrid relay node to upload the information to the destination. In [8] considers different WPCN scenarios composed of three nodes, in which multi-antenna relay not only WET the source node of a single antenna, but also cooperatively forward the information from the source node to the destination of the single antenna. In [9] considers a double-hop WPCN, in which the two-way communication between the user and the access point is assisted by a relay. The piecewise linear energy collection model of user and relay studies the problem of maximizing the total throughput of amplified and decoded relay mode, in which the energy transmitter is multiple antennas, the information access point and the user each have one antenna, and the relay has two antennas. Multi-antenna HAP is capable of downlink energy beamforming, which enables HAP to deliver more power to a specific user. In [9] addition to the multi-antenna system, the access point also adopts full-duplex communication, which improves the total throughput in full-duplex WPCN compared with half-duplex WPCN. Therefore, relay-assisted communication is an essential element in the IoT world. It expands coverage, improves availability, enhances reliability, increases network capacity, and reduces power consumption.

Another factor that affects throughput in WPCN is interference. In actual wireless communication scenarios, the total throughput of all users can be maximized as the number of terminals increases, but interference will increase [10]. The average interference of all HAP in multi-terminal WPCN is considered in [10], and the time allocation problem of achieving maximum and minimum throughput of all users in multi-cell WPCN is studied. A multi-antenna decoding and forwarding WPCN system is considered in [11]. The block descent method is used to jointly optimize transmit beamforming and self-interference cancellation in the process of combining and receiving. After repeated iterations, the optimization result is close to the optimal solution. In recent years, with the proposal and development of time reversal (TR) technology, research shows that TR can effectively alleviate the interference in the communication system. In [12] combines TR with multi-input and multi-output (MIMO) technology, and adopts the method of parameter research to quantitatively study the relationship between the performance of MIMO communication based on TR and the change of multi-path environment. The probability density function and cumulative distribution function of signal-to-noise ratio at the receiver in TR communication system are derived in [13]. Based on the derived probability density function, the traversal capacity, interrupt probability and bit error rate of binary phase shift keying are also obtained. The above research shows that in TR system, the inter-symbol interference decreases with the increase of the up-sampling factor, and the diversity gain increases with the increase of the number of paths. In [14] proposed that TR has tunneling effect in cloud wireless access network, which can effectively restrain the interference in the uplink transmission process and improve the system throughput. Broadband cooperative spectrum sensing based on FDMA technology is considered in [15]. TR generalized linear, TR maximum ratio combining and improved TR maximum ratio combining rules are proposed for decision fusion. The effectiveness of the proposed rules against inter-symbol interference (ISI) and inter-carrier interference(IUI) is tested according to the function of signal-to-interference-noise ratio(SINR).

The research shows that compared with the traditional fusion rules, the performance of the fusion rules based on TR is greatly improved. Through the above research, we can know that TR can effectively alleviate the interference problem in the communication system, thus improving the performance of the system. TR makes use of the advantage of channel reciprocal focusing signal to suppress the transmission interference of uplink. Specifically, it uses the degree of freedom provided by the environment, that is, rich multipath, to use signature waveform design technology to combat interference. The basic idea of signature waveform design is to adjust the amplitude and phase of each tap of the signature waveform based on channel information so that the signal at the receiver can retain most of the useful signals and suppress interference as much as possible.

Through the review of the above work, we can see that the existing research work on WPCN is insufficient, as follows. First of all, in multi-terminal node networks, the nodes are randomly distributed, and it is the basic requirement of network application to ensure users' fair access to communication resources. Therefore, for WPCN with multiple communication nodes, the fairness among users can be ensured by calculating the minimum throughput maximization of terminal nodes. Secondly, resources are allocated only from the perspective of time. In the wireless power supply communication network, both time and power will affect the total throughput of the system, so it is necessary to jointly optimize to improve the network performance. Then, in the Internet of things, a large number of nodes, interference will also reduce network performance. Therefore, for the network with thousands of nodes, it is necessary to consider the combination of new technologies to reduce interference in the network and improve performance. Inspired by the above factors, in order to alleviate the problem of double distance to ensure user fairness and interference, this paper studies the resource allocation algorithm of dual-hop wireless communication network based on time reversal. The main work is as follows：

1. A multi-antenna double-hop WPCN model based on time reversal is proposed. Considering the constraints of time slot resources, HAP transmission power and relay node transmission power, the resource allocation problem of maximizing throughput is established.
2. Because the established optimization problem needs to jointly optimize time slots and power resources, resulting in the coupling of multiple variables, the proposed optimization problem is a non-convex optimization problem, so it is impossible to directly use the existing convex optimization tools to obtain the optimal solution. Therefore, this paper first transforms it into an equivalent convex problem by introducing relaxation variables and auxiliary variables, and then decomposes the optimal convex problem into two single-variable sub-problems, on this basis, the solution of the problem is obtained by using convex optimization theory.
3. Simulation results show that this algorithm can effectively alleviate the impact of double distance and interference problems. Compared with other schemes, it is proved that this algorithm can achieve better throughput performance.

The rest of this paper is organized as follows. A dual-hop WPCN model is introduced in section 2. Section 3 investigate the minimum throughput maximization problem, respectively. Section 4 evaluates the performance of the presented algorithms by conducting numerical simulations and section 5 concludes the paper.

2. SYSTEM MODEL

As shown in Figure 1. , this paper considers a dual-hop WPCN consisting of one HAP, one energy constrained relay node and K terminals. The HAP and relay nodes are equipped with N antennas, and the terminal is equipped with a single antenna. The transmission process of the

whole system is completed by three processes: downlink WET, time reversal detection and uplink WIT. As shown in Figure 1. (a), the HAP transmits energy to the relay node and the terminal node in the downlink WET phase, and the relay node transmits detection signal detection channel state information to different terminals in the time reversal detection phase. The channel gains between the HAP to the relay node and to the terminal node k are defined as H, h_{hk} ; The uplink WIT process is divided into two stages, as shown in Figure 1. (b), After the terminal node records the detected status information and performs TR operation, in the first stage of uplink WIT, the terminal node uses the energy stored in the downlink WET stage to send the information to the relay node, and in the second stage of uplink WIT, the relay centrally processes the received signal and forwards it to the HAP. Define the channel gain from the relay node to the terminal node k as h_{rk} .

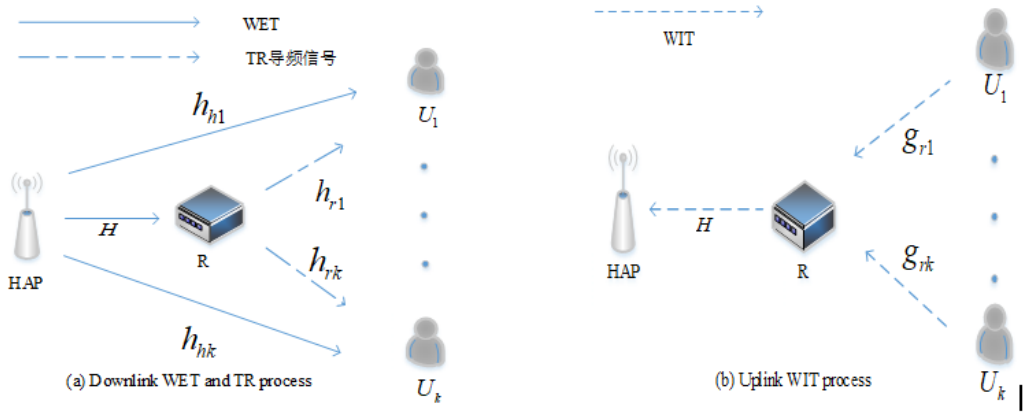

Figure 1. Dual hop WPCN system model

It is assumed that the relay and all terminal nodes are equipped with energy storage devices such as rechargeable batteries or super capacitors to store energy. The relay protocol of "charge than forward" is considered to coordinate energy and information transmission, in which HAP with constant energy supply firstly downlink transmits energy signals to charge relay and terminal nodes, and then relay and terminal nodes collect and store energy for uplink WIT. It is assumed that all channels between HAP, relay and terminal are reciprocal and experience slow, independent and flat Rayleigh fading, so that the channel gain remains unchanged within each transmission block T, but varies independently between different transmission blocks. In the process of uplink WIT, the distance between the terminal and the HAP is very long or the signal is seriously attenuated, and there is no direct link between the HAP and the terminal, so relay assistance is needed to forward the information. Without losing generality, in the rest of this paper, assume that T = 1.

Considering the frame structure of a dual-hop WPCN with K terminals shown in Figure 2., one frame is divided into four time slots. In the first time duration τ_0, HAP broadcasts an energy signal to the relay and all terminal nodes in order to provide energy for their upcoming uplink transmission tasks. In the second time duration τ_{tr}, Firstly, the relay node sends detection signals to different terminals to obtain discrete channel impulse response; Then, each terminal records the discrete channel impulse response and inverts it in the time domain. After receiving the energy signal to supplement the battery, the terminal node uses the collected energy in the third time duration τ_1 to convolute the information signal to be transmitted with the inverted discrete channel impulse response, and then sends it to the multi antenna relay node through

SDMA. In the fourth time duration τ_2, The relay node uses the energy stored in the downlink WET phase to forward the decoded information to the HAP.

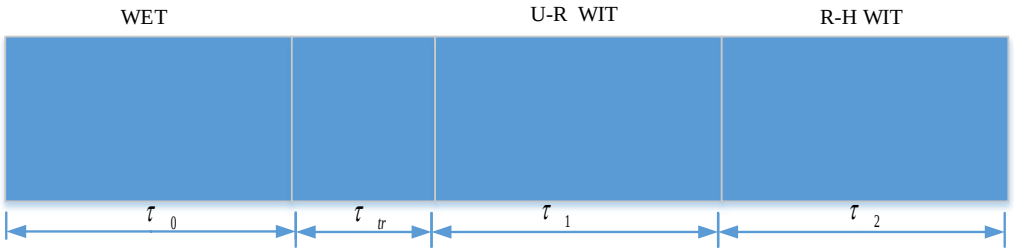

Figure 2. Frame structure of double hop WPCN network model

2.1. Downlink WET Phase

In the WET phase, HAP sends energy beams to the free battery multi antenna relay and all users at the same time. Let us defined w_l as the energy beamforming weight transmitted by the l antenna. The energy beamforming vector is defined as $w=(w_1,...,w_N)^T$. The transmission power of antenna l is $|w_l|^2$. The following limits are imposed on the transmit power of each antenna：

$$|w_l|^2 \le P_{\max} \tag{1}$$

$l=1,...,N$, $P_{\max}$ is the maximum transmit power of each antenna, $|\ |$ representative length. The channel gain from HAP to battery free relay is expressed as $H \in \mathbf{C}^{N\times N}$, The channel gain from HAP to terminal K is defined as $h_{hk} \in \mathbf{C}^{N\times 1}$, $1 \le k \le K$. The signal received by the relay node and terminal is：

$$y_r = H^H w + n_r \tag{2}$$

$$y_k = h_{hk}^H w + n_k \tag{3}$$

where $n_r \sim CN(\mu,\sigma_r^2 I)$ is $N\times 1$ Additive White Gaussian Noise(AWGN), $n_k \sim N(0,\sigma_k^2)$ is AWGN of terminal U_k, Superscript H denotes conjugate transpose. The power obtained by the free battery relay in time slot τ_0 can be expressed as：

$$r_r = \left|y_r^2\right| = \left|H^H w\right|^2 = tr\left(w^H H H^H w\right) = tr\left(H H^H w w^H\right) = tr\left(G_r S\right) \tag{4}$$

where $G_r = HH^H$、$S = ww^H$. The energy obtained by the free battery relay in the τ_0 slot is：

$$E_{Hr} = \varepsilon \tau_0 tr\left(G_r S\right) \tag{5}$$

Where $\varepsilon \in [0,1]$ is energy conversion efficiency, Similarly, the power obtained by the terminal in the τ_0 slot can be expressed as：

$$r_k = \left| y_k^2 \right| = \left| h_{hk}^H w \right|^2 = tr\left(w^H h_{hk} h_{hk}^H w \right) = tr\left(h_{hk} h_{hk}^H w w^H \right) = tr\left(G_k S \right) \tag{6}$$

Where $G_k = h_{hk} h_{hk}^H$

Similarly, the energy obtained by terminal U_k in τ_0 slot can be expressed as：

$$E_{Hk} = \varepsilon \tau_0 tr\left(G_k S \right) \tag{7}$$

Then, the maximum transmission power of the relay node and terminal U_k in the information transmission stage can be expressed as：

$$P_r = \frac{\varepsilon \tau_0 tr\left(G_r S \right)}{\tau_2}, \quad P_k = \frac{\varepsilon \tau_0 tr\left(G_k S \right)}{\tau_1} \tag{8}$$

2.2. TR Phase

In WPCN network, on the one hand, when multiple terminals transmit information uplink in SDMA mode, IUI will occur, resulting in system performance degradation. On the other hand, there are many terminals and their locations are randomly distributed. Pilot pollution will also affect the system performance. Rich multipath will produce inter symbol interference (ISI). Because the unique space-time focusing characteristic of TR technology can alleviate the channel delay and reduce the interference, this paper introduces TR technology and uses the unique space-time focusing characteristic of TR in multipath environment to counter IUI and ISI. Time focusing effect concentrates most of the useful signal energy of each symbol in a short time interval, which effectively suppresses ISI. Spatial focusing effect is to collect the signal energy at the expected position and reduce the leakage to other positions, resulting in the reduction of transmission power consumption and the co-channel interference to other positions.

The relay node and the terminal node receive the energy signal, the relay sends the detection signal to the terminal. The terminal inverts the discrete channel impulse response obtained by detection in the time domain to obtain the equivalent channel response, and normalizes it as a channel signature waveform. Each terminal transmits information through channels with different channel signatures, and effectively suppresses interference by adjusting the parameters of the channel signatures. The normalized channel signature waveform after TR processing is：

$$g_{rk}(m) = \frac{h_{rk}^*(L-1-m)}{\sqrt{\sum_{m=0}^{L-1} \left| h_{rk}(m) \right|^2}} \tag{9}$$

Where $*$ represents conjugation, L is the number of multipath, $m = 0,...,L-1$.

2.3. Uplink WIT Phase

In time slot τ_1, all terminal nodes use the energy acquired in time slot τ_0 to send information to relay nodes in SDMA, as shown in fig.1. According to transmission characteristics, the channel from the terminal to the relay node is regarded as a multi-access channel(MAC), so its capacity can be expressed by the following formula：

$$C_{U-R} = \log_2 \left| I + \sum_{k=1}^{K} SINR \right| \tag{10}$$

Where formula I above is the identity matrix, $SINR$ is：

$$SINR = \frac{P_k g_{rk} R_k^0 g_{rk}^H}{P_k g_{rk} \hat{R}_k g_{rk}^H + \sum_{i \neq k} P_i g_{rk} R_i g_{rk}^H + \sigma^2 I} \tag{11}$$

where $p_k = [P_1, ..., P_K]^T$ is the uplink transmission power vector of the terminal, $R_k^0 = A_k^L A_k^{(L)H}$、$R_i = A_i A^H$, A_i is the channel Topplitz matrix of the terminal node i to relay node, A_k^L represents the L row vector of A_k, $\hat{R}_k = A_k A_k^H - R_k^0$. In (11) $P_k g_{rk} \hat{R}_k g_{rk}^H$ and $\sum_{i \neq k} P_i g_{rk} R_i g_{rk}^H$ represents ISI and IUI. To simplify the calculation, it is assumed that the noise variances of different channels are the same.

In the time slot τ_2, The relay node decodes the message received during τ_1 and sends it to HAP through the MIMO channel. The singular value decomposition of the MIMO channel can be described as $H = UDV^H$, where U and V^H is the $N \times N$ unitary matrix. The diagonal elements of $D \in \mathbf{R}^{N \times N}$ are the eigen values of channel H, denoted by ψ_i. Its capacity is expressed as:

$$C_{R-H} = \sum_{i=1}^{N} \log_2 \left(1 + \frac{P_{ri} \psi_i}{\sigma_h^2} \right) \tag{12}$$

Where σ_h^2 denoes noise power, N denotes the number of Gaussian channels, P_{ri} is the transmission power of the ith equivalent Gaussian channel. P_{ri} can be calculated by water-filling algorithm[16]. The calculation formula is： $P_{ri} = \left[\frac{1}{\gamma \times In2} - \frac{\sigma_h^2}{\psi_i} \right]^+$

Proof: please see Appendix A.

According to the above, the power has the following limitations $\sum_{i=1}^{N} P_{ri} = P_r$.

The throughput of the first hop and the second hop of the model proposed in this paper has been given, so the total throughput of the system can be expressed as:

$$R_{sum} = \min\left(\tau_1 C_{U-R}, \tau_2 C_{R-H}\right) \tag{13}$$

The objective of this paper is to ensure the minimum rate of each terminal and maximize the total throughput of the system through joint optimization of time allocation and HAP downlink transmit energy beamforming matrix S. The minimum throughput maximization problem is shown as follows：

$$OP_1: \quad \max_{P_{ri}, P_k, \tau, w} \quad R_{sum} \tag{14}$$

$$s.t. \quad C_1: tr(S) \le (K+1)P_{\max} \tag{14a}$$

$$C_2: \tau_i \ge 0, \quad \tau_{tr} \ge 0, \quad i = 0,1,2 \tag{14b}$$

$$C_3: \tau_0 + \tau_{tr} + \tau_1 + \tau_2 = 1 \tag{14c}$$

Where C_1 is maximum transmit Power constraint of HAP, C_2 and C_3 is time Duration constraint. This means that the sum of the duration of energy and information transmission must not exceed the length of the transmission block assumed to be 1. OP_1 is non-convex, making it difficult to solve the problem, which will be solved below.

3. OPTIMIZATION PROBLEM SOLVING

In this part, we study and solve problem OP_1. OP_1 is a non-convex problem of strongly coupled variables τ and w, which can not be solved in its original form. Therefore, in order to make the problem easy to deal with, this paper introduces new variables and uses the characteristics of sub-problems to solve. OP_1 is equivalent to the following questions:

$$OP_2: \quad \max_{w_i, \bar{R}, \tau} \quad \bar{R} \tag{15}$$

$$s.t. \quad C_1, \quad C_2, \quad C_3$$

$$C_4: \bar{R} \le \tau_1 \log_2\left|I + \sum_{k=1}^{K} SINR\right| \tag{15a}$$

$$C_5: \bar{R} \le \tau_2 \sum_{i=1}^{N} \log_2\left(1 + \frac{P_{ri}\psi_i}{\sigma_h^2}\right) \tag{15b}$$

Where $\bar{R}$ is an introduced relaxation variable, The constraints of C_4 and C_5 are associated with the objective function expression. Obviously, C_{U-R} and C_{R-H} are joint concave functions in optimization variables, because they are the perspective of concave functions, and the perspective operator maintains concavity[17]. The objective function of OP_2 and the constraints in C_1、C_2, and C_3 are affine functions. So for the new variables w、$\bar{R}$ and τ, problem OP_2 is joint concave, and the problem can be effectively solved by a general convex optimization tool, such as CVX [18]. For the solution proposed above, because it contains matrix

calculation, the computational complexity is very high. Therefore, this paper divides problem OP_2 into Energy beamforming sub problem and time allocation sub problem to solve.

Given the time allocation τ, the optimization method of energy beamforming is as follows:

In the proposed scheme, a beam splitting algorithm is used to separate the energy beams to provide energy to multiple nodes at the same time. Multi beam technology uses beamforming weight vector to realize Pareto optimality in received power domain. Pareto boundary $\left(R^{PF}\right)$ is defined as the set of all Pareto optimals in R. Expressed as follows:

$$R^{PF} = \left\{ x \in R \middle| not\, r \in R\, when\, x \prec r \right\}$$

where"$\prec$" represents element inequality.

In this paper, a point on the Pareto boundary of R is obtained by finding the maximum weighted sum of the received power vector in R. The optimization problem of the received power vector of the relay node and each terminal node is to maximize $\alpha^T x$ under the condition of $x \in R$, where $\alpha = (\alpha_0, ..., \alpha_K)^T$ is the received power weight vector. The received power weight vector α should satisfy $\alpha_k \geq 0$ and $1^T \alpha = 1$. This optimization problem is equivalent to the following optimization problem:

$$\max : \alpha_0 \left\| H^H w \right\|_\infty^2 + \sum_{k=1}^K \alpha_k \left| h_{hk}^H w \right|^2 \tag{16}$$

$$s.t. \quad |w_k|^2 \leq P_{\max} \quad k = 0, ..., K \tag{16a}$$

It is difficult to directly solve the above optimization problem (16). Therefore, by relaxing the limit on the transmission power of each antenna, the total transmission power is limited. That is, $\sum_{k=0}^K |w_k|^2 \leq (K+1) P_{\max}$. Than, the optimization problem (16) is transformed into:

$$\max : \alpha_0 tr(G_r S) + \sum_{i=1}^K \alpha_k K tr(G_k S) \tag{17}$$

$$s.t. \quad tr(S) \leq (K+1) P_{\max} \tag{17a}$$

To solve the optimization problem (17), define:

$$V(\alpha) = \alpha_0 G_r + \sum_{i=1}^K \alpha_k G_k \tag{18}$$

The eigenvalues of $V(\alpha)$ are decomposed into:

$$V(\alpha) = U(\alpha)^H Z(\alpha) U(\alpha) \tag{19}$$

Where $U(\alpha)$ is unitary matrix $U(\alpha) = (u_1(\alpha), ..., u_N(\alpha))$, $Z(\alpha)$ is diagonal matrix $Z(\alpha) = diag(z_1(\alpha), ..., z_N(\alpha))$. The diagonal elements in $Z(\alpha)$ are arranged in descending order. $z_1(\alpha)$ is the principal eigenvalue. $u_1(\alpha)$ is the eigenvector corresponding to the principal

eigenvalue $z_1(\alpha)$. Other eigenvectors are the same one-to-one correspondence. Therefore, the solution of (17) can be defined as $w^{op}(\alpha)=\left(w_1^{op}(\alpha),...,w_N^{op}(\alpha)\right)^T$, where：

$$w^{op}(\alpha)=\sqrt{(K+1)P_{\max}}\,u_1(\alpha) \tag{20}$$

Since (20) is the solution of the relaxation optimization problem (17), the transmission power constraint of each antenna in (16) may not be satisfied and the maximum transmission power of each antenna is not fully utilized. Therefore, the transmission power of each antenna is adjusted in the following way to make the transmission power of each antenna $P_{\max}$ ：

$$w_k^{op}(\alpha)=\frac{w_k^{op}(\alpha)}{\left|w_k^{op}(\alpha)\right|}\sqrt{P_{\max}} \tag{21}$$

Where $W^{op}(\alpha)=\left(w_1^{op}(\alpha),...,w_N^{op}(\alpha)\right)^T$ is the optimal beamforming weight vector given α. The optimal beamforming weight vector in (21) is the solution of (16).

For a given energy beamforming vector, the optimization problem of OP_2 can be described as:

$$\begin{aligned} OP_3：&\max_{\bar{R},\tau}\ \bar{R} \\ &s.t.\ \ C_2,\ \ C_3,\ \ C_4,\ \ C_5 \end{aligned} \tag{22}$$

It is worth noting that OP_3 is a convex optimization problem. In this paper, we fix τ_0 to find the optimal time allocation coefficient. In the case of fixed value τ_0 , the partial Lagrange of OP_3 can be written as:

$$\begin{aligned} &L\left(\bar{R},\tau_0,\tau_{tr},\tau_1,\tau_2\lambda,\mu,\nu\right)=\bar{R}-\lambda\left(\tau_0+\tau_{tr}+\tau_1+\tau_2-1\right) \\ &-\mu\left(\bar{R}-\tau_1\log_2\left|I+\sum_{k=1}^{K}SINR\right|\right)-\nu\left(\bar{R}-\tau_2\sum_{i=1}^{N}\log_2\left(1+\frac{P_{ri}\psi_i}{\sigma_h^2}\right)\right) \end{aligned} \tag{23}$$

Since OP_3 is convex and satisfies Slater condition[17], strong duality is established, and karush Kuhn Tucker (KKT) condition is a necessary and sufficient condition for the optimal solution. In order to obtain the optimal solution of the above dual function, it is necessary to maximize the Lagrange variable.

By considering the KKT condition related to $\bar{R}$, the following lemma is obtained. When finding the optimal throughput, the variable $\bar{R}$ and the dual variable must meet the optimal conditions.

Lemma 1：The optimal dual variables μ^{op} and ν^{op} must satisfy $0<\mu^{op}<1$ and $0<\nu^{op}<1$

Proof: Please see Appendix B

According to the result of lemma 1, the following proposition can be deduced: When the beamforming matrix is given, the throughput of the first hop from the terminal to the relay node and the throughput of the second hop from the relay node to the HAP must be equal. Namely:

$$\tau_1^{op} C_{U-R}\left(\tau^{op}\right) = \tau_2^{op} C_{R-H}\left(\tau^{op}\right) \qquad (24)$$

From lemma 1, dual variables must satisfy $0<\mu^{op}<1$ and $0<\nu^{op}<1$. From the KKT complementary relaxation condition in the proof process of lemma 1, it can be easily proved that the optimal value should be satisfied $\bar{R}^{op} = \tau_1^{op} C_{U-R}\left(\tau^{op}\right) = \tau_2^{op} C_{R-H}\left(\tau^{op}\right)$. Therefore, when the throughput of the first hop is equal to that of the second hop, the maximum total throughput can be obtained. And the optimal time allocation algorithm also needs the above conditions.

Lemma 2： When the energy beamforming matrix S is given, and when τ_1 and τ_2 satisfy $\tau_1^{op} C_{U-R}\left(\tau^{op}\right) = \tau_2^{op} C_{R-H}\left(\tau^{op}\right)$, the throughput of the system is a concave function of τ_0.

Proof: Please see Appendix C

When the beamforming energy matrix and the time τ_0 of downlink transmission energy are given, τ_1 and τ_2 can be determined according to lemma 2, When $\tau_1^{op} C_{U-R}\left(\tau^{op}\right) = \tau_2^{op} C_{R-H}\left(\tau^{op}\right)$ is satisfied, the throughput of the system is a concave function of τ_0, and then the optimal solution is obtained according to the golden section algorithm.

Table 1. Time allocation optimization algorithm

Algorithm 1 Time allocation algorithm steps
1. Calculate τ_{tr} with given d and v_0
2.Input accuracy ε, $\tau_0^{low}=0$, $\tau_0^{up}=1$。
3.While $\left\|\tau_0^{up}-\tau_0^{low}\right\| \geq \varepsilon$ do
4. $\tau_0^{op1} = \tau_0^{up} - 0.618\left(\tau_0^{up}-\tau_0^{low}\right)$
5. $\tau_0^{op2} = \tau_0^{low} + 0.618\left(\tau_0^{up}-\tau_0^{low}\right)$
6. Calculate $R_{sum}\left(\tau_0^{op1}\right)$和 $R_{sum}\left(\tau_0^{op2}\right)$ by soving (24)
7.If $R_{sum}\left(\tau_0^{op1}\right) \leq R_{sum}\left(\tau_0^{op2}\right)$, $\tau_0^{low}=\tau_0^{op1}$. Else $\tau_0^{up}=\tau_0^{op2}$
8.End while
9. $\tau_0^{op} = \dfrac{\tau_0^{op1}+\tau_0^{op2}}{2}$,obtain τ_1^{op} and τ_2^{op} by soving (24)
10.Output R_{sum}^{op}

4. SIMULATION RESULTS

For the multi antenna relay assisted WPCN in this paper, both HAP and free battery relay are equipped with N=6 antennas. Assume the total bandwidth is 10MHz, All channels experience quasi-static flat Rayleigh fading, and the channel power gain is modeled as a $10^{-3}\left(\dfrac{d_{ij}}{d_0}\right)^{-\alpha}$, where d_{ij} is the distance between relay node i and terminal j, set d_0 as a reference distance of 1m and $\alpha=3$ as the path loss factor. The maximum transmit power of HAP is set as

$P_{\max} = 40dBm$, The energy conversion efficiency is set as $\varepsilon = 1$. This paper considers that in a two-dimensional Euclidean space with meter as unit, Place HAP in $(-2,0)$, The relay node is placed in the circle which center is $(0,0)$、The terminal node is uniformly placed in a circle which center is $(2,0)$ and a diameter of 1m. Unless otherwise stated, the channel model and simulation parameters are set the same. In this paper, the influence of this scheme on system performance under different conditions is analyzed and compared with the traditional scheme.

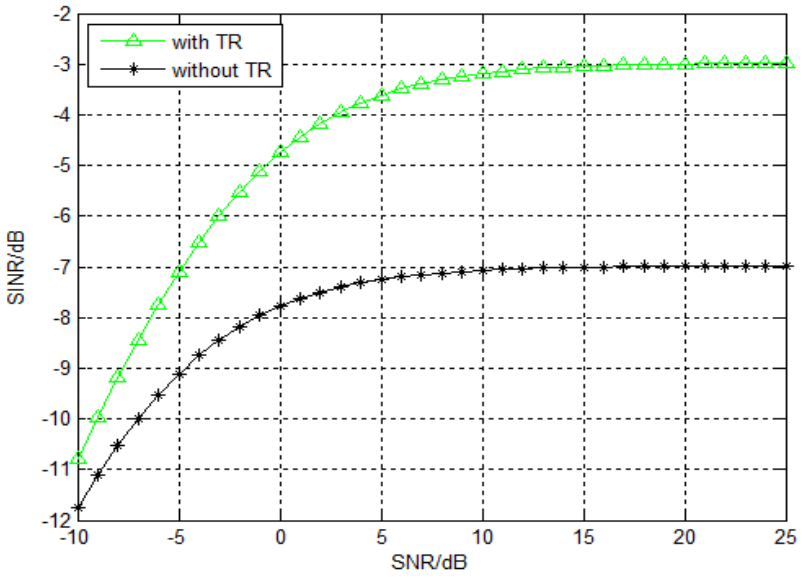

Figure 3. The relationship between SNR and SINR

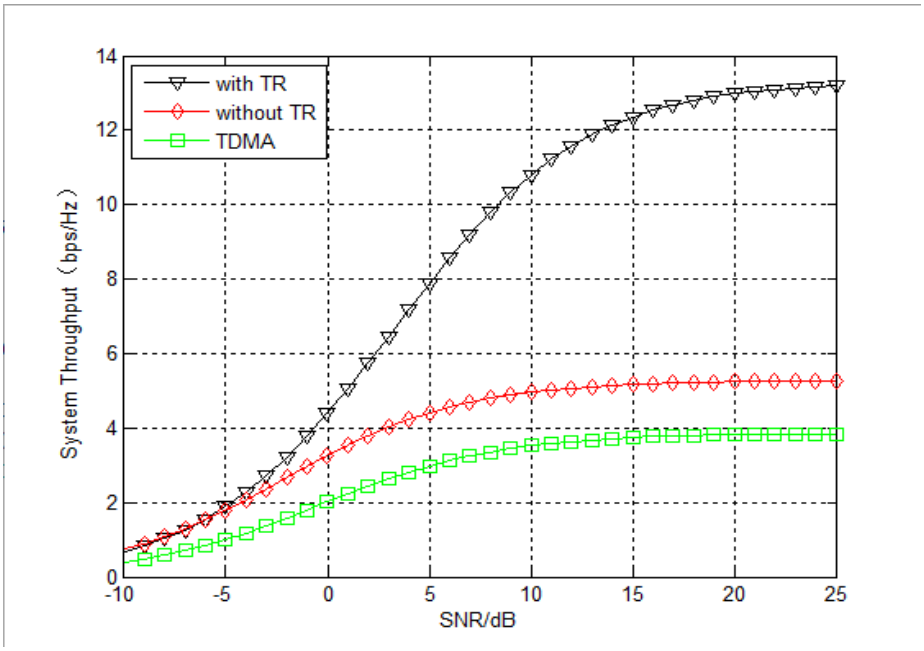

Figure 4. A comparison diagram of the throughput of different solutions

Figure 3. shows the comparative analysis of SNR and SNIR with and without TR technology, it can be clearly seen that TR technology can effectively improve the SNIR. This is due to the fact that the space-time focusing of TR will focus at the same time and adaptively at the target point, resulting in higher signal amplitude and lower signal amplitude in other positions. Figure 4 shows the relationship between different SNR and the total throughput of the system under the same conditions. It can be seen from figure 4. that the proposed scheme can effectively improve the system throughput. Compared with the TDMA scheme with single antenna transmission, the system throughput of this scheme is significantly improved. In this paper, HAP and relay nodes are equipped with multi-antennas, and the downlink WET process adopts MIMO system multi-user beamforming technology, and the optimal beamforming power allocation scheme is obtained through research. Compared with the single antenna scheme, with the increase of the number of transmitting antennas, Higher directional gain can be obtained by beamforming in the direction of downlink energy concentrated transmission make up for the energy loss in the transmission process and improve the energy efficiency of WET. Therefore, the terminal node can get more energy for uplink WIT. And compared with the methods that do not use TR technology, TR technology effectively suppresses IUI and ISI in the uplink information transmission process, thus improving the system throughput and stability. In the scheme using TDMA, although there is no IUI, the uplink transmission consumes too much time, so the spectrum efficiency is low, and the relay is not used for secondary forwarding in the uplink WIT process, resulting in long transmission distance and large loss, which makes the system throughput relatively low. So, compared with the traditional time allocation scheme, the scheme proposed in this paper can effectively improve the total throughput of the system and make the system more stable. It can also be observed from Figure 4. that although the total throughput of the system increases with the increase of the SNR, when the SNR is too large, the throughput will reach balance state and will not change significantly with the increase of the SNR. Because with the improvement of SNR, although large capacity signals can be received more efficiently, the transmitted signals will be more interfered at the same time, so the system throughput reaches a balance state.

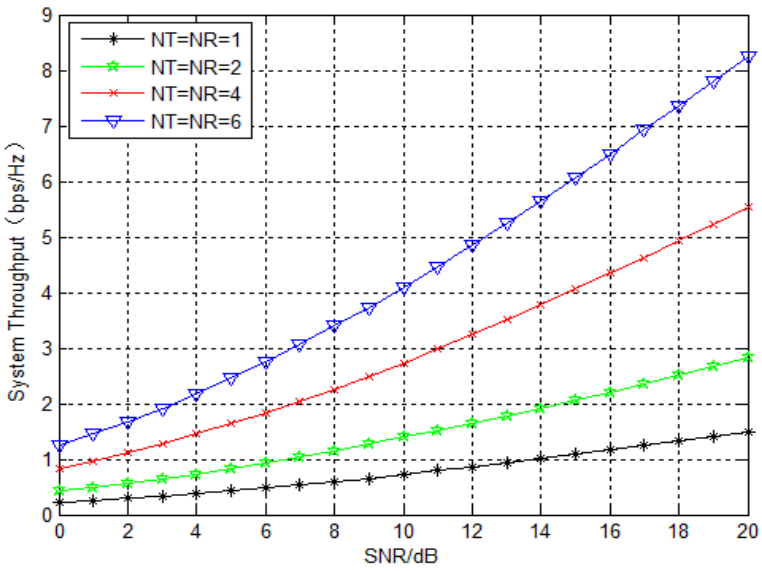

Figure 5. The relationship between total throughput and SNR

Figure 5. depicts the relationship between the total throughput of this paper scheme and the SNR under different antenna configurations of HAP and relay node. Monte Carlo simulation method

is used in the simulation. Number of transmit receive antennas is 1x1、2x2、4x4、6x6, The number of iterations of Monte Carlo simulation is 10000. As can be seen from Fig. 5, in the same case, the throughput increases with the increase of SNR. Compared with a single antenna, the greater the number of antennas, the greater the increase of throughput. It is proved that the application of multi-antenna in WPCN can effectively increase system throughput, and the effectiveness of the proposed scheme is also proved.

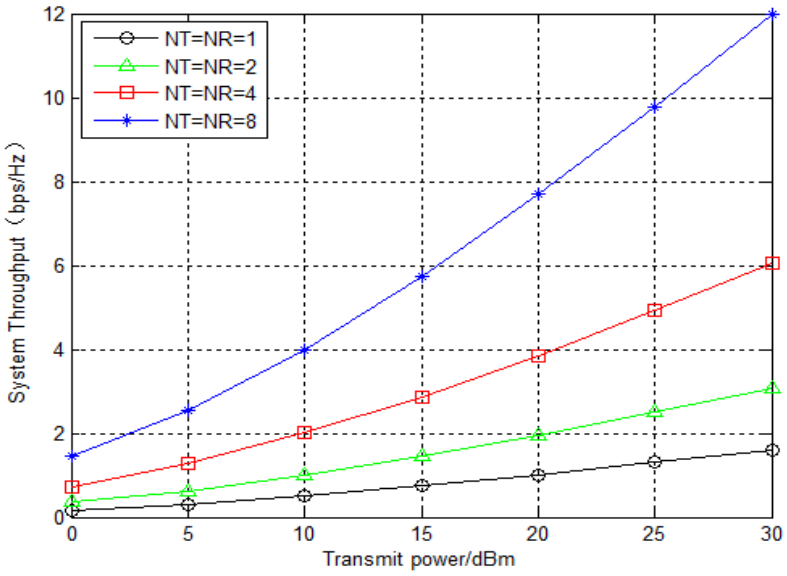

Figure 6. The relationship between total throughput and transmitted power

Figure 6. depicts the relationship between the total throughput of the system and the transmission power under different antenna configurations of HAP and relay nodes. For the scheme in this paper, it can be observed that the throughput increases with the increase of transmission power, and the multi-antenna scheme used in this paper has considerable gain. This is obvious, because by using higher transmission power, devices can collect more energy during WET, thus achieving higher throughput in WIT. It is clear in this figure that the multi-antenna scheme will always bring higher throughput than the single-antenna scheme. Because this paper applies relay to WPCN system, relay divides long-distance transmission into two hop transmission, which can reduce energy loss during transmission. After the first hop is added with TR technology, the space-time focusing characteristics of TR can effectively suppress the interference and information signal energy leakage during transmission, so the system performance can be improved.

5. CONCLUSIONS

For the WPCN, considering the double distance problem of the network itself and the interference will affect the system performance, By introducing relay and combining TR, this paper jointly optimizes the resource allocation problems such as the transmission time of each stage of the network, the transmission power of HAP and the transmission power of relay, in order to maximize the total throughput of the system. According to the model established by the proposed problem, the non-convex problem is transformed into an equivalent convex problem by introducing relaxation variables and auxiliary variables. Then the convex problem is divided into two sub-problems to obtain the suboptimal solution of the original problem. The simulation

results show that the proposed scheme can suppress the interference and improve the system performance on the basis of effectively alleviating the double near far problem. The WPCN system has broad application prospects and can be used in wireless sensor networks. Future research can allocate and optimize resources for multi-antenna HAP and multi-user complex multi-cell environments and vehicle networking environments. The focusing intensity of time reversal technology depends on channel estimation. When there is an error in channel estimation, it will affect the focusing of time reversal. The scheme of this paper assumes that the channel state information is perfect, but in practice, the channel estimation will be affected by many factors in the wireless environment. Therefore, when there is an error in channel estimation, the performance of the system will be affected.

ACKNOWLEDGEMENTS

This work is supported by the Natural Science Foundation of China (NSFC) grant funded by the China government (61771084).

APPENDIX A

PROOF OF CALCULATION FORMULA

$$\max: \ C_{R-H} = \sum_{i=1}^{N} \log_2\left(1+\frac{P_{ri}\psi_i}{\sigma_h^2}\right) \tag{25}$$

$$s.t. \ \ \sum_{i=1}^{N} P_{ri} = P_r \tag{26}$$

The solution of the above problems can be solved by Lagrange multiplier method, and the corresponding Lagrange function can be written as：

$$Z\left(\psi, P_{ri}\right) = \sum_{i=1}^{N} \log_2\left(1+\frac{P_{ri}\psi_i}{\sigma_h^2}\right) + \gamma\left(P_r - \sum_{i=1}^{N} P_{ri}\right) \tag{27}$$

In the above formula, the partial derivative of P_{ri} can be obtained：

$$\frac{\partial Z}{\partial P_{ri}} = \frac{1}{In2} \times \frac{\frac{\psi_i}{\sigma_h^2}}{1+\frac{P_{ri}\psi_i}{\sigma_h^2}} - \gamma = 0 \tag{28}$$

$$P_{ri} = \frac{1}{\gamma \times In2} - \frac{\sigma_h^2}{\psi_i} \tag{29}$$

Since the power cannot be negative, write the above formula as follows：

$$P_{ri} = \left[\frac{1}{\gamma \times In2} - \frac{\sigma_h^2}{\psi_i}\right]^+ \tag{30}$$

Where $[a]^+ = \max(a, 0)$ and γ is constant.

APPENDIX B

PROOF OF LEMMA 1

$$\frac{\partial L}{\partial \bar{R}} = 1 - \mu^{op} - \nu^{op} = 0 \tag{31}$$

$$\mu^{op}\left(\bar{R}^{op} - \tau_1^{op}\log_2\left|I + \sum_{k=1}^{K} SINR\right|\right) = 0 \tag{32}$$

$$\nu^{op}\left(\bar{R}^{op} - \tau_2^{op}\sum_{i=1}^{N}\log_2\left(1 + \frac{P_{ri}^{op}\psi_i}{\sigma_h^2}\right)\right) = 0 \tag{33}$$

Where μ^{op} and ν^{op} are the optimal dual variables of throughput constraint in OP_3. It is obvious from (31) that $\nu^{op} = 1 - \mu^{op}$ makes the Lagrangian dual function bounded, Using this result, $0 < \mu^{op} < 1$ and $0 < \nu^{op} < 1$ will be proved.

When $\mu^{op} = 0$ and $\nu^{op} = 1$.According to the KKT complementary relaxation condition in (33), it can be concluded that $\bar{R}^{op}$ is the optimal solution that satisfies the throughput constraints in (15b) and is equal, i.e., $\bar{R}^{op} = \tau_2^{op}\sum_{i=1}^{N}\log_2\left(1 + \frac{P_{ri}^{op}\psi_i}{\sigma_h^2}\right)$. According to the KKT conditions in $\mu^{op} = 0$ and (32), $\bar{R}^{op} - \tau_1^{op}\log_2\left|I + \sum_{k=1}^{K} SINR\right| = 0$ does not exist, because $\mu^{op} = 0$ means that the constraint in C_4 satisfies a strict inequality without loss of generality, i.e.,

$\bar{R}^{op} < \tau_1^{op}\log_2\left|I + \sum_{k=1}^{K} SINR\right|$. Suppose there is P^1, so that

$\bar{R}^1\left(P^1\right) = \left(\tau_1^{op} C_{U-R}\left(P^1\right)\right)^1 = \left(\tau_2^{op} C_{R-H}\left(P^1\right)\right)^1$ and $P^{op} < P^1$, because $\tau_2^{op}\sum_{i=1}^{N}\log_2\left(1 + \frac{P_{ri}^{op}\psi_i}{\sigma_h^2}\right)$ is an increasing function about P, it is get $\bar{R}^{op} < \bar{R}^1$, which contradicts the assumption that $\bar{R}^{op}$ is the optimal solution. Therefore, μ^{op} cannot be 0.

Than, The other case is $\mu^{op} = 1$, According to (31) know $\nu^{op} = 0$. By taking similar steps as before and considering the KKT conditions in (32) and (33), it is easy to get that $\bar{R}^{op}$ is the optimal solution, so that $\bar{R}^{op} = \tau_1^{op}\log_2\left|I + \sum_{k=1}^{K} SINR\right|$, without losing the generality of $\bar{R}^{op} < \tau_2^{op}\sum_{i=1}^{N}\log_2\left(1 + \frac{P_{ri}^{op}\psi_i}{\sigma_h^2}\right)$. Suppose there is P^2, so that $\bar{R}^2\left(P^2\right) = \left(\tau_1^{op} C_{U-R}\left(P^2\right)\right)^2 = \left(\tau_2^{op} C_{R-H}\left(P^2\right)\right)^2$ and $P^{op} < P^2$, because $\tau_1 C_{U-R}$ is an increasing function about P, it is get $\bar{R}^{op}\left(P^{op}\right) < \left(\tau_1^{op} C_{U-R}\left(P^2\right)\right)^2$, i.e. $\bar{R}^{op} < \bar{R}^2$. Therefore, μ^{op} cannot be

1.If $\mu^{op} > 1$, then $\nu^* < 0$, this violates the inequality constraint in the KKT condition where the dual variable is equal to or greater than zero. Therefore, it can be concluded that the optimal dual variable for the throughput constraint of OP_3 should satisfy $0 < \mu^{op} < 1$ and $0 < \nu^{op} < 1$.

APPENDIX C

PROOF OF LEMMA 2

$$R_{sum}(\tau_0) = \max_{E_r, E_k, \tau_1, \tau_2, \bar{R}} \quad \bar{R} \tag{34}$$

$$s.t. \quad C_2, \quad C_3$$

$$C_4 : E_{Hr} = \varepsilon \tau_0 tr(G_r S) \tag{34a}$$

$$C_5 : E_{Hk} = \varepsilon \tau_0 tr(G_k S) \tag{34b}$$

$$R_{sum}(\tau_0) = \min \max_{E_r, E_k, \tau_1, \tau_2, \bar{R}} \bar{R} - a_1\tau_0 - b_1(\tau_{tr} + \tau_1 + \tau_2 - 1) - b_2(E_{Hr} - \tau_0 \varepsilon tr(G_r S)) - b_3(E_{Hk} - \tau_0 \varepsilon tr(G_k S)) \tag{35}$$

Where $\bar{R}$, τ_1, τ_2, S are the optimal solution. For arbitrary $\rho \in (0,1)$,

$$\begin{aligned} &R_{sum}(\rho x_1 + (1-\rho)x_2) = \min_{a_{1\geq 0}, b \geq 0} \\ &\bar{R} - a_1(\rho x_1 + (1-\rho)x_2) - b_1(\tau_{tr} + \tau_1 + \tau_2 - 1) - b_2(E_{Hr} - (\rho x_1 + (1-\rho)x_2)\varepsilon tr(G_r S)) \\ &- b_3(E_{Hk} - (\rho x_1 + (1-\rho)x_2)\varepsilon tr(G_k S)) \\ &= \min_{a_1 \geq 0, b \geq 0} \rho\left[\bar{R} - a_1 x_1 - b_1(\tau_{tr} + \tau_1 + \tau_2 - 1) - b_2(E_{Hr} - x_1 \varepsilon tr(G_r S)) - b_3(E_{Hk} - x_1 \varepsilon tr(G_k S))\right] \\ &+ (1-\rho)\left[\bar{R} - a_1 x_2 - b_1(\tau_{tr} + \tau_1 + \tau_2 - 1) - b_2(E_{Hr} - x_2 \varepsilon tr(G_r S)) - b_3(E_{Hk} - x_2 \varepsilon tr(G_k S))\right] \end{aligned} \tag{36}$$

Since $\min_x f(x) + g(x) \geq \min_x f(x) + \min_x g(x)$ for $f(x)$ and $g(x) \in R$

So：

$$\begin{aligned} &R_{sum}(\rho x_1 + (1-\rho)x_2) \geq \\ &\rho \min_{a_1 \geq 0, b \geq 0}\left[\bar{R} - a_1 x_1 - b_1(\tau_{tr} + \tau_1 + \tau_2 - 1) - b_2(E_{Hr} - x_1 \varepsilon tr(G_r S)) - b_3(E_{Hk} - x_1 \varepsilon tr(G_k S))\right] \\ &+ (1-\rho) \min_{a_1 \geq 0, b \geq 0}\left[\bar{R} - a_1 x_2 - b_1(\tau_{tr} + \tau_1 + \tau_2 - 1) - b_2(E_{Hr} - x_2 \varepsilon tr(G_r S)) - b_3(E_{Hk} - x_2 \varepsilon tr(G_k S))\right] \\ &= \rho R_{sum}(x_1) + (1-\rho) R_{sum}(x_2) \end{aligned} \tag{37}$$

According to [17], $R_{sum}(\tau_0)$ is concave with τ_0.

REFERENCES

[1] KANG Xin, Ho C K, SUN Sumei. “Full-Duplex Wireless-Powered Communication Network with Energy Causality”. IEEE Transactions on Wireless Communications, 2015, 14(10): 5539-5551.

[2] XU Jei, ZHANG Rui. “Throughput Optimal Policies for Energy Harvesting Wireless Transmitters with Non-Ideal Circuit Power”. IEEE Journal on Selected Areas in Communications, 2014, 32(2): 322-332.

[3] LU Xiao, WANG Ping, NIYATO D, et al. “Wireless Networks With RF Energy Harvesting: A Contemporary Survey”. IEEE Communications Surveys & Tutorials, 2015, 17(2): 757-789.

[4] MIKEKA C, ARAI H. “Design of a cellular energy-harvesting radio”// 2009 European Wireless Technology Conference. Rome, Italy. IEEE, 2009: 73-76

[5] RAMEZANI P, JAMALIPOUR A. “Toward the Evolution of Wireless Powered Communication Networks for the Future Internet of Things”. IEEE Network, 2017, 31(6): 62-69.

[6] YANG Zhaohui, XU Wei, PAN Yijin, et al. “Optimal Fairness-Aware Time and Power Allocation in Wireless Powered Communication Networks”. IEEE Transactions on Communications, 2018, 66(7): 3122-3135.

[7] LIU Mengyu, LIU Yuan. “Charge-Then-Forward: Wireless-Powered Communication for Multiuser Relay Networks”. IEEE Transactions on Communications, 2018, 66(11): 5155-5167.

[8] HU Guojie, CAI Yueming, XU Kui, et al. “Opportunistic Energy Harvesting for Multi-Antenna-Relay-Assisted Wireless Powered Communication Network”. IEEE Communications Letters, 2019, 23(1): 148-151.

[9] RAMEZANI P, JAMALIPOUR A. “Two-Way Dual-Hop WPCN With A Practical Energy Harvesting Model”. IEEE Transactions on Vehicular Technology, 2020, 69(7): 8013-8017.

[10] HE Chunlong, LIANG Jiaqian, QIAN Gongbin, et al. “Optimal Time Allocation in Multi-Cell Wireless Powered Communication Networks”. IEEE Access, 2019, 7: 26519-26526.

[11] ZHENG Yali, HU Jie, YANG Kun. “Sum - Throughput Maximisation in Multi - Antenna aided Full-Duplex WPCNs with Self-Interference”//2020 International Wireless Communications and Mobile Computing (IWCMC). Limassol, Cyprus. IEEE, 2020: 1240-1245.

[12] KIDA Y, DEGUCHI M, SHIMURA T. “The Effects of Interferences on Time Reversal MIMO: An Evaluation of Multipath and Co-Channel Interference”//2018 OCEANS - MTS/IEEE Kobe Techno-Oceans (OTO). Kobe, Japan. IEEE, 2018: 1-5.

[13] LEI Weijia, YAO Li. “Performance Analysis of Time Reversal Communication Systems”. IEEE Communications Letters, 2019, 23(4): 680-683.

[14] MA Hang, WANG Beibei, CHEN Yan, et al. “Waveforming Optimizations for Time-Reversal Cloud Radio Access Networks”. IEEE Transactions on Communications, 2018, 66(1): 382-393.

[15] DEY I, CIUONZO D, ROSSI P S. “Wideband Collaborative Spectrum Sensing Using Massive MIMO Decision Fusion”. IEEE Transactions on Wireless Communications, 2020, 19(8): 5246-5260.

[16] NIGUS H R, KIM K H, HWANG D. “Multi-antenna channel capacity enhancement in wireless communication”// 2015 Seventh International Conference on Ubiquitous and Future Networks. Sapporo, Japan. IEEE, 2015: 77-82.

[17] BOYD S, VANDENBERG L. “Convex optimization”. Cambridge: Cambridge University Press, 2004: 61-256.

[18] GRANT M . CVX: MATLAB software for disciplined convex programming. http://cvxr.com/cvx, 2008.

AUTHORS

Wei Liu Born in 1994, from Inner Mongolia, China. Now he is a master's student of Chongqing University of Posts and telecommunications. His main research interests are wireless power communication network and time reversal.

Fang Wei Li Born in 1960 in Chongqing, China. Professor and doctor of Chongqing University of Posts and telecommunications. His main research interests are electromagnetic field and electromagnetic wave, wireless network security, wireless transmission theory and technology.

Jun Zhou Xiong Born in 1985 in Hubei, China. Now he is a doctoral student of Chongqing University of Posts and telecommunications, and his main research direction is the key technology of wireless communication and physical layer security technology.

Ming Yue Wang Born in 1990 in Chongqing, China. Now her is a doctoral student of Chongqing University of Posts and telecommunications, and her main research direction is wireless transmission theory and technology.

© 2022 By AIRCC Publishing Corporation. This article is published under the Creative Commons Attribution (CC BY) license.

RESEARCH ON THROUGHPUT MAXIMIZATION OF WIRELESS POWERED COMMUNICATION NETWORK BASED ON A RETRO DIRECTIVE MATRIX

Bo Li[1, 2] and Hong Tang[1, 2]

[1]School of Communication and Information Engineering, Chongqing University of Posts and Telecommunications, Chongqing 400065, China;
[2]Chongqing Key Laboratory of Mobile Communications Technology, Chongqing 400065, China

ABSTRACT

Aiming at the problem of limited system throughput caused by double near-far effect in wireless power communication network. In this paper, a retro directive matrix method based on phase conjugation is proposed. In the method, energy base stations and information base stations are depolyed separately, energy base station uses large-scale multiple input multiple output (MIMO) system, when system running point equipment firstly to send a beacon signal to energy base station, the energy base station amplifies its conjugate to form a directional beam to achieve multi-input and multi-output energy gains, thus improving the throughput of information transmission of node devices. Through the optimizated and allocated the time of beacon signal, the time of energy transmission, the time of information transmission and some power parameters, a convex optimization problem is proposed. And it has been solved by Lagrange generalized multiplier method and golden section method. Simulation results show that the proposed method has better performance than others projects.

KEYWORDS

Wireless Powered Communication Network, Matrix Retrodirective Array, Energy Transmision, Information Transmision, System Throughput.

1. INTRODUCTION

Conventional wireless sensor networks, such as those that detect earthquakes, temperature, humidity and noise, are powered by batteries. Battery power comes with a number of pitfalls, such as limited available time. If the replacement is delayed, communication is interrupted and service quality is affected [1]. Wireless power communication network (WPCN)is a new type of network which combines energy transmission and traditional information transmission. Energy transmitter uses Radio frequency (RF) transmission mode to transmit energy, and sensor node equipment can realize self-sustainable information transmission after acquiring these energy [2].

In literature [3], a classical WPCN is studied for the first time and a "acquisition first, transmission later" protocol is proposed, in which the user (device) first obtains energy from the Downlink (DL, Downlink) through mixed node H-AP broadcast. The obtained energy is then used to send information to the hybrid node H-AP on the UL (Uplink).H-AP here refers to the deployment of energy nodes and information nodes together, which results in limited information

David C. Wyld et al. (Eds): NATP, ACSTY, CCCIOT, MLSC, ITCSS - 2022
pp. 51-63, 2022. CS & IT - CSCP 2022 DOI: 10.5121/csit.2022.120104

transmission distance and double near-far effect. Firstly, the limited information transmission distance is due to the difference in the scope of energy transmission and information transmission. Energy transmission is different from the structure and antenna system of information transmission, and has high requirements on the sensitivity of the receiver. Usually, the effective range is about 15 meters. Secondly, "double near and far effect" means that users who are far away from H-AP receive less energy on DL than those who are close to H-AP. However, due to signal attenuation caused by double distance on DL and UL, remote users need to transmit signals at higher power on UL to ensure service quality. Taken together, these problems can result in low system coverage and limited throughput. In order to solve the problem of limited throughput of wireless power supply network, time inversion technology is added in the information transmission stage in literature [4] to increase system capacity by resisting multipath effect. Literature [5] studies the problem of throughput maximization of wireless power supply network based on NOMA technology. All node devices are equipped with multiple antennas, and NOMA technology is adopted for transmission in the information transmission stage to increase system throughput through beamforming. Literature [6] and [7] studied the maximization of wireless power supply network throughput of UAV as node device. Energy transmission is also particularly important in the wireless power supply network. Node devices need to obtain enough energy to ensure good service quality.

Therefore, in order to overcome the problem of reduced system throughput caused by the limited energy of node devices due to significant power loss over long distances in rf wireless energy transmission, Multi-antenna directional transmission of energy or energy beamforming (EB) can be a good solution [8]. However, the practical implementation of EB requires obtaining the perfect CSI in the energy emitter. In order to obtain CSI, literature [9] uses forward link training with CSI feedback of receiver, and literature [10] uses reverse link training with channel reciprocity, as well as training based on energy feedback. However, prior to link training and training methods based on energy feedback and feedback overhead actually is very high, especially in the equipment has a large number of nodes or has a large-scale multiple input multiple output system, but as this article proposed a reverse link training method based on phase conjugate feedback because they don't need node equipment, and the length of training has nothing to do with the number of transmit antennas, This has huge advantages for large-scale MIMO systems[11]. Back in the direction of the array is on the basis of the principle of phase conjugate reverse link training method, has proven to be an effective way of wireless power transmission (WPT), able to position the target under the condition of unknown, automatic back entrance to the direction of the wave, and there's no need to through complex digital signal processing algorithm, this method has been widely used in radio frequency identification, microwave imaging, Radar anti-collision system and other identification systems.

This article introduced in the traditional wireless network reverse beamforming is a kind of low complexity, in this kind of technology, all nodes to equipment energy transmitter launch a public beacon signals at the same time, all energy transmitter antenna on the conjugate amplifier, and on all the nodes radio equipment energy, node equipment using its energy to the base station to send information. The work of this paper is as follows:

1. Improve the traditional wireless power supply network model, deploy energy nodes and information nodes separately, so as to increase the coverage of the system
2. By using channel reciprocity, the directional backtracking matrix method based on phase conjugate is added in the energy transmission stage, and large-scale MIMO system is adopted in the energy base station to make the node equipment obtain energy gain, so as to increase the throughput in the information transmission stage.
3. Considering the quality of service of node equipment, the time allocation and power control of beacon signal, energy transmission and information transmission are jointly

optimized, and the problem model is planned, and the maximum throughput is solved by using Lagrange duality method combined with golden Section algorithm.

4. Other classical wireless power supply network schemes are compared and analyzed by simulation to prove the effectiveness of the proposed scheme.

2. SYSTEM MODEL

As shown in Figure 1, we have studied the multi-user wireless power communication network under a large-scale antenna array. The energy transmitter ET has M_t transmitting antenna, and there are k node device (ER), each node device has an antenna, and the information receiver IR has a receiving antenna. The transmission matrix from ET to ER is represented by H .

$$H = \begin{bmatrix} h_{11} & h_{12} & \dots & h_{1M_t} \\ h_{21} & h_{22} & \dots & h_{2M_t} \\ \vdots & \vdots & & \vdots \\ h_{K1} & h_{K2} & \dots & h_{KM_t} \end{bmatrix} \quad (1)$$

Among them, $h_{ij} = (i = 1,2,...,K; j = 1,2,...M_t)$ represents the channel transmission coefficient from the j^{th} energy transmitter antenna to the i^{th} receiver antenna. Assume that all transmitted signals are narrowband signals:

$$h_{ij} = \sqrt{\beta_{ij}} S_{ij} \quad (2)$$

β_i and g_i represents the large-scale fading coefficient of the channel, and S_{ij} represents the small-scale fading coefficient of the channel. The large-scale fading coefficient is related to the distance between equipment and energy transmitter ET and information receiver IR. The large-scale fading coefficient of each node device and all ET antennas is the same, which can be expressed as:

$$\beta_i = c_0 (r_i / r_0)^{-\alpha} \quad (3)$$

The large scale fading coefficient from each node device to the IR of the information receiver is also the same, g_i can be expressed as:

$$g_i = c_0 (d_i / r_0)^{-\alpha} \quad (4)$$

Where $c_0 = -30dB$ is the constant attenuation factor of path loss at the reference distance $r_0 = 1m$. α is the path loss index, r_i is the distance from the i^{th} antenna of the terminal device to the energy transmitter, d_i represents the distance between the i^{th} antenna of the terminal device and the information receiver. The fading coefficient of small scale S_{ij} independent from antenna of different energy transmitter to antenna of different receiver. It is a complex Gaussian random variable with zero mean unit variance, $S_{ij} \square CN(0, 1)$. The channel from ET to ER_k is represented by $h_k^* \square [h_{k1}.....h_{kM_t}]^T$, $a*, a^T$ represents the conjugate and transpose of the copy vector. It is assumed that the channels from the energy transmitter ET to the node device are reciprocal

and that the information receiver IR has perfect CSI for all node devices, so the channel from ER_k to ET can be represented by h_k^H , a^H represents the conjugate transpose of the vector a .

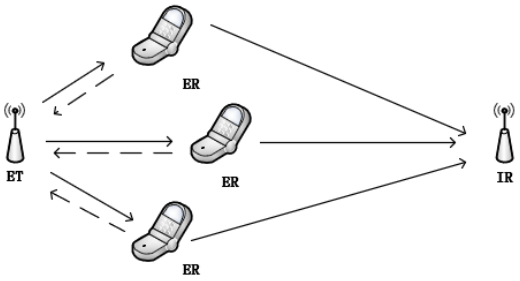

Figure 1. The system model diagram

3. SYSTEM SOLUTION

Based on the reciprocity of channel, a wireless power supply network scheme with low complexity based on phase conjugate directional backtracking array is proposed. Each time transmission block is composed of three time slots. In first time slot τ_1, the node device transmits a beacon signal to the energy transmitter ET. In second time slot τ_2, the energy transmitter ET transmits energy to the node equipment. In third time slot τ_3, the node device sends information to the information receiver IR. And each node device has a certain amount of energy before the system starts to ensure that the node device can send beacon signals to the energy transmitter ET. In the following time block, the energy transmitted by the acquired energy transmitter ET is used to transmit information to the information receiver IR, so as to realize the self-sustainability of node equipment. Figure 2 is the slot allocation diagram of the system model in this paper.

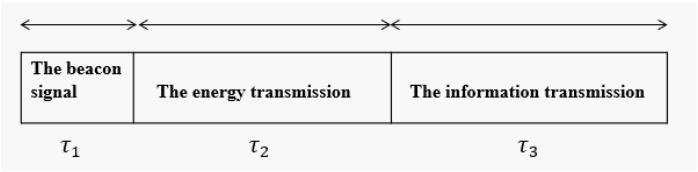

Figure 2. The timeslot allocation diagram of this paper model

3.1. Solution Steps

1) Beacon signal stage: dotted arrow in the system model figure in Figure 1, K node devices simultaneously send beacon signals to the energy transmitter ET when the system is running, which can be expressed as:

$$\Phi_k(t) = \sqrt{2P_k}\cos(2\pi f_c t) \qquad (5)$$

P_k is the power of the node device to send beacon signals, $0 \le P_k \le P_{\max}$, $P_{\max}$ is the maximum transmitting power of beacon signal, f_c is carrier frequency, beacon signal duration is τ_1. So the system bandwidth is $w = 1/\tau_1$. The equivalent baseband signal received by ET is expressed as:

$$y(t) = \sum_{k=1}^{K} \sqrt{P_k} h_k^* + z(t) \qquad (6)$$

$$= g + z(t) \qquad (7)$$

$z(t) \sim [z_1(t),.....z_{M_t}(t)]^T$ represents additive White Gaussian noise (AWGN) with mean value zero and power spectral density is N_0 .At the same time, $g \sim \sum_{k=1}^{K} \sqrt{P_k} h_k^*$ represents the effective weighted linear combination signal received by the energy transmitter ET sent by K nodal devices. Then the energy transmitter ET performs matching filtering operation on the received signal $y(t)$. The result of that is $\hat{g}$, can be expressed as:

$$\hat{g} = \frac{1}{\tau} \int_0^{\tau} y(t)dt = g + \tilde{g} \qquad (8)$$

Among them, $\tilde{g} \sim \frac{1}{\tau} \int_0^{\tau} z(t)dt \sim CN(0, \frac{N_0}{\tau} I)$, $\boldsymbol{0}$, $\boldsymbol{I}$, represents the all zero vector of size $M_t \times 1$, and the identity matrix of size $M_t \times M_t$ .At this point, the energy consumed by node equipment ER_k at this stage can be expressed as:

$$E_k^c = P_k * \tau_1 \qquad (9)$$

2) Energy transfer stage: as shown in the left straight arrow in Figure 1, the energy transmitter ET sends energy to K node device ER, Specifically, $\hat{g}$ is conjugated at the receiving end of the energy transmitter. Each antenna sends a sinusoidal signal using the same carrier f_c as the beacon signal. The transmitted power is P_t . At this time, equivalent baseband transmitting signal of energy transmitter can be expressed as:

$$x = \sqrt{P_t} \frac{\hat{g}^*}{\|\hat{g}\|} \qquad (10)$$

Then the signal received by each node device can be expressed as:

$$r_k = h_k^H x \qquad (11)$$

$k = 1,...,K$, accordingly, the energy received by each node device is E_k , can be expressed as:

$$E_k = |r_k|^2 * \tau_2 = \frac{P_t}{\|\hat{g}\|^2} \left| \sum_{l=1}^{K} \sqrt{P_l} h_k^H h_l + h_k^H \tilde{g}^* \right|^2 \tau_2 \qquad (12)$$

In this phase, the energy transfer time is τ_2 , For the sake of simplicity, we ignore the power lost by the circuit during the actual transmission.

3) Information transmission stage: as shown in the straight arrow on the right in Figure 1, the node device ER adopts the mode of air division multiple access to simultaneously send information to the information receiver IR, and the transmission time is τ_3 . It is assumed that each node device consumes its acquired energy during the information transmission phase, leaving only the next time block for the node device to transmit the energy E_k^c of the detection signal. At this point, the transmitted power ER_k of each node device can be expressed as:

$$P_k^S = \frac{E_k - E_k^c}{\tau_3} \tag{13}$$

At this point, each node device transmits data to the information receiver within the unit time transmission block. In this process, the throughput that a single node device can achieve can be expressed as follows:

$$R_k = W\tau_3 \log_2(1 + \frac{P_k^S g_k}{N_0}) \tag{14}$$

Combined with (9),(12),(13) and (14), it can be further concluded that:

$$R_k = W\tau_3 \log_2(1 + \frac{(\frac{P_t}{\|\hat{g}\|^2}\left|\sum_{l=1}^{K}\sqrt{P_l}h_k^H h_l + h_k^H \tilde{g}^*\right|^2 \tau_2 - P_k * \tau_1)g_k}{\tau_3 N_0}) \tag{15}$$

Therefore, the total system throughput achieved by all nodes within the unit time transmission block is:

$$R = \sum_{k=1}^{K} W\tau_3 \log_2(1 + \frac{(\frac{P_t}{\|\hat{g}\|^2}\left|\sum_{l=1}^{K}\sqrt{P_l}h_k^H h_l + h_k^H \tilde{g}^*\right|^2 \tau_2 - P_k * \tau_1)g_k}{\tau_3 N_0}) \tag{16}$$

4. PROGRAMMING PROBLEM

In order to maximize the system throughput R, we need to allocate time for beacon signal time τ_1, energy transmission time τ_2 and information transmission time τ_3. Without loss of generality, the sum of single transmission fast time is 1. Considering the service quality of single node equipment, the following problems are planned:

$$Max \quad R = \sum_{k=1}^{K} W\tau_3 \log_2(1 + \frac{(\frac{P_t}{\|\hat{g}\|^2}\left|\sum_{l=1}^{K}\sqrt{P_l}h_k^H h_l + h_k^H \tilde{g}^*\right|^2 \tau_2 - P_k * \tau_1)g_k}{\tau_3 N_0}) \tag{17}$$

$$S.t \ \tau_1 + \tau_2 + \tau_3 = 1 \tag{18}$$

$$R_k \geq R_{\min} \ (k = 1.....K) \tag{19}$$

$$\tau_1, \tau_2, \tau_3, P_t, P_k, P_k^S > 0 \tag{20}$$

$R_{\min}$ represents the minimum throughput of node devices. (19) Constraints indicate that the throughput of any node device in the system must be greater than or equal to the minimum throughput requirements to ensure the service quality of node devices.

4.1. Analysis and solution of optimal solution

In order to solve make it easier to solve, we let $\tau_2 + \tau_3 = m$. According to the convex optimization theory, it can be concluded that the system throughput R is a strictly concave function about τ_2 in

the domain $[0,1]$. And it has a maximum in its domain. The simulation assistant is set here to prove this conclusion. Simulation parameters are set as： M_t =120, Power of beacon signal emitted by node device P_k =0.2W, The transmitted power of ET is P_t =2W. τ_1 =0.1s, α =3. Noise power spectral density is N_0 =-55 dBm/Hz.

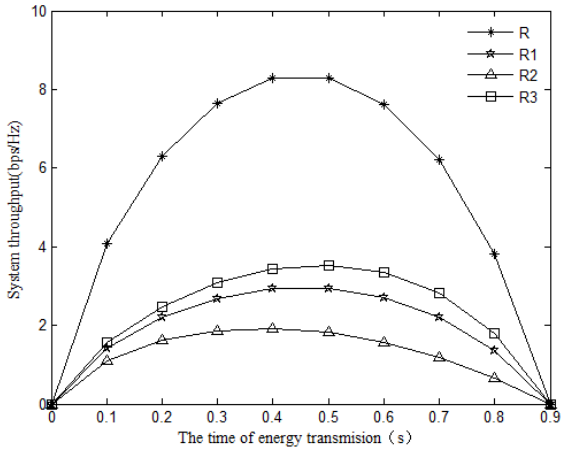

Figure 3. The relationship diagram of energy transfer time and system throughput

Here, nine node devices are set up in the simulation, which are divided into three groups R_1, R_2, R_3 .There are three nodes in each group. R_3 is 4-6 meters away from the energy base station and 100 meters away from the information base station. R_1 is 9~11 meters away from the energy base station and 95 meters away from the information base station. R_2 is 13~15 meters away from the energy base station and 90 meters away from the information base station. According to Figure 3, it can be proved that the system throughput is a concave function of τ_2 and has a maximum value. Therefore, the original problem can be transformed into a standard convex optimization problem, and the objective function becomes:

$$Min \quad R = -\sum_{k=1}^{K} W\tau_3 \log_2(1+\frac{A_k\tau_2 - B_k\tau_1}{C_k\tau_3}) \quad (21)$$

$$S.t \quad \tau_1 + \tau_2 + \tau_3 - 1 = 0 \quad (22)$$

$$R_{\min} - W\tau_3 \log_2(1+\frac{A_k\tau_2 - B_k\tau_1}{C_k\tau_3}) \le 0 \quad (23)$$

$$\tau_1, \tau_2, \tau_3, P_t, P_k, P_k^S > 0 \quad (24)$$

$$(k = 1.....K)$$

These variables including $A_k = \frac{P_t}{\|\hat{g}\|^2}\left|\sum_{l=1}^{K}\sqrt{P_l}h_k^H h_l + h_k^H \tilde{g}^*\right|^2 g_k$, $B_k = P_x * g_k$, $C_k = N_0$. The maximum problem is transformed into a minimum standard convex optimization problem with constraints. In order to further solve, the constraint condition (22) is changed to

$$\tau_2 = m - \tau_3 \quad (25)$$

Here $m = 1 - \tau_1$ is a constant, Substitute (25) into (21) to obtain the new objective function:

$$Min \qquad R = -\sum_{k=1}^{K} W\tau_3 \log_2(1+\frac{A_k(m-\tau_3)-B_k\tau_1}{C_k\tau_3}) \qquad (26)$$

$$S.t \quad R_{\min} - W\tau_3 \log_2(1+\frac{A_k(m-\tau_3)-B_k\tau_1}{C_k\tau_3}) \le 0 \qquad (27)$$

$$\tau_1, \tau_2, \tau_3, P_t, P_k, P_k^S > 0 \qquad (28)$$

Lagrange multipliers $\lambda_k (k = 1.....K)$ are introduced to solve the above convex optimization, then the Lagrange form of the objective function is as follows problems:

$$L(\tau_3, \lambda_k) = \sum_{k=1}^{K} W\tau_3 \log_2(1+\frac{A_k(m-\tau_3)-B_k\tau_1}{C_k\tau_3}) + \lambda_k (R_{\min} - W\tau_3 \log_2(1+\frac{A_k(m-\tau_3)-B_k\tau_1}{C_k\tau_3})) \quad (29)$$

This paper adopts Lagrange duality method to solve the problem, so the dual function is:

$$g(\lambda_k) = \min\ L(\tau_3, \lambda_k) \qquad (30)$$

The dual problem is:

$$\max \quad g(\lambda_k) = \sum_{k=1}^{K} W\tau_3 \log_2(1+\frac{A_k(m-\tau_3)-B_k\tau_1}{C_k\tau_3}) + \lambda_k (R_{\min} - W\tau_3 \log_2(1+\frac{A_k(m-\tau_3)-B_k\tau_1}{C_k\tau_3})) \quad (31)$$

According to the convex optimization theory, (26) and (27) are convex functions, and the dual gap is zero, which meets the strong dual condition. Therefore, the solution of the original function is the solution of the dual function. The duality function is solved below. According to the Karloch-Kuhn-Tucker (KKT) condition, the partial derivative of τ_3 is obtained:

$$\frac{\partial L(\tau_3, \lambda_k)}{\partial \tau_3} = \sum_{k=1}^{K} W[\log_2(1+\frac{A_k(m-\tau_3)-B_k\tau_1}{C_k\tau_3}) - \frac{A_k m + B_k\tau_1}{(A_k(m-\tau_3)-B_k\tau_1+C_k\tau_3)In2}]$$

$$+\lambda_k W[\log_2(1+\frac{A_k(m-\tau_3)-B_k\tau_1}{C_k\tau_3}) - \frac{A_k m + B_k\tau_1}{(A_k(m-\tau_3)-B_k\tau_1+C_k\tau_3)In2}] \qquad (32)$$

Let $\frac{\partial L(\tau_3, \lambda_0, \lambda_k)}{\partial \tau_3} = 0$, get out:

$$\lambda_k^* = \frac{\sum_{k=1}^{K}[\log_2(1+\frac{A_k(m-\tau_3^*)-B_k\tau_1}{C_k\tau_3^*}) - \frac{A_k m + B_k\tau_1}{(A_k(m-\tau_3^*)-B_k\tau_1+C_k\tau_3^*)In2}]}{[\log_2(1+\frac{A_k(m-\tau_3^*)-B_k\tau_1}{C_k\tau_3^*}) - \frac{A_k m + B_k\tau_1}{(A_k(m-\tau_3^*)-B_k\tau_1+C_k\tau_3^*)In2}]} \qquad (33)$$

τ_3^* and λ_k^* are the optimal solutions of the original problem and the dual problem. On this basis, the maximum value of $g(\lambda_k)$ can be calculated by combining the golden section method. That is to calculate the maximum throughput of the system, the algorithm table for solving the maximum throughput of the system is proposed here.

Table 1. Table of time allocation and throughput algorithms

1）Calculate m with given τ_1，
2）Initialize $\tau^{low}=0,\tau^{up}=m$，
1. $\tau_3^a=\tau^{low}+0.382*(\tau^{up}-\tau^{low})$
2. $\tau_3^b=\tau^{low}+0.618*(\tau^{up}-\tau^{low})$
3.Compute $\bar{\lambda}_{k1},\bar{\lambda}_{k2}$ by substituting τ_3^a and τ_3^b into (33)
4.Compute $\bar{g}_1$ by substituting τ_3^a and $\bar{\lambda}_{k1}$ into (31)
5.Compute $\bar{g}_2$ by substituting τ_3^b and $\bar{\lambda}_{k2}$ into (31)
6.if $\bar{g}_1<\bar{g}_2$， let $\tau^{up}=\tau_3^b$， else $\tau^{low}=\tau_3^a$。
3)if $\tau^{up}-\tau^{low}\le\varepsilon$， let $g(\lambda_k)=(\bar{g}_1+\bar{g}_2)/2$，
$\tau_3^*=(\tau^{up}+\tau^{low})/2$， otherwise go to step 1， until
$\tau^{up}-\tau^{low}\le\varepsilon$

4.2. Complexity Analysis

In this paper, the maximum throughput and the allocation of each time slot per unit time block are obtained by Lagrange duality method combined with golden Section algorithm, namely the algorithm in Table 2, whose complexity is approximately $O(n\log N)$.Under the condition of maintaining the same accuracy, compared with references [4], [5]($O(n\wedge k)$) and [3]($O(n^2)$), the complexity of the algorithm in this paper is greatly reduced, and the convergence effect can be achieved more quickly, so as to quickly calculate the beacon signal detection time, energy transmission time and the allocation of information transmission time, thus improving the system performance.

5. SIMULATION RESULTS

Simulation results demonstrate the excellent performance of the proposed retrodirective matrix method based on phase conjugation in wireless power communication networks. In order to better highlight the performance of the proposed scheme, a comparative analysis is made with the traditional wireless power supply network scheme [3], the scheme based on time inversion [4] and the scheme based on NOMA [5]. As shown in Figure 4, the simulation analyzes the influence of the energy transmitter ET transmitting power on the system throughput, where the simulation parameter is set as,

$M_t=40$， $P_k=P_{\max}=0.1W$， $\tau_1=0.05s$， $f_c=910MHz$， $W=20kHz$，

$N_0=-55\,\mathrm{dBm/Hz}$， $\alpha=3$，

The system contains nine nodes. Their relative positions are as follows:

$r_1=d_1$ ranges from 4 to 6 meters. $r_2=d_2$ ranges from 9 to 11 meters. $r_3=d_3$ is in the range of 13 to 15 meters. Three nodes are randomly placed in each range. In order to compare simulation parameters uniformly, energy transmitter and information receiver are placed together. It can be seen from the figure that as the power of the energy transmitter increases, the system throughput of the retrospective matrix method based on phase conjugate and all other schemes increases

accordingly. However, when the power of energy transmitter is the same, the throughput of the backtracking array based on phase conjugate is greater than that of other schemes, and the throughput of the scheme in this paper, the scheme based on time inversion and the scheme based on NOMA are all greater than that of the traditional scheme. It is proved that the system throughput can be improved by increasing ER transmitting power in practical application.

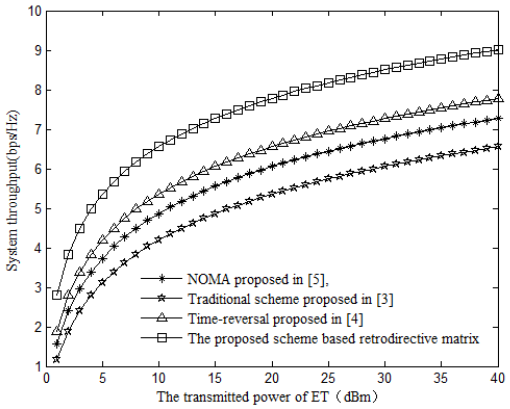

Figure 4. The relationship diagram of transmitted power and system throughput

As shown in Figure 5, the simulation studies the influence of different positions of node equipment on system throughput. It should be noted that node equipment is between energy transmitter and information receiver. $P_t = 30dBm$, A node device is set up in the simulation, and the energy transmitter and the information receiver are placed separately. $d_1 = 80$. As can be seen from the figure, when the distance from the energy transmitter is less than 4 meters, the throughput of the transmission scheme based on NOMA is slightly higher than that of the proposed scheme and other schemes at the same distance. When the distance is beyond 4 meters, the throughput of the proposed scheme is greater than that of other schemes with the increase of the distance. It is proved that the proposed scheme has a wider coverage and better robustness.

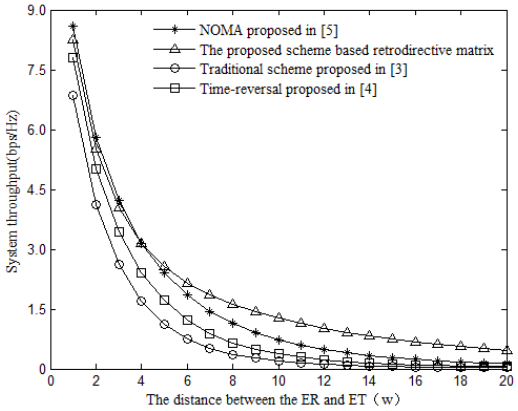

Figure 5. The relationship diagram of the location of node devices and system throughput

As shown in figure 6, the simulation research of antenna number influence on system throughput, because this article scheme and is based on NOMA scheme adopts large-scale MIMO antenna array, the traditional solutions and based on the time reversal USES a single antenna, so it can be seen in the picture increase number of traditional antenna scheme and had no effect on the solution of inversion based on time, Here, 9 node devices are set, and the relative positions of nodes are the same as in simulation figure 5.As can be seen from the figure, it can be noted that,

$P_t = 30dBm$, when the number of antennas is the same and both are greater than 10, the system throughput of the scheme proposed in this paper is greater than that of other schemes. As the number of antennas increases, the throughput of both the proposed scheme and the NOMA scheme increases, but the growth rate gradually slows down. Therefore, the throughput of the system can be increased by increasing the number of antennas of the energy transmitter, and the appropriate number of antennas can be selected considering the cost.

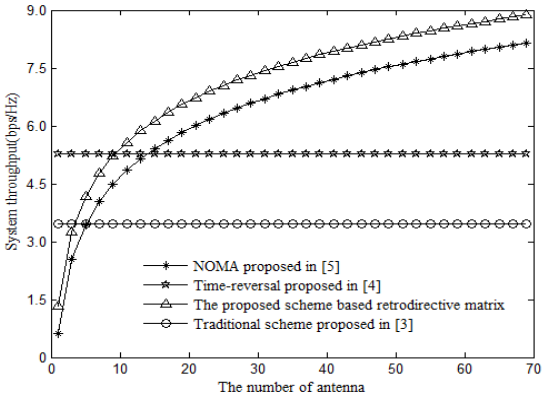

Figure 6. The relationship diagram of the number of antennas and system throughput

As shown in Figure 7, the simulation studies the relationship between beacon signal power transmitted by node equipment and the respective throughput of each node. Consider nine node devices, of which three are in a group, and the first group is in a relative position: r_1 ranges from 4 to 6 meters, $d_1 = 80$. The second group of relative positions: r_2 ranges from 9 to 11 meters, $d_2 = 75$. The relative position of the third group: r_3 is in the range of 13 to 15 meters, $d_3 = 70$.

As can be seen from the figure, under the same beacon signal power, the throughput of the node device close to the energy transmitter is higher. With the increase of beacon signal power, the throughput of all three node devices increases, but the increase of R_1 is obviously larger than that of R_2 and R_3, that is to say, the node device closer to the energy transmitter is more sensitive to beacon signal power transmitted by the node device. Therefore, the throughput of the system can be increased by increasing beacon signal power as much as possible.

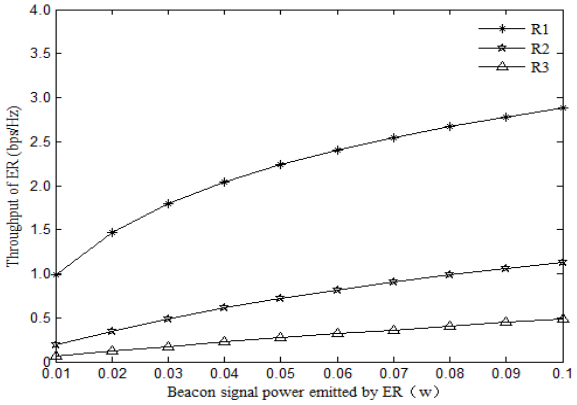

Figure 7. The relationship diagram of the node device transmits beacon signal power and the node harvests energy

As shown in figure 8, the simulation research is the influence of channel attenuation index of system throughput, can see from the picture, with the increase of channel fading index, all solutions achieve throughput decreases, on the same channel fading coefficient, based on the time inversion scheme to realize the throughput of the plan and this article is higher than other schemes, At the same time, the scheme based on time inversion is higher than the scheme in this paper, because the scheme based on time inversion adopts the time inversion technology, uses the channel reciprocity to focus the signal and restrains the interference in the information transmission stage at the same time, showing good adaptability in different environments. As the channel fading index increases, the system throughput decreases in all schemes. When $\alpha \geq 5$, as the channel attenuation index continues to increase, the throughput gap achieved by the scheme in this paper, the scheme based on time inversion and the scheme based on NOMA is not obvious. As the channel attenuation index continues to increase, the throughput decreases sharply, which can no longer meet the service demand. Therefore, in the area of low signal attenuation, the proposed scheme can achieve higher throughput than other schemes based on time inversion.

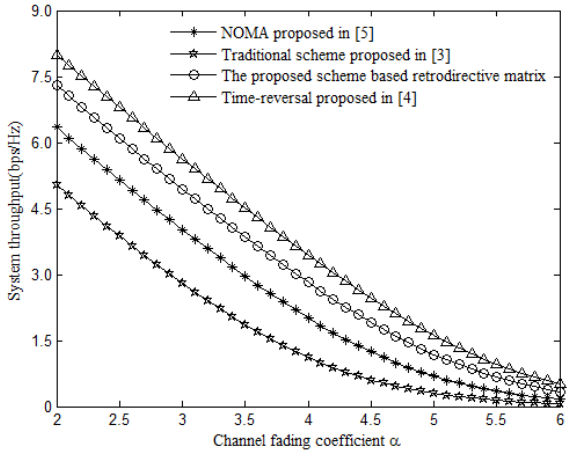

Figure 8. The relationship diagram of channel fading coefficient and system throughput

6. CONCLUSION

Improve traditional wireless power supply network model, this paper studied on the basis of the principle of phase conjugate wireless power supply network, adopt the way of energy transmitter and receiver information segregated, increases the system coverage, transmitter using MIMO antenna arrays, and the energy in the energy transfer process to generate directional beam, node equipment node equipment gain energy gain, This increases node device and system throughput. This paper also studies the maximization of system throughput under the condition of ensuring the minimum throughput of each node device, and uses the generalized Lagrange multiplier method combined with the golden section algorithm to solve the value of maximum throughput and the time allocation of each slot. This scheme has the advantages of flexible installation and low complexity of equipment. Various wireless sensor networks are suitable for indoor wireless charging in the future, such as sensors used to monitor temperature, humidity, pressure, soil ph and other data in agricultural greenhouse planting, and sensors used to monitor noise in industrial indoor. At the same time, the system energy efficiency is also an important indicator of wireless power supply network, which needs further research in the future.

ACKNOWLEDGEMENTS

This work is supported by Program for Changjiang Scholars and Innovative Research Team in University (IRT_16R72).

REFERENCE

[1] BI Suzhi, ZENG Yong, ZHANG Rui. “Wireless powered communication networks: An overview. IEEE Wireless Communications”, 2016, 23(2): 10-18.
[2] BI Suzhi, HO C K, ZHANG Rui. “Wireless powered communication: Opportunities and challenges”. IEEE Communications Magazine, 2015, 53(4): 117-125.
[3] JU H, ZHANG Rui. “Throughput maximization in wireless powered communication networks”. IEEE Transactions on Wireless Communications, 2014, 13(1): 418-428.
[4] LI Fangwei, WU Yue, NIE Yifang, et al. “Time allocation and optimization in time-reversal wireless powered communication networks”. International Journal of Antennas and Propagation, 2020, 2020: 1-8.
[5] SONG D, SHIN W, LEE J, et al. “Sum-throughput maximization in NOMA-based WPCN: A cluster-specific beamforming approach”. IEEE Internet of Things Journal, 2021, 8(13): 10543-10556.
[6] MIAO Jiansong, WANG Pengjie, ZHANG Qian, et al. “Throughput maximization for multi-UAV enabled millimeter wave WPCN: Joint time and power allocation”. China Communications, 2020, 17(10): 142-156.
[7] PARK J, LEE H, EOM S, et al. “UAV-aided wireless powered communication networks: Trajectory optimization and resource allocation for minimum throughput maximization”. IEEE Access, 2019, 7: 134978-134991.
[8] ZENG Yong, CLERCKX B, ZHANG Rui. “Communications and signals design for wireless power transmission”. IEEE Transactions on Communications, 2017, 65(5): 2264-2290.
[9] YANG Gang, HO C K, GUAN Yong liang. Dynamic resource allocation for multiple-antenna wireless power transfer. IEEE Transactions on Signal Processing, 2014, 62(14): 3565-3577.
[10] ZENG Yong, ZHANG Rui. “Optimized training design for wireless energy transfer. IEEE Transactions on Communications”, 2015, 63(2): 536-550.
[11] KASHYAP S, BJÖRNSON E, LARSSON E G. “On the feasibility of wireless energy transfer using massive antenna arrays”. IEEE Transactions on Wireless Communications, 2016, 15(5): 3466-3480.

AUTHORS

Bo Li received the B.S. degree from Shanxi Institute of Engineering and Technology, China, in 2018. Currently studying for a master’s degree in Chongqing University of Posts and Telecommunication. His major is information and communication engineering. His research interest covers Internet of Things and wireless power communication network.

Hong Tang received the B.S.degree from Sichuan University, China, in 1990 and Dr. degree from Chongqing University, China ,in 2003. He is currently a professor at Chongqing University of Posts and Telecommunications. His research interests include key technologies and implementation of mobile Internet.

© 2022 By AIRCC Publishing Corporation. This article is published under the Creative Commons Attribution (CC BY) license.

A CONTEXT-AWARE INTELLIGENT SYSTEM TO ASSIST USER PROFILEFILTERING USING AI AND DEEP LEARNING

Xinrui Que[1] and Yao Pan[2]

[1]Crean Lutheran High School, 12500 Sand Canyon Ave,
Irvine, CA 92618, USA
[2]Department of Computer Science, Vanderbilt University, USA

ABSTRACT

Community based websites such as social networks and online forums usually require users to register by providing profile information and avatars. It is important to ensure these user uploaded information comply with the website policy. This includes the information being personal, related and clear, as well as not containing unhealthy/disturbing content. A review or censorship system is usually deployed to review new user registration. Nowadays, many platforms still use manual review or rely on 3rd party APIs. However, manual review is time-consuming and costly. While 3rd party services are not tailored to the specific business needs thus do not provide enough accuracy.

In this paper, we developed an automatically new user registration review system with deep learning. We apply the state-of-art techniques such as CNN and BERT for an end-to-end evaluation system for multi-modal content. We tested our system in E-pal, a freelancing platform for gaming companionship and conducted a qualitative evaluationof the approach. The results show that our system can evaluate the quality of avatars, voice descriptions, and text profiles with high accuracy. The system can significantly reduce the effort of manual review and also provides input for the recommendation ranking.

KEYWORDS

Deep learning, Image classification, BERT, CNN.

1. INTRODUCTION

Community based websites such as social networks and online forums usually require users to register by providing profile information and avatars. It is important to ensure these user uploaded information comply with the website policy. This includes the information being personal, related and clear, as well as not containing unhealthy/disturbing content. A review or censorship system is usually deployed to review new user registration. Nowadays, many platforms still use manual review or rely on 3rd party APIs. However, manual review is time-consuming and costly. While 3rd party services are not tailored to the specific business needs thus do not provide enough accuracy.

In this paper, we focus on a specific user case: E-pal.gg [1], which is a freelancing platform for gaming companionship. But the framework can be applied to other platforms/websites as well. The E-pal community is composed of E-pal and gamer. E-pal gets commissions by playing games with other gamer or teaching others to play games. The platform, as an intermediary,

David C. Wyld et al. (Eds): NATP, ACSTY, CCCIOT, MLSC, ITCSS - 2022
pp. 65-75, 2022. CS & IT - CSCP 2022 DOI: 10.5121/csit.2022.120105

connects E-pals and gamer. When an E-pal registers an account on the platform, they are asked to use their photos as avatars, speak a language describing themselves, and write a self-descriptive text. We need to ensure that the avatar is personal and clear. There is no racial discrimination in language and writing, no sexual suggestion, and no unhealthy content.

Nowadays, many platforms still use manual review, where some contractors are hired to perform evaluation tasks and decide whether the image/text comply with policy. However, there are several limitations with manual review: 1) manual review is slow. Manual review takes significant time and adds operational cost. The member base of popular platforms can easily get to millions or even billions of users. For many start-ups, they could also face a user growth explosion where thousands of new users register every day. It could require multiple people. 2). manual review is not accurate. When a person repeats a simple action, they will get tired and make mistakes. Auditing is a process that is easy to make people tired and make mistakes. Denying qualified registration impacts the platform's user growth. Passing unqualified registration could damage company reputation and cause a serious public relation crisis. 3). Manual review is very subjective. It is easy to have personal emotions in the review and fail to maintain a consistent standard.

Besides manual review, There are companies/websites providing 3rd party image/video/audio evaluation services. For example, Amazon has a Rekognition API. These services are general purpose and train on millions of images across many categories. They work well for general image classification but they are not tailored to the specific business needs thus do not provide enough accuracy.

Open Problem: Review new user registration information in multi-modal (image/audio) accurately and scalable to enable a fast and healthy platform development.

Solution: Automatically new user registration information review with deep learning.

In this paper, we developed an automatic user registration review system with deep learning. Deep learning [2] has received great success in recent years in the domain of computer vision [3] and natural language processing [4]. AI has shown to outperform humans in various tasks such as image classification [5], question answering and reading comprehension [6]. The advancement of deep learning comes from both algorithm improvement such as more advanced network structure, as well as hardware advancement, where the scale and parameters of the model can grow very large.

One particular important advancement for deep learning is transfer learning. Where a model is trained on a large number of labeled data first. The model will learn some general patterns and is then fine-tuned on a specific task.

For our new registration review tasks, we are dealing with multi-modal data from image, text to audio. We apply the state-of-art technique such as MobileNet and BERT to learn an embedding first and then use supervised learning to fine-tune it on our specific task.

Compared to manual review, AI based review systems are more scalable since the system can be easily extended by leveraging cloud computing service. The AI system is also more subjective since it has a unified standard.

In order to evaluate our system, we conducted quantitative experiments to evaluate the effectiveness and accuracy of our proposed solution. First, we experiment with the profile image evaluation where the task is 1) Detect if the image contains any sexual/unhealthful content 2)

detect if the image contains a face and the quality of the image in terms of lighting, clearness, Aesthetic, etc. Our system will generate a score from 0 to 100 as well as a reason. The score was compared with a manually reviewed score to calculate the mismatch.

Second, we experiment with audio intro evaluation. The task is 1) Evaluate the audio quality in terms of noise level, speech speed, clearness, appeal/aesthetic value. 2) Detect if the speech content is relevant to the context and does not contain any racial/offensive language. In the end, We also picked some samples for qualitative analysis and demonstrated the effectiveness of the system.

The rest of the paper is organized as follows: Section 2 gives the details on the challenges thatwe met during the experiment and designing the sample; Section 3 focuses on the details of our solutions corresponding to the challenges that we mentioned in Section 2; Section 4 presents the relevant details about the experiment we did, following by presenting the related work in Section 5. Finally, Section 6 gives the conclusion remarks, as well as pointing out the future work of this project.

2. CHALLENGES

Currently, we rely on humans to manually review users' uploaded profiles to determine if they are eligible. However, there are several challenges with the existing approach.

2.1. Review needs to be scalable and fast to keep up with the growth of the platform

Manual review takes significant time and adds operational cost. The member base of popular platforms can easily get to millions or even billions of users. For many start-ups, they could also face a user growth explosion where thousands of new users register every day. It could require multiple people.

Also it's hard to predict the user growth. If we hire more people but not enough registration, we add unnecessary operational cost, Or if the user growth outpaces the human review, the review will be delayed and lead to poor member experience.

2.2. Review needs to be accurate

We need to minimize the probability of denying qualified registration as well as passing unqualified registration. Manual review is not accurate as humans make mistakes. Review is a repetitive and tedious process where it's easy for a person to get tired and make mistakes. Denying qualified registration impacts the platform's user growth. Passing unqualified registration could damage company reputation and cause a serious public relation crisis.

2.3. Review needs to be subjective and maintain a consistent standard

Manual review is very subjective and the criteria differs from person to person. On one hand, this inconsistency creates inaccuracy, on the other hand, this could also confuse users if their registration is rejected but similar or even worse registration has passed.

3. SOLUTION

The overall system consists of a front end which is a website built with HTML/JSS, and a back-end which is powered by Flask and Tensorflow. Users will be able to upload images or audio file and the evaluation results will be given for each uploaded item.

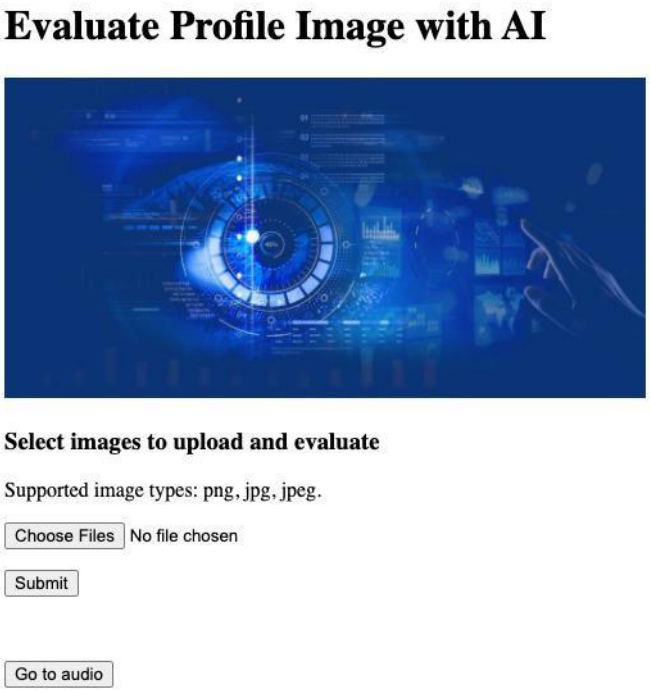

Figure 1. Overview of the system

The system consists of two sub-systems: profile image scoring system and audio introduction scoring system.

Profile image scoring system. The system has the following components:

1. Image pre-process. In this step, images of various formats and dimensions will need to be converted to a unified format to facilitate following processing.
2. Nudity Content detection. To ensure a healthy platform environment, images with explicit sexual content are not allowed. We utilize an existing library NudeNet (https://github.com/notAI-tech/NudeNet) to help us with nudity censoring.
3. Face detection. Required by the platform, profile images should contain the selfie of the user himself. Sometimes users will upload arbitrary images of landscape, object or cartoon. We perform face detection to distinguish those that have human faces versus those that don't.
4. Image quality evaluation. This subsystem is to evaluate the quality of the profile image. The quality here refers to lighting, clearness, Aesthetic, etc. Here we use a pretrained model efficientnetv2-s to get the embedding of the image. Then we feed the embedding into a deep neural network to train a regression model.

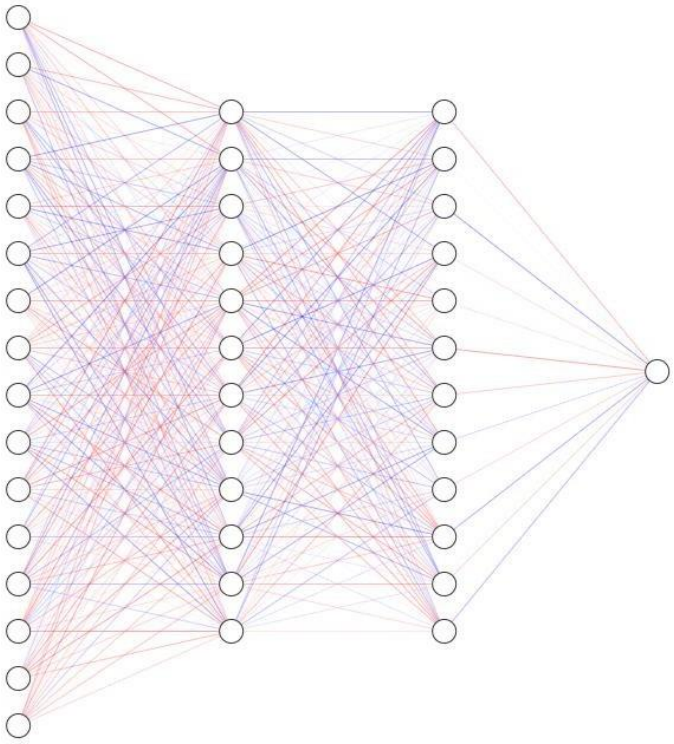

Figure 2. Image classification

Deep learning has received great success in recent years in various tasks such as image classification, object detection, etc. One important advancement is transfer learning. Where a model is trained on a large number of labeled data first. This is very helpful as image models can easily have millions of parameters and require a lot of computing power to rain. The model will learn some general patterns and is then fine-tuned on a specific task. In our case, efficientNet is trained on ImageNet for general image classification. The embedding is then used to fine-tune for profile image scoring.

We rely on efficient v2-s (EfficientNetV2: Smaller Models and Faster Training) to learn an embedding of image. EfficientNet is a new family of CNN with faster speed and better parameter efficiency. It was able to achieve comparable performance compared to some large scale models while remaining small in size and fast serving speed.

Our neural network has 4 layers.

1. The first layer is input layer: it has 384*384*3 dimensions. Where 384*384 is the image sizeand 3 is the color channel.
2. The second layer is the embedding layer. efficient v2-s will return an embedding of dimension1280.
3. The third layer is the dropout layer. It is used for preventing overfitting. A certain percentageof connection is randomly dropped during training.
4. The fourth layer is the output layer. It has a dimension of 1. It outputs a score of 0-100.

Audio introduction scoring system. The system has the following components:

1. Audio preprocess. In this step, images of various format dimensions will need to be converted to a unified format to facilitate following processing. We use ffmpeg which is a widely used audio conversion tool.
2. Speech recognition. We rely on Google cloud Speech recognition API.
3. Text evaluation. We trained a Bert based model.
4. Audio evaluation. This subsystem is to evaluate the quality of the intro audio. The quality here refers to noise level, speech speed, clearness, appeal/aesthetic value.

For the text-based evaluation, we build a neural network based on BERT [7] embedding. BERT (Bidirectional Encoder Representations from Transformers) is a transformer based model which has been widely successful on a variety of NLP tasks such as text classification, sentiment analysis, questions answering, machine translation, etc [8].

The text from the speech recognition API is fed into the input of the neural network. Then there is a pre processing layer which computes the input_words_id, input_mask and input_type_ids. The three inputs are passed to the BERT encoder, and output a 1024 length vector. We then include a 512 length dense layer with relu activation function. Anally an output layer of length 2. The network diagram is shown in Fig 3.

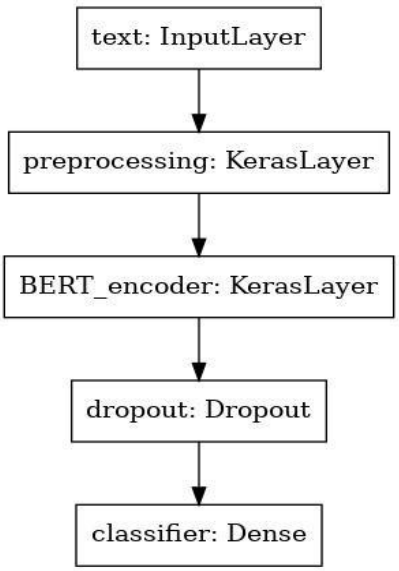

Figure 3. Network diagram

Although the BERT model can capture the content of the audio and determine whether it is relevant, it cannot capture other factors of audio such as whether the voice is clear, speech speed is appropriate or the voice is appealing. So we build a second neural network to evaluate the voice features of the audio. We use YAMNET [9] to generate embedding of the audio and usethe embedding to train a classifier to predict whether the audio is high quality or low quality.

YAMNET is a deep network that predicts audio from 521 classes. It returns a score vector indicating the probability for each of the 521 classes. It also returns an embedding of shape (N, 1024) where N is the number of 0.96 second frames. We conducted an average-pooling to get a vector of 1024.

4. EXPERIMENT

In this section, we described experiments to evaluate the effectiveness and accuracy of our proposed solution.

We collected 2137 real user profile avatar and audio from E-pal. Both image and audio are manually reviewed and given a score from 0 to 100. The score distribution is image and audio are given in Fig 4 and 5. We use 80% of the data as training and the remaining 20% as evaluation.

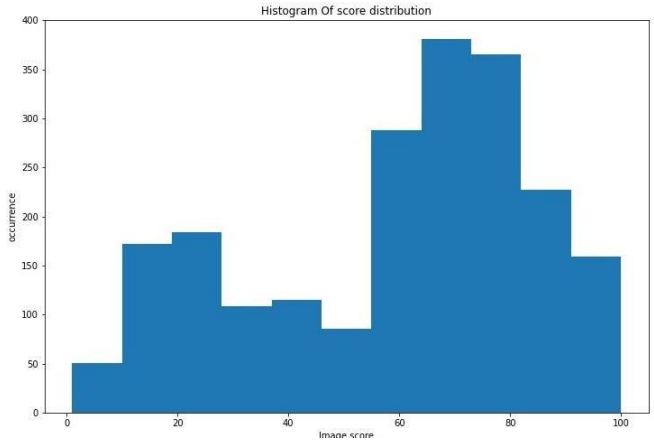

Figure 4. Histogram of image score distribution

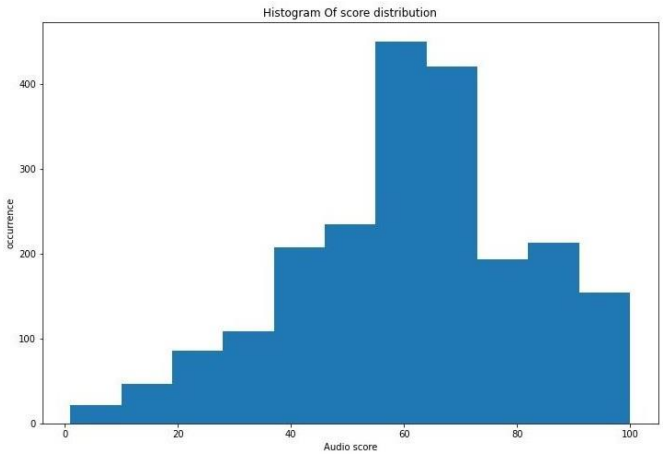

Figure 5. Histogram of audio score distribution

1. Evaluate Profile Image Scoring System.

First, we experiment with the profile image evaluation where the tasks is 1) Detect if the image contain any sexual/unhealthy content 2) detect if the image contain face and the quality of the image in terms of lighting, clearness, Aesthetic, etc. Our system will generate a score from 0 to 100 as well as a reason. The score was compared with a manually reviewed score to calculate the mismatch.

We use two metrics to evaluate our results. 1. mean absolute error (MAE). MAE is defined as

100ni=1n|Ai-Fi|. At is the actual score and Fi is the predicted score. This is to get a sense of how close the predicted score is compared to actual score. 2. Accuracy. Here we treat the problem as a classification problem where we only have 3 classes: High quality (score above 70), medium quality (40-70) and low quality (score below 40). The metric is used to give a rough estimation of the evaluation quality.

The hyper-parameters are set as follows:

learning rate = 0.005, momentum =0.9, L2 regularization = 0.0001, batch size =16, dropout rate =0.2.

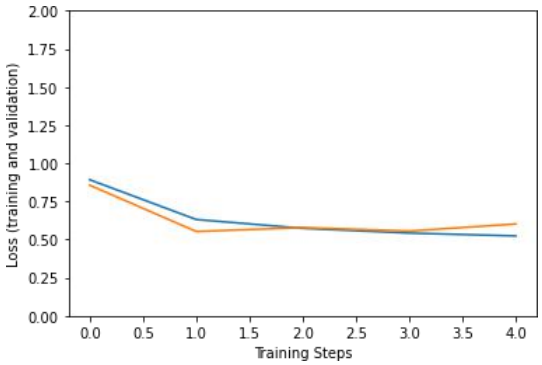

Figure 6. Loss vs. Training steps

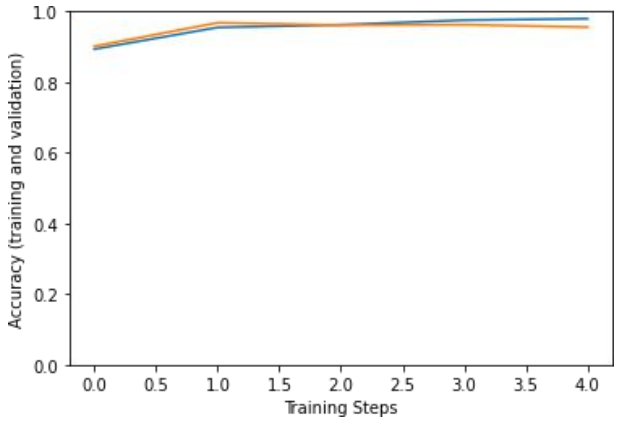

Figure 7. Accuracy vs. Training steps

Table 1. Performance of image evaluation on different embedding

Image embedding	MAE	Accuracy
Efficientnet v2-s	12.4	91.2%
Efficientnet v2-m	12.2	91.5%

In table 1, we compare the performance of two embedding (efficient netv2-m has a larger dimension than efficient netv2-s). We can see that both can accurately evaluate the image. The absolute score difference is around 12 and over 90% of the samples are correctly classified into the high/medium/low classes. Having a larger embedding model slightly improves the performance, but also at the cost of longer training and scoring execution time.

We also look at the images where the predicted score and labeled score differs the most. We found this most happened for certain cartoon images where the cartoon character is very human. The algorithm tends to give them higher scores while human reviewers give lower scoresbecause they are not the photo of users. One of our future work would be to improve the algorithm to better distinguish cartoon characters.

We also tested the performance of scoring with the train model. We deployed the tensor-flow model on an Amazon ec2 t3.large instance. The average scoring time is 0.83s, which means that it can score over 100000 images per day. And the cost is only 0.0832*24=$2 per day, which is much lower than human reviewers.

2. Evaluate Audio Intro Scoring System.

For the audio intro evaluation, the task is 1) Evaluate the audio quality in terms of noise level, speech speed, clearness, appeal/aesthetic value. 2) Detect if the speech content is relevant to the context and does not contain any racial/offensive language.

For the BERT model, the hyperparameters are set as follows:

learning rate = 2e-5, batch size = 32, max sequence length = 64, epsilon = 1E-8.For audio neural network, the hyperparameters are set as follows:

learning rate = 0.005, momentum =0.9, L2 regularization = 0.0001, batch size =16, dropout rate =0.2.

Table 2. Performance of audio evaluation on different training sample size

Training samples	MAE	Accuracy
500	16.4	83.9%
1500	14.1	85.3%

We can see that the proposed solution can also accurately evaluate the audio. The absolute score difference is around 14 and over 85% of the samples are correctly classified into the high/medium/low classes. The performance improves as we increase the training sample size (from 500 to 1500). If more training data is available, the accuracy could be further improved.

5. RELATED WORK

Face detection is a topic that has been widely discussed in computer vision over the past few decades. Viola. et al [10] propose a detection framework based on Haar features and Adaboost classifier. In recent years, deep learning has achieved great success in many computer vision tasks. Deep convolutional neural net based methods [11] [12] have outperformed traditional machine learning methods in face detection in both the accuracy and ease of use.

Image aesthetic assessment [13] aims to computationally distinguish high-quality image from low-quality image based on photographic rules. Different approaches have been developed, some based on hand-crafted features [14] and some based on deep features [15].

Speech or Audio quality assessment has been studied by several researchers [16]. Some rely on signal-to-noise ratio measures. Some rely on spectral distance measures. However, these low

level features cannot capture the semantic meaning of the speech. Text classification is the tasks of classifying text into multiple classes based on the semantic meaning. The applications range from email spam classification [17], sentiment classification, There are companies/websites providing 3rd party image/video/audio evaluation services. For example, Amazon has Rekognition API [18]. These services are general purpose and train on millions of images across many categories. They work well for general image classification but are not tailored to specific business needs.

6. CONCLUSIONS

In this paper, we proposed an automatically new user registration review system with deep learning. We apply the state-of-art techniques such as CNN and BERT for an end-to-end evaluation system for multi-modal content. The system can be used in various Community based websites such as social networks and online forums. We conducted an experiment using real world profile data from E-pal, a freelancing platform for gaming companionship. The results indicate deep learning can accurately classify low quality profile vs. high quality profile.

There are still some limitations with the current approach. The algorithm generally performs well at identifying low quality or high quality input, but accuracy can become lower for border line cases. Also, there are some corner cases not well covered by the current algorithm. For example, cartoon characters are sometimes still detected as human faces.

For future work, we plan to continue to improve the algorithm accuracy by cleaning andobtaining more training data, experiment with more complex models and improve the explainability of the scoring results.

REFERENCES

[1] E-pal. https://www.epal.gg/
[2] Goodfellow, Ian, Yoshua Bengio, and Aaron Courville. Deep learning. MIT press, 2016.
[3] Voulodimos, Athanasios, Nikolaos Doulamis, Anastasios Doulamis, and Eftychios Protopapadakis. "Deep learning for computer vision: A brief review." Computational intelligence and neuroscience 2018 (2018).
[4] Young, Tom, Devamanyu Hazarika, Soujanya Poria, and Erik Cambria. "Recent trends in deep learning based natural language processing." ieee Computational intelligenCe magazine 13, no. 3 (2018): 55-75.
[5] He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Delving deep into rectifiers: Surpassing human- level performance on imagenet classification." In Proceedings of the IEEE international conference on computer vision, pp. 1026-1034. 2015.
[6] Raffel, Colin, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. "Exploring the limits of transfer learning with a unified text-to-text transformer." arXiv preprint arXiv:1910.10683 (2019).
[7] Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805 (2018).
[8] Young, Tom, et al. "Recent trends in deep learning based natural language processing." ieee Computational intelligence magazine 13.3 (2018): 55-75.
[9] https://tfhub.dev/google/yamnet/1
[10] Viola, Paul, and Michael Jones. "Rapid object detection using a boosted cascade of simple features." In Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001,vol. 1, pp. I-I. Ieee, 2001.
[11] Albawi, Saad, Tareq Abed Mohammed, and Saad Al-Zawi. "Understanding of a convolutional neural network." In 2017 International Conference on Engineering and Technology (ICET), pp. 1-6. Ieee, 2017.
[12] Sun, Xudong, Pengcheng Wu, and Steven CH Hoi. "Face detection using deep learning: An improved

fasterRCNN approach." Neurocomputing 299 (2018): 42-50.
[13] Deng, Yubin, Chen Change Loy, and Xiaoou Tang. "Image aesthetic assessment: An experimental survey."IEEE Signal Processing Magazine 34, no. 4 (2017): 80-106.
[14] Datta, Ritendra, Dhiraj Joshi, Jia Li, and James Z. Wang. "Studying aesthetics in photographic images usinga computational approach." In European conference on computer vision, pp. 288-301. Springer, Berlin, Heidelberg, 2006.
[15] Peng, Kuan-Chuan, and Tsuhan Chen. "Toward correlating and solving abstract tasks using convolutional neural networks." In 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), pp. 1-9. IEEE, 2016.
[16] Loizou, Philipos C. "Speech quality assessment." In Multimedia analysis, processing and communications, pp. 623-654. Springer, Berlin, Heidelberg, 2011.
[17] Abdulhamid, Shafi'I. Muhammad, Maryam Shuaib, Oluwafemi Osho, Idris Ismaila, and John K. Alhassan. "Comparative Analysis of Classification Algorithms for Email Spam Detection." International Journal of Computer Network & Information Security 10, no. 1 (2018).
[18] Amazon Rekognition. https://aws.amazon.com/rekognition/

© 2022 By AIRCC Publishing Corporation. This article is published under the Creative Commons Attribution (CC BY) license.

SHORTCOMINGS OF THE FUNDAMENTAL MATRIX EQUATION TO RECONSTRUCT 3D SCENES

Tayeb Basta

College of Engineering and Computing, Al Ghurair University, Dubai, UAE

ABSTRACT

In stereo vision, the epipolar geometry is the intrinsic projective geometry between the two views. The essential and fundamental matrices relate corresponding points in stereo images. The essential matrix describes the geometry when the used cameras are calibrated, and the fundamental matrix expresses the geometry when the cameras are uncalibrated. Since the nineties, researchers devoted a lot of effort to estimating the fundamental matrix. Although it is a landmark of computer vision, in the current work, three derivations of the essential and fundamental matrices have been revised. The Longuet-Higgins' derivation of the essential matrix where the author draws a mapping between the position vectors of a 3D point; however, the one-to-one feature of that mapping is lost when he changed it to a relation between the image points. In the two other derivations, we demonstrate that the authors established a mapping between the image points through the misuse of mathematics.

KEYWORDS

Fundamental Matrix, Essential Matrix, Stereo Vision, 3D Reconstruction.

1. INTRODUCTION

In computer stereo vision, the 3D object shape reconstruction from two 2d images can be defined as follows:

The object to be reconstructed is a set of 3D points M, it is depicted by two cameras from two different standpoints. Left and right coordinate systems are defined in each of these standpoints. And every 3D point is projected on the left and right images as two 2D points m_l and m_r, respectively.

The epipolar geometry is the intrinsic projective geometry between the two views. It is independent of scene structure, and only depends on the cameras' internal parameters and relative pose. The fundamental matrix F encapsulates this intrinsic geometry [1].

A 3D point M is represented in the left and right coordinate systems by two position vectors $M_l = [X_l \quad Y_l \quad Z_l]^T$ and $M_r = [X_r \quad Y_r \quad Z_r]^T$. And $m_l = [x_l \quad y_l]^T$ and $m_r = [x_r \quad y_r]^T$ are the position vectors of the projective points m_l and m_r in the left and right coordinate systems, respectively, as in Figure 1.

3D shape reconstruction is performed in the following steps [1]

1. Compute the fundamental matrix from point correspondences.

David C. Wyld et al. (Eds): NATP, ACSTY, CCCIOT, MLSC, ITCSS - 2022
pp. 77-86, 2022. CS & IT - CSCP 2022 DOI: 10.5121/csit.2022.120106

2. Compute the camera matrices from the fundamental matrix.
3. For each point correspondence $m_l \leftrightarrow m_r$, compute the point in space that projects to these two image points.

Thus, the first step is to compute the fundamental matrix and the eight-point algorithm is the most used method to do so. In practice the number of image points is large; so, the fundamental matrix can only be estimated rather calculated. Researchers keep developing methods that overcome previously devised ones in terms of accuracy and mitigating noise effects. Only few researchers thought that the bad performance of the eight-point algorithm would requires the revision of the projective geometry approach itself.

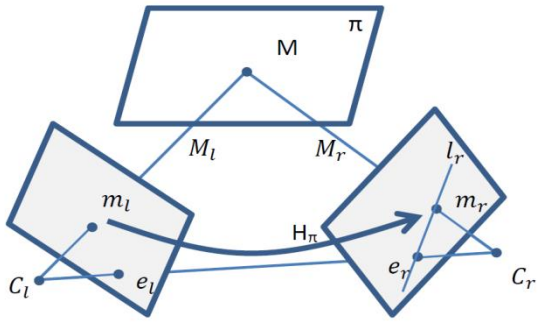

Figure1. The epipolar geometry. A point m_l in one image is transferred via the plane π to a matching point m_r in the second image. The epipolar line l_r through m_r is obtained by joining m_r to the epipole e_r.

The main objective of the current work is to revise the theory underpinning the derivation of the essential and fundamental matrices equations. Thus, clarify the reason behind the bad performance of the projective geometry application to 3D reconstruction from 2D views.

The rest of the paper is organized as follows: Section 2 introduces the motivation of addressing a classic problem like the fundamental matrix of stereo vision. Sections 3 exposes some related work. Section 4 demonstrates the shortcoming of the essential matrix equation. Section 5 shows the mathematical flaws of two derivations of the fundamental matrix. And the paper concludes in section 6.

2. WHY SHOULD WE ADDRESS SUCH A CLASSIC PROBLEM?

The epipolar geometry application in computer stereo vision represented by the fundamental matrix is still part of computer vision courses in most universities around the world. On top of that, researchers are still spending time to develop methods to estimating the fundamental matrix [2, 3, 4, 5, 6]. Table 1 shows a sample of outstanding universities with links to their computer vision courses that include at least one chapter on epipolar geometry and the fundamental matrix.

Table 1 Sample universities teaching the epipolar geometry to reconstruct 3D shape from two views.

University	Course Title	Course Link
Stanford University, USA	Computer Vision, From 3D Reconstruction to Recognition	http//web.stanford.edu/class/cs231a/syllabus.html
The University of Washington, USA	Computer Vision	https//courses.cs.washington.edu/courses/cse455/
MIT, USA	Computer Vision and Applications	www.ai.mit.edu/courses/6.891/lectnotes/lect8/lect8-slides.pdf

University College London, UK	Machine Vision	https//www.ucl.ac.uk/module-catalogue/modules/machine-vision/COMP0137
University of Toronto, Canada	Foundations of Computational Vision	http//www.cs.toronto.edu/~kyros/courses/2503
Tokyo Institute of Technology, Japan	Computer Vision	http//www.ocw.titech.ac.jp/index.php?module=General&action=T0300&JWC=201804591&lang=EN&vid=03
Sorbonne Université - Télécom Paris	Master Informatique - Parcours IMA	https//perso.telecom-paristech.fr/bloch/P6Image/VISION.html

3. RELATED WORK

Though the fundamental matrix theory is considered as a landmark achievement of computer vision, certain researchers called it into question [7, 8, 9,10, 11, 12]. In a series of research work, Basta demonstrated that many of the derivation methods of the essential and fundamental matrix equations are flawed [13, 14, 15, 16, 17, 18, 19].

In [17] and [19], the author presented extensive experimental results of two real images of a building captured from two standpoints. The building (Figure 2) is composed of two parts with different depths with respect to the camera lens. In [17], the author used a MATLAB Toolbox [20] that contains several methods for estimating the fundamental matrix using the eight-point algorithm. In [19], he implemented the solution in Python and used the findFundamentalMat() function of the cvonline package to estimate the fundamental matrix.

Figure 2. The building image used to estimate the fundamental matrix in [17] and [19].

In both works [17] and [19], the author estimated the fundamental matrix that satisfies the equation $m_r^T F m_l = 0$. Then, he calculated the values of the expression $m_r^T F m_l$ for several pairs of corresponding points (m_l, m_r). Such values are supposed to be equal to zero. The matrix F is calculated from different regions of the images (whole images, back part of the images, and front side of the images) and the pairs of corresponding points are selected arbitrarily from the images. Table 2 shows that the values of $m_r^T F m_l$ are sometimes very far away from 0; greater than 10 for some cases.

Table 2. the values of the expression $m_r^T F m_l$ calculated for selected points from the whole images, the back side, and the front side of the images. As it is apparent the image is composed of components with different depth with respect to the camera lens. This result is published in [19].

F matrix calculated from		
Whole	Back	Front
0.322	0.121	-0.504

0.084	1.496	0.557
-0.026	0.545	0.684
0.234	3.978	0.748
0.328	7.314	-0.726
0.135	16.158	-0.508
-0.165	9.001	-0.784
0.184	13.800	2.989
0.070	12.401	-0.109
0.135	10.794	-1.970

In the current work, three main publications where the essential and fundamental matrices are derived as a product of a skew matrix and a rotation transformation matrix are scrutinized. One of these is where the first time the essential matrix introduced to the computer vision community by Longuet-Higgins [21]. Next section shows how Longuet-Higgins succeeded in securing a one-to-one mapping between the position vectors of world points of a scene and that mapping is lost when he transformed it to a relation between the image points. In the other two derivations, the authors try to directly establish a one-to-one relation between the image points. Such a relation is represented by the fundamental matrix. The current work shows the mathematical flaws in these two derivations.

4. LONGUET-HIGGINS' DERIVATION OF THE ESSENTIAL MATRIX

4.1. The Equation Derivation

In [21], Longuet-Higgins created a matrix $Q = RS$ where $S = \begin{bmatrix} 0 & t_3 & -t_2 \\ -t_3 & 0 & t_1 \\ t_2 & -t_1 & 0 \end{bmatrix}$. The matrix R and the vector t are the rotation and translation of the right coordinate system with respect to the left coordinate system. M_l and M_r are the position vectors of a world point M on the left and right coordinate systems, respectively. The author formed the expression $M_r^T Q M_l$ and after some arithmetic manipulations he found out that

$$M_r^T Q M_l = 0 \tag{1}$$

For every 3D point there are exactly two position vectors; one represents that point in the left coordinate system and the other represents the point in the right coordinate system. Thus, Q in (1) is a one-to-one mapping between M_l and M_r.

In terms of coordinates, $M_l = (X_l, Y_l, Z_l)$ and $M_r = (X_r, Y_r, Z_r)$. And the coordinates of the projective points m_l and m_r of the point M in the left and right coordinate systems, respectively are

$$\begin{aligned} m_l &= (X_l/Z_l\,, Y_l/Z_l\,, 1) \\ m_r &= (X_r/Z_r\,, Y_r/Z_r\,, 1) \end{aligned} \tag{2}$$

Finally, the author divided the left-hand side of (1) by $Z_l Z_r$ to conclude the essential matrix equation

$$m_r^T E m_l = 0 \tag{3}$$

4.2. Shortcoming of Longuet-Higgins's derivation

Longuet-Higgins approached the problem from an algebraic perspective, he used matrix product as the main operation to derive the essential matrix equation. So, he has not been faced with the problem of transformation from one coordinate system to the other.

He formed the expression $M_r^T Q M_l$. And because the matrix product is an associative operation, the expression $M_r^T Q M_l$ is the product of 1×3 row matrix and a 3×3 matrix and a 3×1 column matrix which led to equation (1).

The problem of Longuet-Higgins' derivation started when he divided equation (1) by $Z_l Z_r$. As it is known, the position vector of a point is the unique vector from the origin of the coordinate system to the point itself. So, for every point M, equation (1) holds for exactly two position vectors M_l and M_r in the left and right coordinate systems, respectively. Once (1) is divided by $Z_l Z_r$ we will get the following equation

$$\frac{M_r^T}{Z_r} \cdot Q \cdot \frac{M_l}{Z_l} = 0 \qquad (4)$$

Where $m_l = \frac{M_l}{Z_l}$ and $m_r = \frac{M_r^T}{Z_r}$ are the projection of the vectors M_l and M_r on the left and right camera planes, respectively.

In projective geometry, m_l could be the projection of a single world point or multiple world points (Figure 3). It is the projection of all world points laying on the ray drawn from the camera lens centre to the point M.

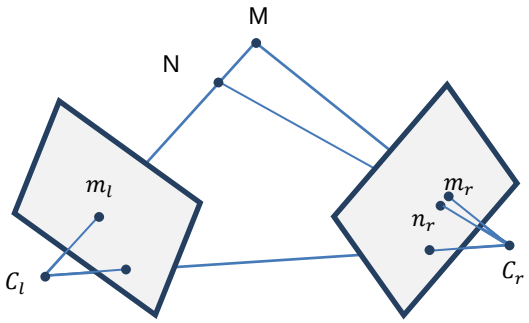

Figure 3. the image point m_l is the projection of two world points M and N. m_l is a corresponding point to two image points m_r and n_r.

Furthermore, there are world points visible to one camera and invisible to the other. This could be because these points are hidden by 3D objects in the scene. This is one of the characteristics of 3D scenes. So, they are projected on the first camera plane and does not have an image on the other camera. However, when you plug this image point into m_l or m_r and solve equation (3), you will get a false corresponding point.

Recall the 3D shape reconstruction as described in [1]

1. Compute the fundamental (essential) matrix from point correspondences.
2. Compute the camera matrices from the fundamental matrix.

3. For each point correspondence $m_l \leftrightarrow m_r$, compute the point in space that projects to these two image points.

Assuming the point $p = (1,2,1)$ is on the left camera plane (image). The matrix E is already calculated or estimated. To compute the corresponding point of p, we plug the value of p into equation (3).

$$[x_r \quad y_r \quad 1]E[1 \quad 2 \quad 1]^T = 0, E = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix} \tag{5}$$

Substituting for the matrix E, we will get the following equation

$$[x_r \quad y_r \quad 1]\begin{bmatrix} A_1 \\ A_2 \\ A_3 \end{bmatrix} = 0 \text{ , where} \begin{bmatrix} A_1 \\ A_2 \\ A_3 \end{bmatrix} = \begin{bmatrix} a_{11} + 2a_{12} + a_{13} \\ a_{21} + 2a_{22} + a_{23} \\ a_{31} + 2a_{32} + a_{33} \end{bmatrix} \tag{6}$$

which leads to the following equation

$$A_1 x_r + A_2 y_r + A_3 = 0 \tag{7}$$

There are infinite values of (x_r, y_r) satisfying equation (7). Geometrically, this means that any point p has many corresponding points. Which is incorrect; the certainty is each image point has at most one corresponding point in each other image except the case of occlusion when two different points have the same corresponding point.

Consequently, the essential (fundamental) matrix equation does not ensure the recovery of the right shapes of 3D scenes.

5. ESTABLISHING A DIRECT MAPPING BETWEEN THE IMAGE POINTS

Because the above essential matrix derivation suffers from the drawback of an image point can have unlimited number of corresponding points, computer vision researchers try to directly draw a mapping between the image points without passing through position vectors of the 3D point. The next sections explore the flaws of two well-known derivations of the essential and fundamental matrices equations.

5.1. Luong-Faugeras derivation of the essential matrix

In [22], Luong et al. assert that because the vector from the first camera optical centre to the first imaged point m_l, the vector from the second optical centre to the second imaged point m_r, and the vector from one optical center to the other t are all coplanar. In normalized coordinates, this constraint can be expressed simply as

$$m_r^T (t \times R m_l) = 0 \tag{8}$$

where R and t capture the rotation and translation of the right cameras coordinate system with respect to the left one. In [23], Birchfield explicitly stated that the multiplication by R is necessary to transform m_l into the second camera's coordinate system. The authors [22] defined $[t]_\times$ as the matrix such that $[t]_\times y = t \times y$ for any vector y, and they rewrite equation (8) as a linear equation

$$m_r^T\ ([t]_\times R m_l\ ) = m_r^T E m_l = 0, \tag{9}$$

Where $E = [t]_\times R$ is called the Essential matrix and $[t]_\times = \begin{bmatrix} 0 & -t_3 & t_2 \\ t_3 & 0 & -t_1 \\ -t_2 & t_1 & 0 \end{bmatrix}$.

5.2. The flaw in Luong-Faugeras derivation

Let us closely examine equation (8), $m_r^T\ (t \times R m_l) = 0$.

We have the following facts. The point m_l is on the left image, so the vector m_l is defined in the left coordinate system and not defined in the right one. The point m_r is on the right image, then the vector m_r is defined in the right coordinate system and not defined in the left one. And the vector t, the translation of origin of the right coordinate system with respect to the left coordinate system; so, t is defined in the left coordinate system and not defined in the right one.

The left hand-side of (8) consists of three vector operations. The term inside the parenthesis is evaluated first which includes a vector product and a matrix product.

Let assume that Rm_l is to be evaluated first; it is the product of a rotation transformation matrix and a vector. So, $v = Rm_l$ is the vector m_l expressed in the right coordinate system. Therefore $t \times Rm_l = t \times v$ is the cross product of t defined in the left coordinate system and v defined in the right coordinate system. Thus, $t \times Rm_l$ is the cross product of two vectors not defined in the same coordinate systems; so, it is invalid.

Now, let us consider that the cross-product operation $t \times R$ is to be evaluated first.

DEFINITION

The cross product (or vector product) of two vectors

$x = \langle x_1, x_2, x_3 \rangle$ and $y = \langle y_1, y_2, y_3 \rangle$ in $\mathbb{R}^3$ is the vector $x \times y = \langle x_2 y_3 - x_3 y_2\ , x_3 y_1 - x_1 y_3\ , x_1 y_2 - x_2 y_1 \rangle$.

The cross product of two vectors x and y in $\mathbb{R}^3$ is a vector orthogonal to both x and y [24].

The cross product of a 3D vector and a 3×3 matrix is UNDEFINED [24].

Therefore, there is no operation called cross product of a vector and a matrix; therefore, the term $t \times R$ is undefined. Thus, equation (8) that is the premise of the current derivation of the essential matrix is invalid. And the current derivation of the essential matrix is flawed.

One could claim that Rm_l is a product of a matrix and a vector which produces a vector defined in the same coordinate system. Then the cross-product $t \times Rm_l$ is a vector defined in the left coordinate system. in this case, $m_r^T \cdot\ (t \times Rm_l)$ is a dot product of two vectors, one from the right coordinate system and the other from the left coordinate system. it is an undefined operation.

5.3. Hartley-Zisserman derivation of the fundamental matrix

In the geometric derivation of the fundamental matrix equation, the authors [1] assert the existence of 2D homography H_π mapping each point m_l from the left image to a point m_r on the right image, because the set of all such points m_l in the left image and the corresponding points m_r in the right image are projectively equivalent, since they are each projectively equivalent to the planar point set M (Figure 1). Thus, $F = [\mathrm{e}_r]_\times H_\pi$ that is a matrix product of a skew matrix and a transformation from left to right.

5.4. The flaw in Hartley-Zisserman derivation

The points M in the above statement are the world points of the 3D scene to be reconstructed from a pair of its images. If the 3D scene is planar, why are we constructing a planar scene from two of its planar images in the first place. Thus, the existence of a homography mapping points of the left image to points on the right image is on condition that the 3D scene is planar. Because typical 3D scenes might contain objects with different depths (i.e., distance from the camera centre). So, some points on these objects can be visible to one camera and hidden from the other. Therefore, some image points on the left camera plane will not have corresponding points on the right camera plane and points on the right image will not have corresponding points on the left image. Furthermore, researchers recognize the existence of the occlusion problem [25] where two 3D points or more are projected onto the same image point as in Figure 3. At the same time, they assert the existence of a homography between points of the left image and those on the right image. These facts, confirm that points on the left and right images are not projectively equivalent and no homography exists between them. In conclusion, the expression $F = [\mathrm{e}_r]_\times H_\pi$ where H_π is a homography is irrational.

6. CONCLUSION

In this work, we demonstrated that the first ever derivation of the essential matrix that has been introduced to the computer vision community is free of flaws; however, it does not ensure a one-to-one mapping between the image points of the two views. Later, researchers tried to address such shortcoming through deriving the essential and fundamental matrices equation as a mapping between the image points. We showed that two of the well-known of these derivations are mathematically flawed.

The current work establishes a rigorous scrutiny of a theory that claims to be mathematically founded. The trend for solving computer vision problems uses machine learning tools to obtain good solutions without requiring any mathematical basis.

REFERENCES

[1] Richard Hartley & Andrew Zisserman, (2004). *Multiple View Geometry in Computer Vision*, 2nd edn. Cambridge University Press, chapter 9.

[2] Omid Poursaeed, Guandao Yang, Aditya Prakash, Qiuren Fang, Hanqing Jiang, Bharath Hariharan & Serge Belongie, (2018) "Deep fundamental matrix estimation without correspondences", In *Proceedings of the European conference on computer vision workshop*, pp1–13.

[3] Daniel Barath, (2018) "Five-Point Fundamental Matrix Estimation for Uncalibrated Cameras", *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (CVPR), pp235-243.

[4] Yi Zhou, Laurent Kneip & Hongdong Li, (2015) "A revisit of methods for determining the fundamental matrix with planes", In 2015 *International conference on digital image computing techniques and applications* (DICTA), pp 1–7. *IEEE*. http//ieeexplore.ieee.org/document/7371221/.

[5] Sushil Pratap Bharati, Feng Cen, Ajay Sharda & Guanghui Wang, (2018) "RES-Q Robust Outlier Detection Algorithm for Fundamental Matrix Estimation", *IEEE Access*, 6, 48664–48674.

[6] Nurollah Tatar & Hossein Arefi, (2019) "Stereo rectification of pushbroom satellite images by robustly estimating the fundamental matrix", *International Journal of Remote Sensing*, 40(23), 8879-8898.

[7] Andrew Zisserman & Stephen Maybank, (1993) "A Case Against Epipolar Geometry", In *2nd Europe-U.S. Workshop on Invariance*, pp 69–88. Ponta Delgada, Azores.

[8] Richard Hartley, (1997) "In Defense of the Eight-Point Algorithm", *IEEE Transaction on Pattern Recognition and Machine Intelligence*, 19(6), 580–593.

[9] Hugh Christopher Longuet-Higgins (1984) "The Reconstruction of a Scene from Two Projections-Configurations that Defeat the 8-Point Algorithm", In *Proceedings of 1st Conf. Artificial Intelligence Applications*, pp395-397.

[10] Quang-Tuan Luong & Olivier Faugeras, (1993) "Determining the Fundamental Matrix with Planes Instability and New Algorithms", *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, CVPR 93, pp489-494. New York.

[11] Thomas Marill, (1988) "A Counterexample to the Theory That Vision Recovers Three-Dimensional Scenes", *A. I. Working Paper, MIT Artificial Intelligence Laboratory*, 319.

[12] Berthold Horn, (1999) "Projective Geometry Considered Harmful", Copyright © people.csail.mit.edu/bkph/articles/Harmful.pdf, last accessed 2021/6/15.

[13] Tayeb Basta, (2009) "Mathematical flaws in the essential matrix theory", In *Proceedings of WSEAS International Conference*, in *Recent Advances in Computer Engineering*, 9, pp215-220. Budapest, Hungary.

[14] Tayeb Basta, (2010) "An Invalid Derivation of the Fundamental Matrix Based on the Equation of a Point Lying on a Line", Workshop on Frontiers of Computer Vision FCV 2010, pp83 – 88. Hiroshima, Japan.

[15] Tayeb Basta, (2012) "Does the Fundamental Matrix Define a One-to-One Relation between the Corresponding Image Points of a Scene?", *Journal of Image and Graphics*, 1(3), 125-128.

[16] Tayeb Basta, (2014) "Is the Fundamental Matrix Really Independent of the Scene Structure?", *International Journal of Signal Processing, Image Processing and Pattern Recognition*, 7(5), 149-168.

[17] Tayeb Basta, (2017) "The Eight-Point Algorithm is not in Need of Defense", *ARPN Journal of Engineering and Applied Sciences*, 12(3), 753-760.

[18] Tayeb Basta, (2019) "The Controversy Surrounding the Application of Projective Geometry to Stereo Vision", In *Proceedings of the 5th International Conference on Computer and Technology Applications*, ICCTA, pp 15–19. Istanbul.

[19] Tayeb Basta, (2020) "Experimental and Theoretical Scrutiny of the Geometric Derivation of the Fundamental Matrix", In *Proceedings of the 3rd International Conference on Artificial Intelligence and Pattern Recognition*, pp25-29. Huaqiao University, Xiamen, China.

[20] Joaquin Salvi, "Fundamental Matrix Estimation toolbox", http//eia.udg.es/~qsalvi/recerca.html, last accessed 3/3/2016.

[21] Hugh Christopher Longuet-Higgins, (1981) "A computer algorithm for reconstructing a scene from two projections", *Nature*, 293, 133-135.

[22] Quang-Tuan Luong & Olivier Faugeras, (1996) "The Fundamental matrix theory, algorithms, and stability analysis", *International Journal of Computer Vision*, 17(1), 43-76.

[23] Stan Birchfield, (1998) "An introduction to projective geometry (for computer vision)", Stanford university https//www.hhi.fraunhofer.de/fileadmin/Departments/VIT/IMC/Team/schreer/doc_pub/1998_Lecture_ProjGeometryIntro_Birchfield.pdf, last accessed 2021/04/12.

[24] Ron Larson, (2016) *Elementary Linear Algebra*. 8th edn. Cengage Learning.

[25] Jianguo Liu, Xuesong Li, Qin Qin & Hao Zhang, (2016) "Study of occlusion problem in stereo matching", *17th IEEE/ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing* (SNPD), pp195-199. Shanghai, China.

AUTHORS

Tayeb Basta graduated with a B. Eng. in computer science in 1983 from the University of Annaba in Algeria. In 1994, he obtained his PhD in Computer Science from the Victoria University of Manchester in UK. Basta is now an associate professor at Al Ghurair University in Dubai, UAE.

© 2022 By AIRCC Publishing Corporation. This article is published under the Creative Commons Attribution (CC BY) license.

SCALABLE LINK PREDICTION IN TWITTER USING SELF-CONFIGURED FRAMEWORK

Nur Nasuha Daud[1], Siti Hafizah Ab Hamid[1], Chempaka Seri[1], Muntadher Saadoon[1] and Nor Badrul Anuar[2]

[1]Department of Software Engineering, Faculty of Computer Science and Information Technology, University of Malaya, 50603, Kuala Lumpur, Malaysia
[2]Department of Computer System and Technology, Faculty of Computer Science and Information Technology, University of Malaya, 50603, Kuala Lumpur, Malaysia

ABSTRACT

Link prediction analysis becomes vital to acquire a deeper understanding of events underlying social networks interactions and connections especially in current evolving and large-scale social networks. Traditional link prediction approaches underperformed for most large-scale social networks in terms of its scalability and efficiency. Spark is a distributed open-source framework that facilitate scalable link prediction efficiency in large-scale social networks. The framework provides numerous tunable properties for users to manually configure the parameters for the applications. However, manual configurations open to performance issue when the applications start scaling tremendously, which is hard to set up and expose to human errors. This paper introduced a novel Self-Configured Framework (SCF) to provide an autonomous feature in Spark that predicts and sets the best configuration instantly before the application execution using XGBoost classifier. SCF is evaluated on the Twitter social network using three link prediction applications: Graph Clustering (GC), Overlapping Community Detection (OCD), and Redundant Graph Clustering (RGD) to assess the impact of shifting data sizes on different applications in Twitter. The result demonstrates a 40% reduction in prediction time as well as a balanced resource consumption that makes full use of resources, especially for limited number and size of clusters.

KEYWORDS

Self-configured Framework, Link Prediction, Social Network, Large-scale.

1. INTRODUCTION

Link prediction is essential for better understanding how individual interactions and connections evolve in social networking platforms. Statistics highlighted that current number of social network users are increasing linearly every year, where as of April 2020 there are almost 3.9 billion people were active internet users [1]. This evolution inspired many researchers over the past few years to explore new research areas of studies related to link prediction in large-scale social networks. Several efforts are available to address the issues of link prediction in large-scale social networks [2]. Hence, the framework in a distributed computing environment Spark was already successfully adopted in recent link prediction works for good prediction performance of large-scale social networks. Link prediction analysis resolved in less time with Spark by multiple computing resources given the in-memory computation, parallel job processing over master-slave

David C. Wyld et al. (Eds): NATP, ACSTY, CCCIOT, MLSC, ITCSS - 2022
pp. 87-96, 2022. CS & IT - CSCP 2022 DOI: 10.5121/csit.2022.120107

architecture, and its scalability features. Spark also provides numerous properties to configure the computation process such as application properties.

Nonetheless, the difficulty of manually configuring the application properties for execution emerges when the performance of application is degraded and the resource utilization is imbalanced. Further, manual work is extremely difficult and time-consuming for users with less knowledge of how to use the framework. The configuration of properties is critical to the analysis that makes our processing system performs efficiently. Besides the need for efficient link prediction execution, the presence of automatically and correctly configured properties encouraged this research. Using the autonomic computing concept, we proposed a novel self-configured framework based on a trained XGBoost classifier to select the best configuration parameters suited to each submitted application in Spark. The proposed framework attempts to demonstrate performance and efficiency improvements in link prediction analysis in large-scale social networks while consuming resources efficiently and effectively.

The Self-Configured Framework (SCF) is further examined using three link prediction applications which include Graph Clustering (GC), Overlapping Community Detection (OCD), and Redundant Graph Detection (RGD) in Apache Spark. The evaluation results show that in terms of time performance, SCF is able to improve almost 40% of time performance in comparison with default configuration without SCF. Furthermore, we discovered that SCF contributes to balancing the resource utilization presented by the resource utilization rate. Accordingly, the next Section 2 provided an overview of related works in scalable link prediction. Then, in Section 3 we present the proposed framework with its implementation details. Section 4 presented the framework evaluation and highlighted the findings. Finally, Section 5 summarize the important key observations and conclude the proposed work.

2. RELATED WORKS

Link prediction research has evolved in recent years from single computing to distributed computing in an effort to provide scalable link prediction with today's vast data. CBRA [3] invented a novel MapReduce-based method for predicting future links proposed to achieve efficiency in large-scale networks. The parallel CBRA showed great efficiency compared with a traditional single computing link prediction. Although, its computation performance is limited due to the map-reduce procedure involving heavy I/O that consequently affects the prediction performance. Later, PCLP [4] proposed to support parallel link prediction using the Pregel model, a Bulk Synchronous Parallel (BSP) abstraction. BSP supports parallel computing adopted as a major technology for graph analytics at massive scale via Pregel and MapReduce [5]. Link prediction performs better when utilizing Pregel with a big data processing framework like Spark. DTLPLP [6] is the most recent work implementing scalable link prediction on Spark framework conducted using three real-world networks, Enron email, Collaboration Ca-Gr, and Facebook network. Majority of studies in scalable link prediction utilized the Pregel model in Spark framework to handle parallel and iterative processing on large data. The key benefit is that it makes algorithm implementation simple and provides scalability features for enormous datasets. Despite the success demonstrated, the parallel process entails a lot of message generations and transfers, which degrades system performance.

Additionally, there are prevailing works that proposed autonomic computing concept with self-configure feature in big data processing framework. Starfish [7] introduced self-tuning system on Hadoop based on user needs and system workloads for better performance using cost-based modeling and simulation. [8] used cron with python automated script to provide self-configure feature in Hadoop. The cron schedules its jobs automatically and scales the computation based on

the load of the cluster for maximum efficiency of the cluster. Utilizing a cluster reconfiguration algorithm, [9] proved the algorithm is able to dynamically scale according to workload and conserve resource energy in cloud computation up to 54%. InSTechAH scales smart computing tasks on clusters automatically by using a workload prediction algorithm in a KVM-based cloud. There is still a scarcity of studies on how to correctly configure properties of the Spark framework. [10] presented an auto-tuning using a machine learning method, neural network model applied in Spark streaming applications to predict the increase or decrease of cluster configuration. The authors in [11] presented a simulation-driven prediction model for Spark that predicts job performance with high accuracy by anticipating the execution time and memory usage of Spark applications.

3. METHODOLOGY

3.1. Application

Three common link prediction applications are chosen as the benchmark in our experiment, we developed three selected applications based on prevailing works that utilized the commonly used algorithms for scalable link prediction as discussed in Section 2. The three applications are Graph Clustering (GC), Overlapping Community Detection (OCD), and Redundant Graph Detection (RGD) use clustering-based, parallel label propagation-based, and path-based algorithms correspondingly. Therefore, the following is the pseudocode of the applications involved in our framework execution:

```
Algorithm 1: Graph Clustering, GC Application
Input: Edge lists file from HDFS pathFromHDFS, parallelism n, Spark context sc
Output: clusters of predicted link
Begin
val usersRDD, relationshipRDD = sc.textFile(pathFromHDFS)     // RDD creation
val graph = relationshipRDD.outerJoinVertices(usersRDD)     // generate new graph
pageRankGraph = graph.COMPUTE_PAGERANK(sc)     // calculate PageRank value of each vertex
adamicAdarGraph = graph.COMPUTE_ADAMICADAR(sc)     // calculate similarity of each vertex
Repeat until convergence
clusteredGraph = adamicAdarGraph.COMPUTE_POWERITERATIONCLUSTER(k,sc)
clusteredGraph.foreach.print()     // list categorized cluster
```

```
Algorithm 2: Overlapping Community Detection, OCD Application
Input: Edge lists file from HDFS pathFromHDFS, parallelism n, Spark context sc
Output: overlapped communities
Begin
val usersRDD, relationshipRDD = sc.textFile(pathFromHDFS)     // RDD creation
val graph = relationshipRDD.outerJoinVertices(usersRDD)     // generate new graph
MaxIteration = m, Communities = n
graph = graph.AFFILIATE_COMMUNITIES(sc, n)     // community labelling on each vertex
Repeat until m<0
overlappedCommGraph = graph.DETECT_OVERLAPPINGCOMMUNITIES(m)     // detect overlapped comm
overlappedCommGraph.foreach.print()     // list detected communities
```

```
Algorithm 3: Redundant Graph Detection, RGD Application
Input: Edge lists file from HDFS pathFromHDFS, parallelism n, Spark context sc
Output: overlapped communities
Begin
val lines = sc.textFile(pathFromHDFS)     // RDD creation represent graoh records by lines
Edges = lines.flatMapToPair()     // second RDD with JavaPairRDD API to form a connecting link
triads = edges.groupByKey()     // combined both RDD to form triads (graph)
trianglesWithDuplicates = triads.flatMap()
uniqueTriangles = trianglesWithDuplicates.distinct()     // detect duplicated triads
sout(uniqueTriangles)     // list out unique triangles
```

3.2. Self-configured framework

The Self-Configured Framework (SCF) is developed to be a part of the existing open-source Apache Spark framework as illustrated in Figure 1. For each submitted application to the framework, it will undergo an automated configuration beginning with collecting application details and cluster information, decision making for the best configuration, and completing with updating the best identified configuration. SCF used XGBoost classifier in the decision making process where we identify feature sets for XGBoost classifier, which classifies the feature sets into a suitable number of executors per node to be used by the application during execution. Based on the best classification results, the application is ready to be executed automatically through the update module.

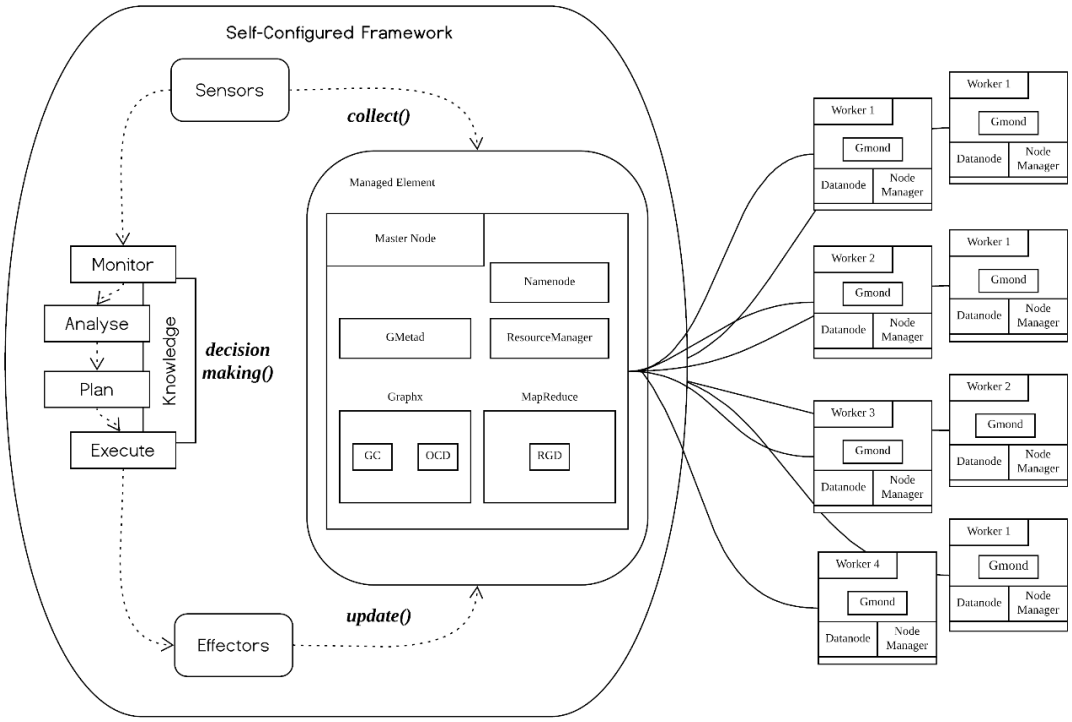

Figure 1: Full architecture of proposed Self-Configured Framework integrated with Apache Spark cluster

SCF consists of three modules and the pseudocode of the implemented modules are as follows:

Algorithm 1: Collect, DecisionMaking, Update modules

Module 1: Collect

Input: P, path of the data file used in the analysis and application
Output: User configurations, application details, and cluster specification

1. From the submitted link prediction application, compute P into size in megabytes, number of lines in the main class, and level of workload
2. Acquire metadata of runtime cluster specification from the registered master server

Module 2: Decision Making

Input: M: application metadata (*mm, mc, wn, wmn, wcn, ds, ac, mec*)
Output: New configuration of *driverCores*, *overheadDriverMemory*, *driverMemory*, *totalInstance*, *overheadMemoryPerExecutor*, *memoryPerExecutors*

1. Assign requested data from collect module to *M*
2. Dispatch *M* into http request API to access xgboost trained model
3. Update predicted value of *EPN* given *M* values in json form

4. if result = 0, make *EPN* = 1 // this is to handle small data cases
5. if result = 1, make *EPN* = 2
6. predefine upper bound values *MOC*, *EM*, *EC*, *ORC*, *ORM*, *PPC*
7. recalculate new configuration for application properties of *dc*, *odm*, *dm*, *ti*, *ompe*, *mpe* given *EPN* without trespassing upper bound

Module 3: Update

Input: New configuration properties

Output: Log updated configuration

1. Check cluster manager
2. If *cm* of submitted application is local, then
3. set smaller spark.driver.core to 2
4. If *cm* is standalone or YARN or Mesos or kubernetes, then
5. Set all recalculated configurations; spark.driver.core, spark.driver.memoryOverhead, spark.driver.memory, spark.executor.instances, spark.executor.memoryOverhead, spark.executor.memory, spark.executor.cores, spark.default.parallelism

---

3.3. XGBoost in Decision Making module

XGBoost is an implementation of gradient boosted decision tree algorithms, a sequential technique designed for speed and performance [12]. In Self-Configured Framework, we implement XGBoost trained model in Python as an API that returns Executor Per Node (EPN) value to DecisionMaking() module as shown in Figure 2. Technically, the input features for our XGBoost model include *masterMemory*, *masterCore*, *workerNode*, *workerMemoryNode*, *workerCoreNode*, *dataSize*, *applicationComplexity*, *memoryCapacity*, and the output or classified feature is predicted value of *executorPerNode*. Tuning parameters that we used are Learning Rate = *Range [0,1]*, Max_depth = *3*, N_estimators = *3*, Objective = *multi:softprob*. The training data used is 70 percent from a total of 16 million configuration records. Once the data set is divided into training and test sets, we build our model with randomly selected data points from the train set. Then we test the model using train set and achieve 100% accuracy. The trained model is then inserted into our storage space to be used by decision making module to request predicted EPN value.

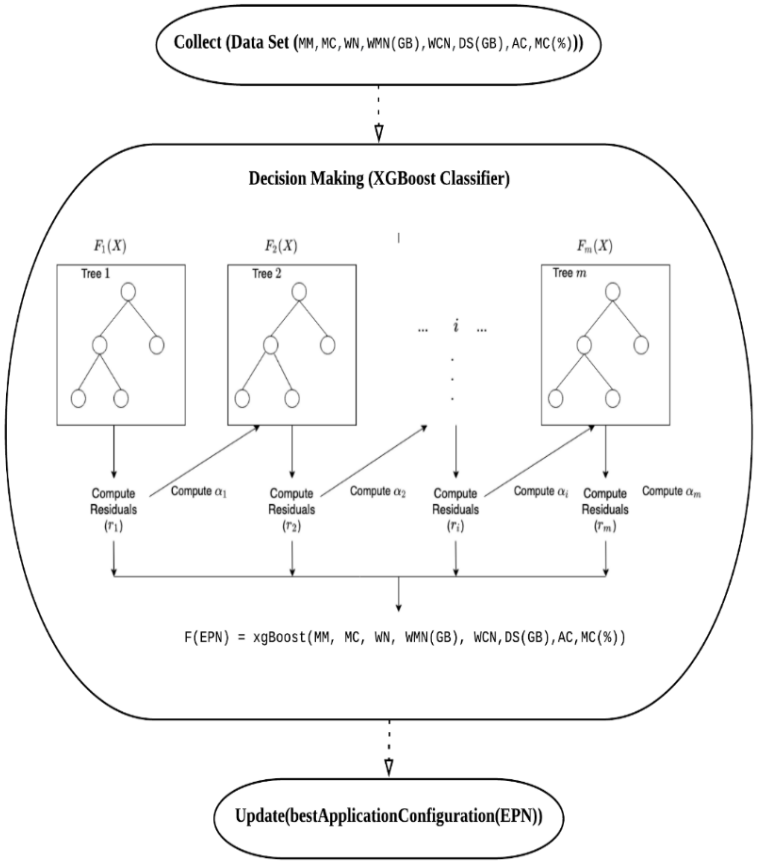

Figure 2: XGBoost model implementation as part of DecisionMaking() module in SCF

3.4. XGBoost vs other classifier

There are several classifications model we used namely; Decision Tree, Logistic Regression, K-Nearest Neighbors. Naïve Bayes, RF, and XGBoost. The best performance, accuracy of training in less time is the XGBoost model. As seen in Table 1, we list the accuracy, precision, recall F1 score, and training time taken for each classifier model. All of the models can achieve the correct classification of the executor per node.

Table 1: Accuracy and Training time comparison of XGBoost with other classifiers.

Classifier	Accuracy	Precision	Recall	F1 Score	Training time(s)
XGBoost	100%	1.0	1.0	1.0	839
DT	100%	1.0	1.0	1.0	133
LR	87%	0.8	0.8	0.79`	8682
KNN	100%	1.0	1.0	1.0	3600
NB	86%	0.81	0.73	0.74	6315
RF	100%	1.0	1.0	1.0	2365

Three classifier models achieved 100% of accuracy, includes Random Forest, Decision Tree, and XGBoost. The reason for 100% accuracy is that the data used has no machine learning problem, we proposed to use a machine learning classifier to enhance the speed of self-configuring properties in SCF in comparison if we only apply a rule-based algorithm. The random forest has the longest training time that is more than half an hour. Decision Tree has the shortest training

time, however, we picked XGBoost as our main classifier because XGBoost is the optimized version of decision tree and claimed to provide advanced learning techniques to yield superior results using fewer computing resources in the shortest amount of time [13]. Although it trained 10 minutes more than decision tree, is the best with multiple decision trees computation that requires more time than a single decision tree classifier. For the SCF dataset, both DT and XGBoost classifiers are reliable and suit tree-based algorithms.

4. SCF EVALUATION

The evaluation was conducted using Apache Spark with two different environments setup Azure Ubuntu 16.04 server for big cluster and Centos 7 ARM x86 server for a smaller cluster. Three types of cluster setup are 2, 4, and 8 node clusters configured in each environment as shown in Figure 3. Memory configurations range from 2GB to 32GB. Data used are 24 million edges and 73 thousand nodes in the form of edge lists. The applications considered in this evaluation are GC, OCD, and RGD written in Scala. The applications are all implementation of link prediction analysis in social networks detailed in Section 3.1.

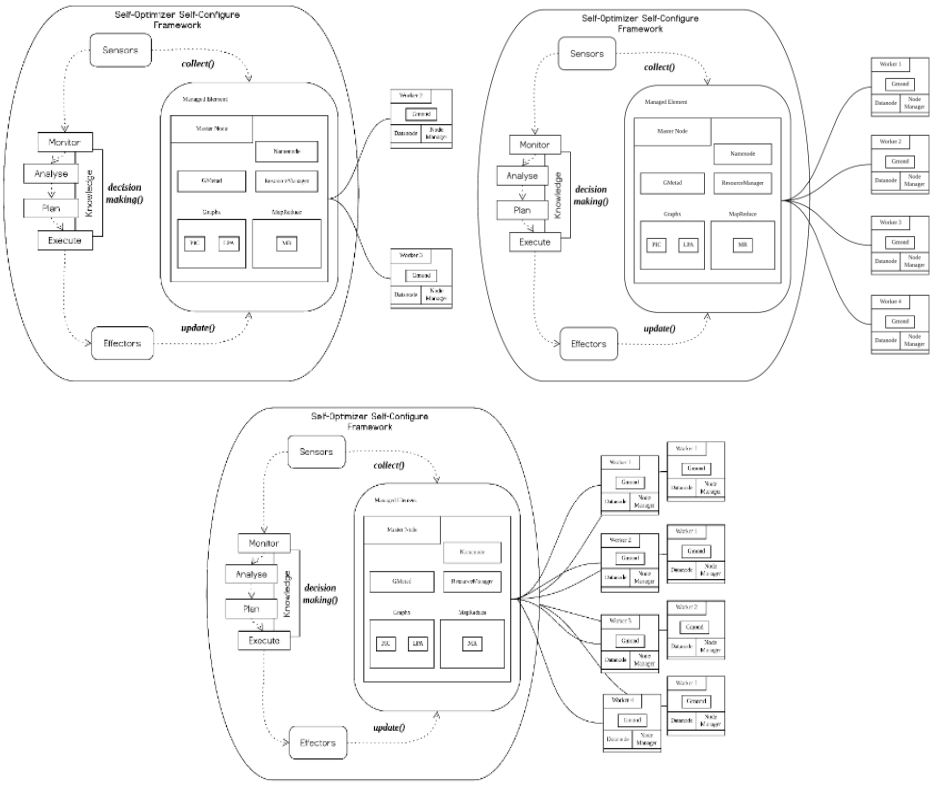

Figure 3: 2, 4, and 8 nodes of spark cluster setup for evaluation

We executed the evaluation with two different methodologies,

- The first evaluation was a comparison of the time performance and efficiency of the three applications in the default configuration and best configuration predicted by SCF.
- The second evaluation conducted a comparison of the resource utilization efficiency of each application.

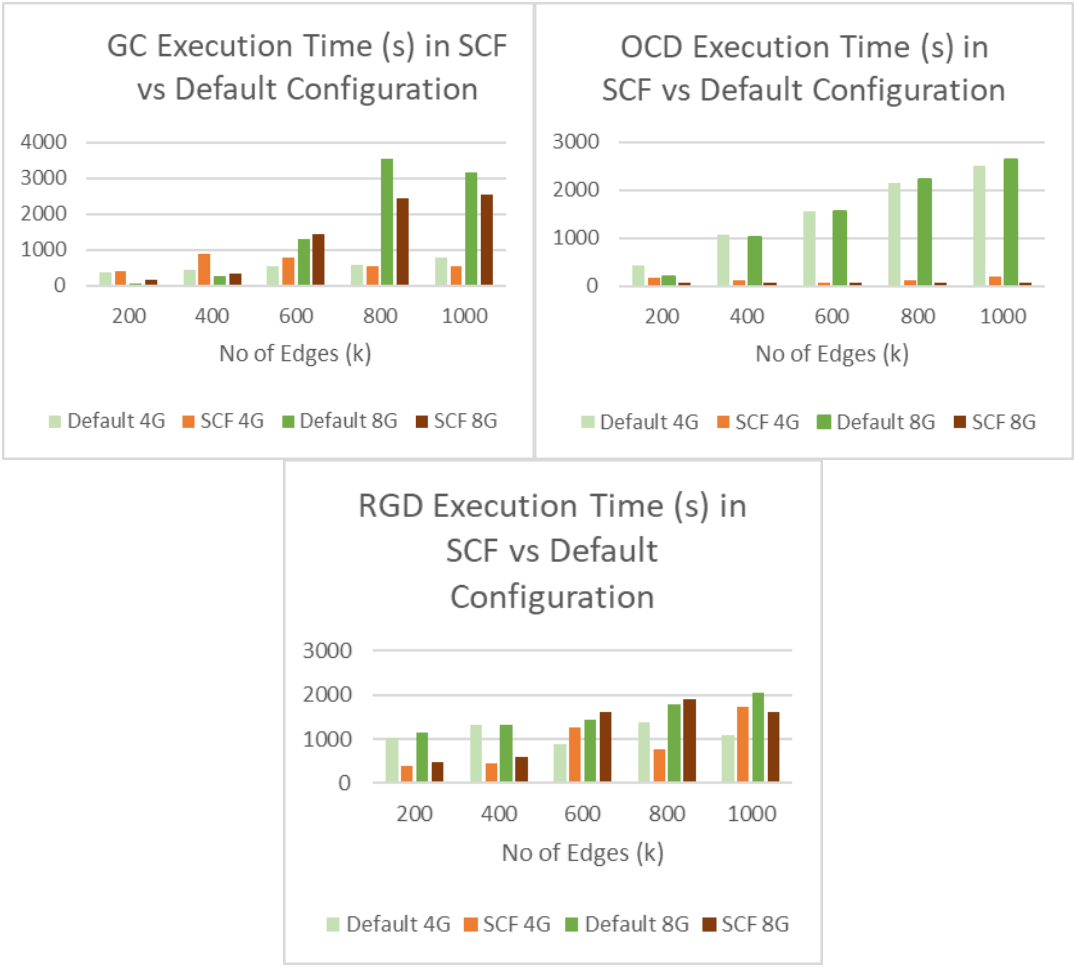

Figure 4: GC, OCD, and RGD execution time for each data used (K) in SCF vs Default Configuration

Figure 4 illustrated the time execution for each application executed in varying no of edges from 200k to 1000k with 4 different case studies which includes: Default configuration of 4GB and 8GB memory, SCF configuration of 4GB and 8GB memory. The plotted time shows that for OCD application, almost 40% lesser time taken to complete the execution when compare with the default configuration 4GB and 8GB memory case studies.

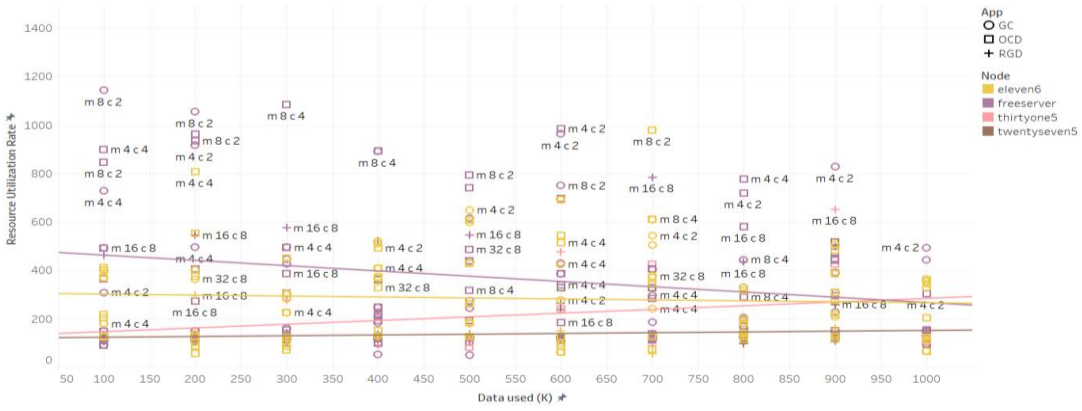

Figure 5: Resource utilization rate vs data used (K) for each application broken down by nodes

We also show how is the rate of all resource utilization in Figure 5. The resource utilization rate is the sum of all resource metrics monitored that are CPU utilization, memory usage in percent, and packet delivery ratio. We observe the resource utilization trend is almost similar for every resource used. This shows that SCF contributed to balancing resource utilization even in varying case studies.

5. CONCLUSIONS

This paper presented a Self-Configured Framework (SCF) integrated with Spark environment to refine the performance and inefficiency of link prediction in large-scale social networks, Twitter. The framework automatically configures the best configuration that suits to particular link prediction application given varying dataset size, workload, and cluster specification in Spark. To provide a general understanding of link prediction, this paper presented state-of-the-art of scalable link prediction in social networks. Further, this research proposed a newly generated dataset for spark clusters in order to invent a new combination set of features for predicting the best configuration of a certain case. Based on the set of features, SCF used XGBoost model to predict a suitable value of executor per node for a submitted link prediction application. The Self-Configured Framework is able to increase the execution time performance at almost 40% and balance the resource utilization when massive application is submitted to the framework for execution. However, SCF is developed as a framework-dependent because as of now it is integrated only with Apache Spark. Finally, this study can provide baseline information on the recent scalable link prediction applications in large-scale social networks the future of SCF would be to integrate with another framework like Hadoop, Storm for streaming analysis.

ACKNOWLEDGEMENTS

We gratefully thank Dr. Zati Hakim Azizul Hasan for providing access to the Robotic Lab and ARM-based server. We are also fortunate to be allowed to conduct the experiment and development in paid servers, Azure Cloud sponsored by Assoc. Prof. Ts. Dr. Nor Badrul Anuar Bin Juma'at. Finally, we also thank Muntadher Saadoon for assistance in accessing and setting up the virtual environment on the ARM server.

REFERENCES

[1] Matt Ahlgren, "40+ Twitter Statistics 2019: Must-Know User Demographics & Facts," 2019. [Online]. Available: https://www.websitehostingrating.com/twitter-statistics/. [Accessed: 10-May-2019].

[2] N. N. Daud, S. H. Ab Hamid, M. Saadoon, F. Sahran, and N. B. Anuar, "Applications of link prediction in social networks: A review," *Journal of Network and Computer Applications*. 2020.

[3] H. Yuan, Y. Ma, F. Zhang, M. Liu, and W. Shen, "A Distributed Link Prediction Algorithm Based on Clustering in Dynamic Social Networks," in *2015 IEEE International Conference on Systems, Man, and Cybernetics*, 2015, pp. 1341–1345.

[4] A. Mohan, R. Venkatesan, and K. V. Pramod, "A scalable method for link prediction in large real world networks," *J. Parallel Distrib. Comput.*, vol. 109, pp. 89–101, Nov. 2017.

[5] "Bulk synchronous parallel - Wikipedia." [Online]. Available: https://en.wikipedia.org/wiki/Bulk_synchronous_parallel. [Accessed: 01-Oct-2021].

[6] X. Xu *et al.*, "Distributed temporal link prediction algorithm based on label propagation," *Futur. Gener. Comput. Syst.*, vol. 93, pp. 627–636, Apr. 2019.

[7] H. Herodotou *et al.*, "Starfish: A Self-tuning System for Big Data Analytics," *undefined*, 2011.

[8] S. N. S. Toomas Roomer, "Autoscaling Hadoop Clusters Supervisor," no. may, 2010.

[9] N. Maheshwari, R. Nanduri, and V. Varma, "Dynamic energy efficient data placement and cluster reconfiguration algorithm for MapReduce framework," *Futur. Gener. Comput. Syst.*, vol. 28, no. 1, pp. 119–127, 2012.

[10] J. Gu, Y. Li, H. Tang, and Z. Wu, "Auto-Tuning Spark Configurations Based on Neural Network," in *IEEE International Conference on Communications*, 2018.
[11] K. Wang and M. M. H. Khan, "Performance prediction for apache spark platform," *Proc. - 2015 IEEE 17th Int. Conf. High Perform. Comput. Commun. 2015 IEEE 7th Int. Symp. Cybersp. Saf. Secur. 2015 IEEE 12th Int. Conf. Embed. Softw. Syst. H*, pp. 166–173, 2015.
[12] T. Chen and T. He, "xgboost: eXtreme Gradient Boosting," 2020.
[13] "Use case on the Buzzinsocialmedia_Twitter dataset | MLJAR," 2018. [Online]. Available: https://mljar.com/machine-learning/use-case/buzzinsocialmedia-twitter/. [Accessed: 14-Sep-2021].

AUTHORS

Nur Nasuha Daud received Bachelor degree in Computer Science and Information Technology (Software Engineering) from University of Malaya, Malaysia. Nasuha is currently pursuing a PhD program from the same university in the Department of Software Engineering, Faculty of Computer Science and Information Technology. Her Ph.D research is in Social network analysis with specific focus on Link Prediction.

Siti Hafizah Ab Hamid received BS (Hons) in Computer Science from University of Technology, Malaysia, MS in Computer System Design from Manchester University, UK., and the PhD in Computer Science from University of Malaya, Malaysia. She is currently an Associate Professor with the Department of Software Engineering, Faculty of Computer Science & Information Technology, and University of Malaya, Malaysia. She has authored over 80 research articles in different fields, including mobile cloud computing, big data, software testing, software engineering, machine learning and IoT.

Nor Badrul Anuar received his Master of Computer Science from University of Malaya in 2003 and a PhD at the Centre for Information Security & Network Research, University of Plymouth, UK in 2012. He is currently an Associate Professor with the Faculty of Computer Science and Information Technology, University of Malaya. He has authored over 128 research articles and a number of conference papers locally and internationally.

© 2022 By AIRCC Publishing Corporation. This article is published under the Creative Commons Attribution (CC BY) license.

A SURVEY OF CLOUD SERVICE EVENTS AND THEIR CONNECTIONS

Hangping Hu, Zhen Zhang, Weijian Qin, Yuan Wang and Xiaojian Li

School of Computer Science & Engineering Guangxi
Normal University, Guilin, China

ABSTRACT

Any unexpected service interruption or failure may cause customer dissatisfaction or economic losses. To distinguish the rights and interests or security disputes between cloud service providers and customers, explore the essence and rules of cloud service events and their various connections, such as: Normal contact of service scheduling, normal contact of service dependence, abnormal contact of resource competition, abnormal contact of service delay, abnormal contact of service dependence, etc., as well as their rules in time, resources, scheduling and other aspects, and the form of the rules; The purpose is to provide the above abnormal connections, as well as the rule and presentation form in terms of time, resources and load, for the study of violation determination and failure tracing in the cloud service accountability mechanism.

KEYWORDS

Cloud service, Event connections, Correlation, Label adaptation.

1. INTRODUCTION

After receiving user service requests, cloud service providers implement their requirements into one or more cloud service jobs, which may have sequence and dependency relationships. Each cloud service job consists of one or more tasks, and there are dependencies among the tasks, which are represented by the DAG (Directed Acyclic Graph) of the tasks. After the task is put into operation, one or more instances will be generated. Since there is a dependency relationship between tasks, instances of two tasks with a dependency relationship must have a dependency connection. It would be helpful to trace the source more accurately if we could point out which instances the connection is caused by, when it ends, and the type and strength of the connection.

This paper summarizes the research status of cloud service events and their connections, mainly including: firstly, this paper reviews the research status of cloud service events and their relationship. Secondly, summarizing and giving a classification of cloud service events and their connections from the current state of affairs. Thirdly, it points out that the common problem at present is interpretable label adaptive labeling. Finally, the research approaches of cloud service events and their correlation are suggested.

The following four sections are as follows: Section 2 summarizes the current situation of cloud service events and their connection; Section 3 gives the classification of current cloud service events and their connection; Section 4 summarizes the remaining problems in the current research and gives suggestions on research approaches; Section 5 is the summary of the full text.

David C. Wyld et al. (Eds): NATP, ACSTY, CCCIOT, MLSC, ITCSS - 2022
pp. 97-104, 2022. CS & IT - CSCP 2022 DOI: 10.5121/csit.2022.120108

2. RESEARCH STATUS

The relevant research status is summarized in two sections: Section 2.1 discusses cloud service events and Section 2.2 discusses the connection between cloud service events.

2.1. Cloud service events

For the lack of labeled data in abnormal detection, paper [2] presents an unsupervised abnormal detection method, which can identify three types of anomalies: Service timeout, network delay and data loss. This method ignores the dependency between events, and independently examines the abnormal conditions of events, and ignores the abnormal conditions such as the delay time of bad services and the dissatisfaction of resources.

Paper [3] makes an empirical analysis on abnormal events collected by 18 online service systems of Microsoft, identify accidental abnormal events, such as wrong procedures submitted by customers, so as to effectively deal with abnormal events, but this method relies on engineers to manually mark accidental abnormal events, which is costly.

Papers [4,6] only focus on the abnormal resource consumption event of the job. Paper[5] only focus on job scheduling failure events, papers [7-10] only focus on the abnormal consumption of container resources, and paper [11] only focus on instance authorization failure and Instantiate abnormal event. They all assume that each event is independent of other events, pay attention to the abnormal of events in terms of resources and scheduling, but ignore the abnormal caused by the dependency between events.

Papers [27-31] analyze the failure characteristics of nodes, but they separate the connection between nodes and independently investigate the failure characteristics of nodes, ignoring the abnormal connection between nodes.

It can be seen that there are still abnormal in the research of cloud service events, which ignore the delay time of bad services and the dependency between events, and rely on manual labeling, which is inefficient.

2.2. Cloud Service Event Connections

Papers [12-16] rely on manual annotation, Give the two types of connections between services "with-without", so as to trace the source of service interruption.

Paper [18] only focus on resource competition between jobs, papers [19-21] only focus on dependencies between containers. They all focus on the connection between events from a single aspect, without considering multiple abnormal connections between events from various aspects such as time, resources and scheduling.

Papers [22-24] can analyze whether there is a connection between events, and give the general characteristics of the connection, but they fail to distinguish the types of connections.

Papers [33, 34] based on subspace method, and papers [35-37] based on feature selection methods, extract or construct a new low-dimensional feature space in the high-dimensional space to detect the abnormal connection, but due to the unknown and heterogeneity of the abnormal connection, these methods are difficult to ensure that the newly constructed feature space already

contains the semantics required for the Multi-Classification and Multi-Label abnormal connections discovery [32].

It can be seen that the research on event connection still relies on manual annotation, which is inefficient, and does not integrate time, resources, scheduling, dependency and other aspects at the same time, and the problem of unclear understanding of multi-Classification connection characteristics.

3. CLOUD SERVICE EVENTS AND THEIR CONNECTIONS CLASSIFICATION

This section summarizes and classifies cloud service events and their connections from the current situation.

3.1. Cloud service events classification

According to the papers [2-11, 27-31], it is summarized that there are normal events, unknown events and abnormal events in cloud service events, this article focuses on abnormal events. The cloud service events classification diagram is shown in Figure 1.

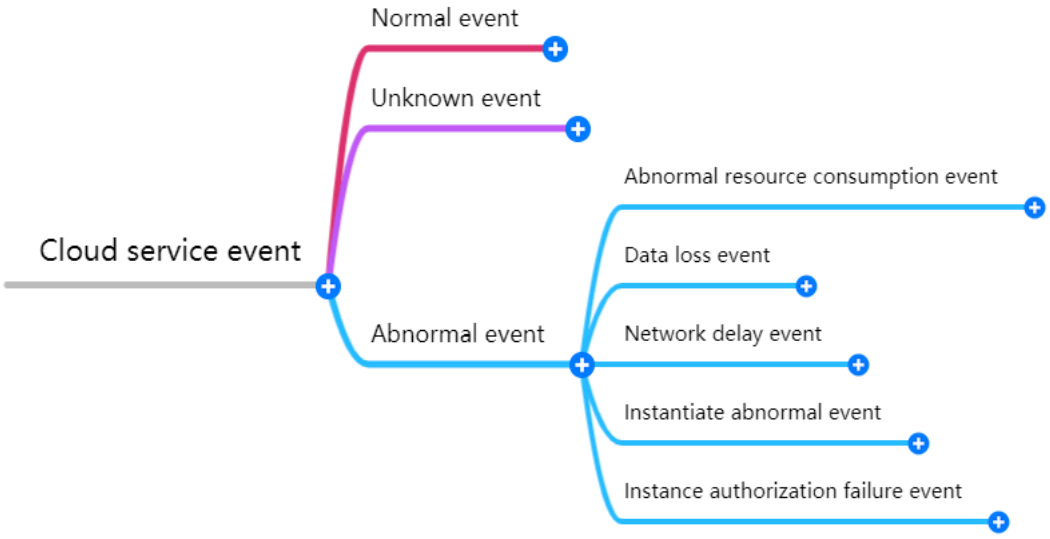

Figure 1. Cloud service events classification

We classify cloud service events into normal events, unknown events, and abnormal events based on their completion. Abnormal events include abnormal resource consumption events, data loss events, network delay events, instantiation abnormal events, and instance authorization failure events.

3.2. Cloud service event connections classification

According to the papers [12-16, 18-24, 32-37], it can be concluded that cloud service event connections includes normal connection, unknown connection and abnormal connection. This article focuses on abnormal connection. The cloud service event connections classification diagram is shown in Figure 2.

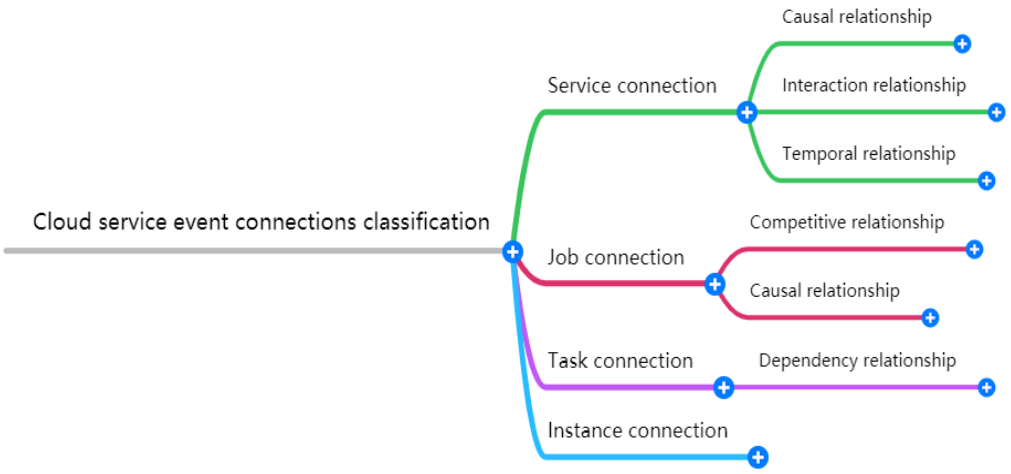

Figure 2. Cloud service event connections classification

According to the granularity of cloud service events, we classify event connections into service connection, job connection, task connection and instance connection. Among them, service connection includes causal relationship, interaction relationship and temporal relationship among services. Job connection includes competitive relationship and causal relationship between jobs. Task connection includes dependency relationship between tasks.

4. REMAINING PROBLEMS AND SUGGESTIONS FOR RESEARCH APPROACHES

4.1. Remaining problems

Problem 1: Adaptive labeling problems of interpretable labels.
Problem 2: Difficulties in understanding the nature of multi-classification connections in cloud service events.
Problem 3: The method challenge of multi-classification connection discovery of cloud service events.

4.2. Research approaches and suggestions

Based on the above urgent problems, the prospective system shown in Figure 3 is proposed.

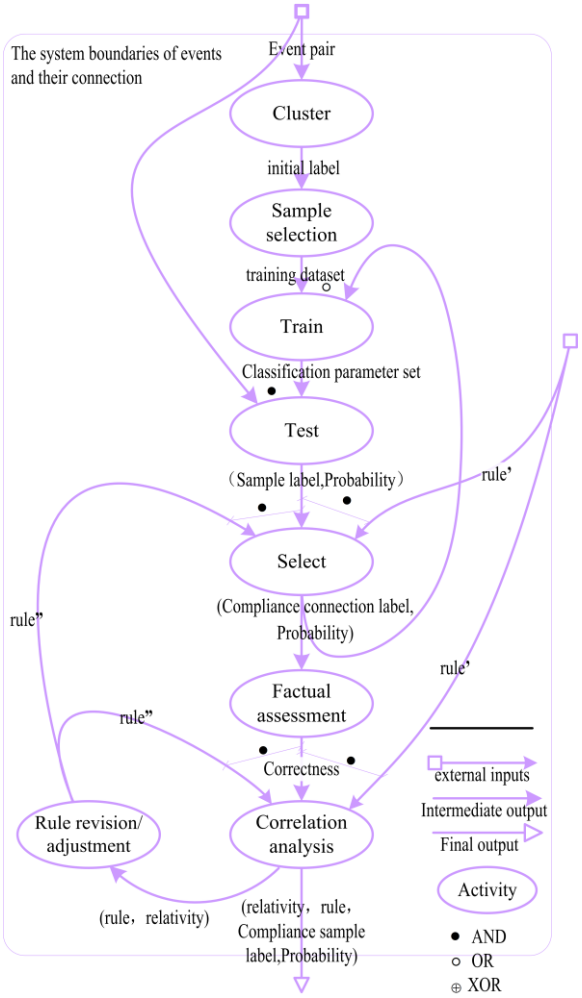

Figure 3. Prospective system

The ellipse in Figure 3 represents the activity, and the directed solid line represents the entity. Directed lines starting from hollow rectangles represent external inputs, and directed lines with solid arrows at the end represent intermediate outputs, and directed lines with hollow arrows at the end represent final outputs. The analysis of cloud service failure factors includes 8 activities: cluster, sample selection, train, test, select, factual assessment, correlation analysis and rule revision/adjustment.

According to the prospective system shown in Figure 3, we give the research scheme. Firstly, explore the adaptive annotation method of event connection labels, such as the initial label of the connection between events is given by clustering. Secondly, analyze the initial label meaning of the connections between events, and get the types of connections and some characteristics of each connection. Thirdly, part of the characteristics of all kinds of connections are regarded as semi-supervised classification rules, such as "do not accept the connection labels that start only after the subsequent events have ended", "do not accept the normal association labels that abnormal end of the subsequent events and abnormal end of the previous events", etc. Then, through semi-supervised classification, classify the types of connections between events.

Fourthly, according to the connections of events, get more characteristics and laws of the connections of various events, as well as the presentation forms of laws, and construct the event network. Finally, evaluate the above process, and give the evaluation results.

5. CONCLUSIONS

This article reviews the current research status of cloud service events and their connections, and points out the following three common problems:

1）Deep learning relies on labels, moreover, event connections label of cloud service is people's qualitative prior knowledge of event connections, it is usually manually annotation in advance and then classified, but manual annotation is labor intensive and inefficient.

2）The explanatory nature of the label depends on the recognition of multi-classification connection characteristics, however, the current cloud service event connection has not considered time, resources, scheduling, and dependencies at the same time, and the problem of unclear understanding of the nature of multi-classification of connection;

3）Machine adjustment labels depends on the understanding of the laws of multi-classification connection, however, the current research is not clear about the characteristics of multi-classification connections, and part of it contains the semantic problems required for multi-classification connections discovery.

For the above problems, we give the prospective system of cloud service events and their connections, which basically meet the requirements of cloud service events and their connections.

ACKNOWLEDGEMENTS

We would like to thank the National Natural Science Foundation of China (61862008, U1636208) for its support.

REFERENCES

[1] Raju B K, Geethakumari G. Event correlation in cloud: a forensic perspective[J]. Computing, 2016, 98(11): 1203-1224.
[2] Nedelkoski S, Cardoso J, Kao O. Anomaly detection and classification using distributed tracing and deep learning[C]//2019 19th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGRID). IEEE, 2019: 241-250.
[3] Chen J, Zhang S, He X, et al. How Incidental are the Incidents? Characterizing and Prioritizing Incidents for Large-Scale Online Service Systems[C]//2020 35th IEEE/ACM International Conference on Automated Software Engineering (ASE). IEEE, 2020: 373-384.
[4] Lu C, Chen W, Ye K, et al. Understanding the Workload Characteristics in Alibaba: A View from Directed Acyclic Graph Analysis[C]//2020 International Conference on High Performance Big Data and Intelligent Systems (HPBD&IS). IEEE, 2020: 1-8.
[5] Jassas M S, Mahmoud Q H. Failure characterization and prediction of scheduling jobs in google cluster traces[C]//2019 IEEE 10th GCC Conference & Exhibition (GCC). IEEE, 2019: 1-7.
[6] Chen X, Lu C D, Pattabiraman K. Failure analysis of jobs in compute clouds: A google cluster case study[C]//2014 IEEE 25th International Symposium on Software Reliability Engineering. IEEE, 2014: 167-177.
[7] Ren R, Li J, Wang L, et al. Anomaly Analysis and Diagnosis for Co-located Datacenter Workloads in the Alibaba Cluster[C]//International Symposium on Benchmarking, Measuring and Optimization. Springer, Cham, 2019: 278-291.
[8] Scheinert D, Acker A. Telesto: A graph neural network model for anomaly classification in cloud services[C]//International Conference on Service-Oriented Computing. Springer, Cham, 2020: 214-227.

[9] Gulenko A, Schmidt F, Acker A, et al. Detecting anomalous behavior of black-box services modeled with distance-based online clustering[C]//2018 IEEE 11th International Conference on Cloud Computing (CLOUD). IEEE, 2018: 912-915.
[10] Acker A, Schmidt F, Gulenko A, et al. Online Density Grid Pattern Analysis to Classify Anomalies in Cloud and NFV Systems[C]//2018 IEEE International Conference on Cloud Computing Technology and Science (CloudCom). IEEE, 2018: 290-295.
[11] Vedurumudi P V V, Morusupalli P. System and method of providing post error analysis for instances of applications in cloud service environments on a per user basis: U.S. Patent 10,379,934[P]. 2019-8-13.
[12] Wang Y, Li G, Wang Z, et al. Fast Outage Analysis of Large-scale Production Clouds with Service Correlation Mining[C]//2021 IEEE/ACM 43rd International Conference on Software Engineering (ICSE). IEEE, 2021: 885-896.
[13] Meng Y, Zhang S, Sun Y, et al. Localizing failure root causes in a microservice through causality inference[C]//2020 IEEE/ACM 28th International Symposium on Quality of Service (IWQoS). IEEE, 2020: 1-10.
[14] Jia T, Chen P, Yang L, et al. An approach for anomaly diagnosis based on hybrid graph model with logs for distributed services[C]//2017 IEEE International Conference on Web Services (ICWS). IEEE, 2017: 25-32.
[15] Nedelkoski S, Cardoso J, Kao O. Anomaly detection from system tracing data using multimodal deep learning[C]//2019 IEEE 12th International Conference on Cloud Computing (CLOUD). IEEE, 2019: 179-186.
[16] Chen Y, Yang X, Lin Q, et al. Outage prediction and diagnosis for cloud service systems[C]//The World Wide Web Conference. 2019: 2659-2665.
[17] Reguieg H. Using mapreduce to scale event correlation discovery for process mining [D]. Université Blaise Pascal-Clermont-Ferrand II, 2014.
[18] Rosa A, Chen L Y, Binder W. Understanding unsuccessful executions in big-data systems[C]//2015 15th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing. IEEE, 2015: 741-744.
[19] Weng J, Wang J H, Yang J, et al. Root cause analysis of anomalies of multitier services in public clouds [J]. IEEE/ACM Transactions on Networking, 2018, 26(4): 1646-1659.
[20] Raju B K, Geethakumari G. Event correlation in cloud: a forensic perspective [J]. Computing, 2016, 98(11): 1203-1224.
[21] Scheinert D, Acker A, Thamsen L, et al. Learning Dependencies in Distributed Cloud Applications to Identify and Localize Anomalies[J]. arXiv preprint arXiv:2103.05245, 2021.
[22] Khan S, Parkinson S. event log analysis: an association rule mining and automated planning approach[J]. Expert Systems with Applications, 2018, 113: 116-127.
[23] Fedorchenko A, Kotenko I, El Baz D. Correlation of security events based on the analysis of structures of event types[C]//2017 9th IEEE international conference on intelligent data acquisition and advanced computing systems: technology and applications (IDAACS). IEEE, 2017, 1: 270-276.
[24] Li G, De Carvalho R M, van der Aalst W M P. Configurable event correlation for process discovery from object-centric event data[C]//2018 IEEE International Conference on Web Services (ICWS). IEEE, 2018: 203-210.
[25] Jiang C, Han G, Lin J, et al. Characteristics of co-allocated online services and batch jobs in internet data centers: a case study from Alibaba cloud[J]. IEEE Access, 2019, 7: 22495-22508.
[26] Nguyen V, Dang T. CloudTraceViz: A Visualization Tool for Tracing Dynamic Usage of Cloud Computing Resources[C].2019 IEEE/ACM Industry/University Joint International Workshop on Data-center Automation, Analytics, and Control (DAAC). IEEE, 2019: 1-6.
[27] Bhattacharyya A, Singh H, Jandaghi S A J, et al. Online characterization of buggy applications running on the cloud[C]. 2016 12th International Conference on Network and Service Management (CNSM). IEEE, 2016: 282-286.
[28] Jassas M, Mahmoud Q H. Failure analysis and characterization of scheduling jobs in google cluster trace[C].IECON 2018-44th Annual Conference of the IEEE Industrial Electronics Society. IEEE, 2018: 3102-3107.
[29] Rosa A, Chen L Y, Binder W. Catching failures of failures at big-data clusters: A two-level neural network approach[C]. 2015 IEEE 23rd International Symposium on Quality of Service (IWQoS). IEEE, 2015: 231-236.

[30] El-Sayed N, Zhu H, Schroeder B. Learning from failure across multiple clusters: A trace-driven approach to understanding, predicting, and mitigating job terminations[C].2017 IEEE 37th International Conference on Distributed Computing Systems (ICDCS). IEEE, 2017: 1333-1344.
[31] Tang H, Li Y, Jia T, et al. Analysis of Frequently Failing Tasks and Rescheduling Strategy in the Cloud System[J]. International Journal of Distributed Systems and Technologies (IJDST), 2018, 9(1): 16-38.
[32] Pang G, Shen C, Cao L, et al. Deep learning for anomaly detection: a review [J]. ACM Computing Surveys (CSUR), 2021, 54(2): 1-38.
[33] Keller F, Muller E, Bohm K. HiCS: High contrast subspaces for density-based outlier ranking[C]//2012 IEEE 28th international conference on data engineering. IEEE, 2012: 1037-1048.
[34] Lazarevic A, Kumar V. Feature bagging for outlier detection[C]//Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining. 2005: 157-166.
[35] Azmandian F, Yilmazer A, Dy J G, et al. GPU-accelerated feature selection for outlier detection using the local kernel density ratio[C]//2012 IEEE 12th International Conference on Data Mining. IEEE, 2012: 51-60.
[36] Pang G, Cao L, Chen L, et al. Sparse modeling-based sequential ensemble learning for effective outlier detection in high-dimensional numeric data[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2018, 32(1).
[37] Pang G, Cao L, Chen L, et al. Learning Homophily Couplings from Non-IID Data for Joint Feature Selection and Noise-Resilient Outlier Detection[C]//IJCAI. 2017: 2585-2591.

© 2022 By AIRCC Publishing Corporation. This article is published under the Creative Commons Attribution (CC BY) license.

AN INTELLIGENT SYSTEM TO AUTOMATE THE INQUERY IN LOGISTICS INDUSTRY USING AI AND MACHINE LEARNING

Leo Liao[1] and Ang Li[2]

[1]Crean Lutheran High School, 12500 Sand Canyon Avenue,
Irvine, CA 92618, USA
[2]California State University, Long Beach, USA

ABSTRACT

Operator and sales employees in the logistics industry often have to submit the same inquiry repetitively to different vendors and opt in for the quotation that will generate the greatest profit for the company [4]. This process can be very laborious and tedious. Meanwhile, for smaller companies that do not have a well-constructed database for quotation information, monitoring employee's work is simply difficult to achieve [5]. To increase the efficiency of sales' workflow in this particular industry, this application devises a platform that automates the inquiry process, analyzes quotations from different vendors, retrieves the most profitable one, and documents all inquiries an employee has committed [6].
The results, after a series of intensive testing, prove to be promising and satisfying. The machine learning model can successfully fetch the most cost-effective price after analyzing a list of emails containing common languages used in the industry. All histories of an employee's inquiry can be correctly displayed on any front-end device. Overall, the obstacle presented above is largely solved.

KEYWORDS

Automation, Quotation, Analysis.

1. INTRODUCTION

The logistics industry is an industry that provides transportation services from one geological point to another [7]. Logistics companies manage and process service requests from upstream parties (usually importing and exporting corporations) [8]. When clients seek these companies for a service, companies return a quotation. If clients accept the quotation they receive, logistics companies then process these requests by arranging specific transportation services, communicating with customs, and finally carrying the goods to the destination. Although this general workflow is divided into different tasks and is much standardized, specific responsibility assigned to individual workers in the working process sometimes is not structured and generalized [9]. It means that there is still a potential way to optimize some steps in the workflow. More specifically about the problem itself, when sales and operators render a quotation to clients, they have to contact different trucking companies to request a quote and service for clients. Traditionally, sales and operators have to email different parties with the same information several times and receive the quotation they have obtained. By comparing various costs and previous impressions of these companies, logistics industry employees make an independent decision about which trucking companies to cooperate with. This working experience is, first of

David C. Wyld et al. (Eds): NATP, ACSTY, CCCIOT, MLSC, ITCSS - 2022
pp. 105-113, 2022. CS & IT - CSCP 2022 DOI: 10.5121/csit.2022.120109

all, not productive and controllable for a company. Second of all, it is not pleasing for all employees due to its repetitive and redundant nature.

One solution is then proposed to address this awkward situation for both logistics industry companies and employees. Our platform develops an interface and server to allow users to input information only for one time and submit a request individually to different trucking companies. By implementing some pricing algorithms, the server will retrieve the most profitable one and suggest that to the user from the email response sent back. In one word, one key contribution this platform makes is the reduction of time-consuming working experience and increase in productivity of the entire company [10].

Employees in this particular workflow still primarily rely on phone and emails. They will make a call or send an email to different trucking companies to obtain the information about the quotations. Based on their working experience and consultations, employees will make an individual decision on which trucking companies to choose. Although emails reduce some redundancies of the workflow, inputting the same information several times is not convenient for all involved parties, especially for employees. Moreover, there is a lack of standardization of which trucking cooperators. Employees make decisions that might not be the most secure or profitable ones for a company. In the same logic, although sometimes experience can pave the way for the most fit option, newer, inexperienced employees choosing a profitable and stable cooperator is also hard to accomplish. Often a company faces a deficit on a single service due to an too expensive cooperating trucking service that leads to higher costs but brings a low income at the end. They also have to accept the risk when choosing a low-cost cooperator because cargos can get lost due to an unqualified service. Additionally, obtaining the costs through email can be redundant to perceive because employees have to process these prices by adding them together and then comparing. Different trucking companies' quotations require employees to do different calculations everytime, which is extremely inefficient and repetitive. Finally, it is difficult for companies to record the history of quotations because all information is processed by hand and email for which no one is accounting. Due to the lack of recording, unqualified trucking companies cannot be detected and filtered out for future cooperation, leaving another risk of a deficit. In one word, conventional communication methods are not modern and advanced enough to carry out current workflow and make the most profit for logistic companies [11].

Our method is to simplify the information inputting process to a one-time work. The platform allows employees to input the information about the inquiry only once. After clicking the submit button placed at the bottom of the screen, all emails will be distributed to each destination. Compared with traditional methods, employees do not have to compose an email with each service provider, waiting for them to reply. Another is automation of calculation in emails. Conventionally, employees have to do a lot of calculations by adding other surcharges indicated by trucking companies to obtain a final price. On this platform, the server implements algorithms to analyze each email and extract all prices from the email to calculate a final cost. This new method saves employees time to complete other important tasks. In addition to automatic calculation, this platform also helps compare each price and keep them in record. Employees do not have to then rely on paper and pencil to record. Lastly, employees are free from the burden of the management of all inquiries as the platform will save and display their work immediately after they submit their requests. Companies also enjoy this benefit as well because prior to the invention of this platform, they did not have the opportunities to check individual employee's working progress [12]. Constructing a management structure will help businesses to monitor each inquiry and its status of completion.

Proving the outcome from the platform is relatively simple. We employ a number of previous emails, feeding it to the model and checking if it will compute the desired calculation and return

results. Using approximately fifty emails sent back to the company, the machine model can detect the price ($+number) in each clause and extract it from the text sample. Although some little errors remain, most of the algorithm is successful and effective. The second stage of proof is to input information by oneself to stimulate a real application of the platform. Throwing in what employees usually compose in an email from the front-end, the retrieving result is satisfying.

As a comparison, employees manually send individual emails to vendors to request a quotation. Email sent back will be artificially analyzed to extract a price. Although it is mostly successful, the speed for which this information is processed is very slow and redundant. It usually takes ten minutes or more for an employee to complete a task that can be handled in a second by the platform.

The rate of success for each trail, by average, is greater than 70% with multiple distractions adding in, such as zip code and phone numbers. All experiments and evaluations, we intend to stimulate the real working situations, since this is the primary goal of the platform. We even modify harder situations to accommodate the possibility that any emails are not formatted ideally or content contains errors in syntax. Nevertheless, the success rate is acceptable and keeps increasing as more updates are made on the algorithm and model.

The rest of the paper is organized as follows: Section 2 gives the details on the challenges that we met during the experiment and designing the sample; Section 3 focuses on the details of our solutions corresponding to the challenges that we mentioned in Section 2; Section 4 presents the relevant details about the experiment we did, following by presenting the related work in Section 5. Finally, Section 6 gives the conclusion remarks, as well as pointing out the future work of this project.

2. CHALLENGES

In order to build the tracking system, a few challenges have been identified as follows.

2.1. Regular expression implemented to extract prices

One challenge is the regular expression implemented to extract prices. Since each email contains surcharges, models need to identify those surcharges and add them together. For the first models, the algorithms do not identify that extra information, but only retrieve the first price it detects. For example, the email will contain a main price for a particular shipment of $40.00. However, it will also list other surcharges for any additional need such as fuel and other service fees. The first model only extracts the first price it sees, abandoning other costs. For revision, I add components from regular expressions so that whenever the model sees a dollar sign in the email, it knows the next string is a cost that needs to be picked up to store. This time the issue is fixed. All numbers with a dollar sign will be grabbed and accumulated to calculate a final price. Even with some kinds of constraints (for any price must start with a dollar sign as a signal), this method is mostly successful because it models how humans perceive and comprehend an email.

2.2. Classifying emails to different users

Another challenge is classifying emails to different users. Since there will be multiple users to send emails to different destinations, the data structure of the system must be efficient enough to store and locate emails. Even though the system does not report an error when emails are not directed to a specific destination, it is mostly due to the fact there is only one tester on the platform. In real situations in which numerous emails will be processed and results distributed to

each end user, managing these emails is extremely crucial as it is individual work that needs to be monitored. To this end, we develop the idea of “quote-id”, “minor-id”, and “user-id”. Minor-id keeps track of different vendors. Quote-id records each quote submission in general. User-id records from whom the quote is submitted. Once these labels are created, emails can be managed more structurally for business and intensive use.

2.3. Price comparison

The last challenge is price comparison. The comparison's goal is to extract, among multiple costs, the median one. The difficulty of this problem is not as high as others. However, many scenarios need to be considered because only one implementation can handle this task. For example, many vendors will return the same price. And there are chances that even numbers of quotations are returned, making it impossible for the model to choose a price to display. To solve this issue, algorithms will display all prices with minor-id’s and recommend the relatively higher price to the users. At this time, human participation is needed to make a decision for the machine model, opting into the best choice based on their experience. This is the only process in which human power is needed. But it will not decrease the efficiency overall because it can connect to later workflow, which requires human participation, nonetheless.

3. SOLUTION

The entire platform is built in two parts: user interface and server. In the user interface, employees input information from an inquiry [13]. After filling out all information needed, it will be sent to the server that is going to compile it into a professional email and send it individually to vendors. These vendors will be labeled with an id, respectively, and stored in an email-list. Before sending each email, the server will assign it with a unique id (quote-id + minor-id + user-id) so that responses given back can be distributed to each end user correctly. To continue, once an email is back from a trucking company, it will be fed into the model. There it will be decomposed and analyzed in smaller components. After several emails are back, the comparison model will be prompted to extract the most favorable cost, which will then be listed as the recommended one in the query results. To ensure stability and security of each process in the workflow, other prices will also be listed for employees to confirm before making a decision. The same idea applies to other quotes. After submitting multiple inquiries and receiving some quotations, that information will be organized under the “My” tab. Each inquiry is labeled as the time it is submitted to the server so that employees have an easier time to identify how many and when they submit an inquiry [14]. For a company, managers can have an intuitive idea of the performance of individual employees on this particular workflow.

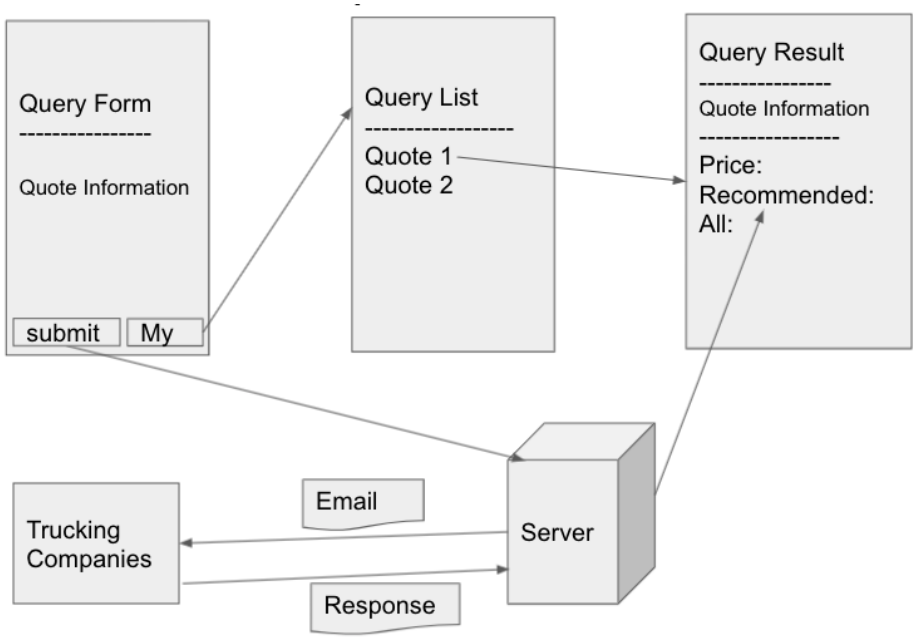

Figure 1. Overview of the project

For the user interface, I employ the flutter SDK to implement a number of widgets, including text fields, buttons, list-view, and alert-logs. Text field will collect information about user input. Buttons will be prompted to send requests to the server. List-view will be used to display quote names. Alert-logs will be used to display specific information about a quotation. After a user logs in to the system, information about an inquiry will be provided. When clicking the confirm button, all information above will be sent to a specific function implemented in the server. In that function, information will be stored in a json dictionary. Then an email is prompted in that function to send the inquiry to different email addresses listed in the email list. After that, the server will scan the inbox every few minutes, checking if any of the sent emails are responded to. If there are, another server function will fetch this email and input it to a function that will parse the price from the email. This process is the most important part of the platform. I use regular expressions and the NLTK package to accomplish this task. Examining over fifty email samples, we develop a pattern of expressions in the emails. Since all samples have common characteristics like prices and other professional words, regular expressions can accommodate those and pick up the price after a particular signal string. Even if sometimes the target is lost because of over-complicated clauses and exceptional expressions with numbers, the success rate for the processing is satisfying overall. Most regular quotations can be withdrawn correctly. After a price is retrieved, information about the inquiry will be accessed and sent back to the user interface. There, users can check the status of each inquiry they make, including the information they have input previously, all prices with vendor number, and a recommended cost derived from the algorithm. The user will make an independent choice from that point, pertaining to which party the company shall opt in with.

Above is the workflow of this platform, the data structure is presented below. Since it is used by employees, I set individual users at the top of the structure. From there, each user will develop independent quotes. Each quote contains its respective information. Until now, this structure is perceived as the most efficient and easy to manage one. Since we do not have to delete or transplant information in this particular dataset, this structure is the most secure and stable one to maintain and sustain. Future modification is definitely needed; however, only addition will apply, which is the most safest measure to take. Other data structures might also apply in this project,

but it will not be as efficient and stable as the current one because data has to be transferred over to another point and deleted in its current place, which makes it not ideal as there are dangers of losing data.

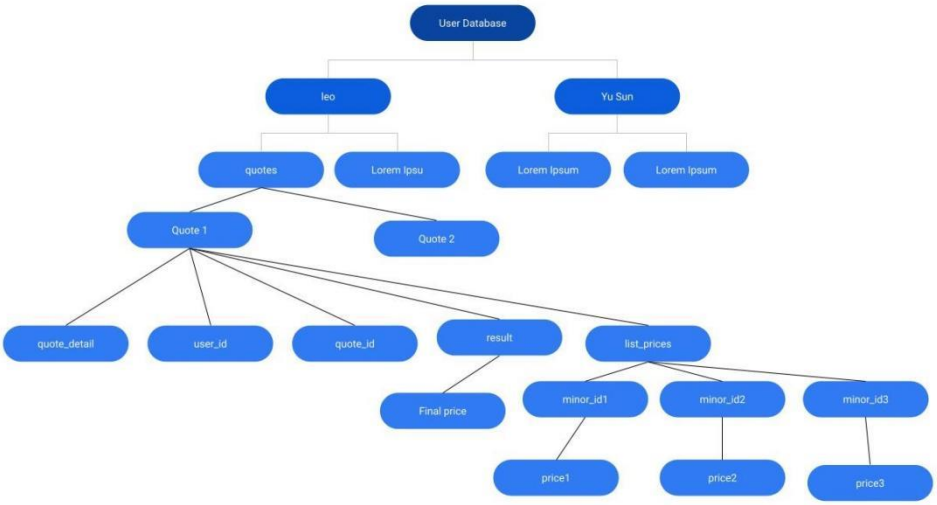

Figure 2. Data structure

4. EXPERIMENT

4.1. Experiment 1

We developed a platform to receive quotes from different logistics industry employees, extract all quote information and automatically send customized emails to the different trucking companies. Our solution reduces misleading communication between trucking companies and logistics industry employees since there are mediators between them. Also, it reduces the amount of time to send the information to trucking vendors since emails are sent instantaneously after logistics industry employees send the quote requests.

For our experiment, we utilized fifty quote samples that were sent to a group of three people. Also, these fifty samples were sent to our platform. For some of the samples, we sent them every minute while others were sent 15 - 20 minutes between them. For the trucking vendors, we used 5 email accounts to simulate trucking vendors.

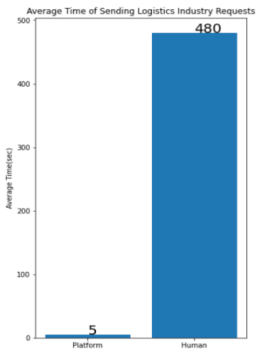

Figure 3. Average time of sending logistics industry requests

The above chart bar shows that the average time of the platform is less in comparison to the human method. Our platform method takes about 5 seconds to send the emails to the simulated trucking vendor emails while the human method takes on average 480 seconds to send the emails to the vendors.

4.2. Experiment 2

We use Natural Language Processing and Python regular expressions to extract the price from the email. Natural Language Processing is a powerful tool that can extract the parts of speech of a sentence. Thus, we use the NLTK package to create a grammar that contains a regular expression of part of speech, so we can extract the correct part of the tree that contains the description and price of the service. Then, we use Python regular expressions to select the desired price of the service. By doing this method, we were able to analyze and extract information from the email.

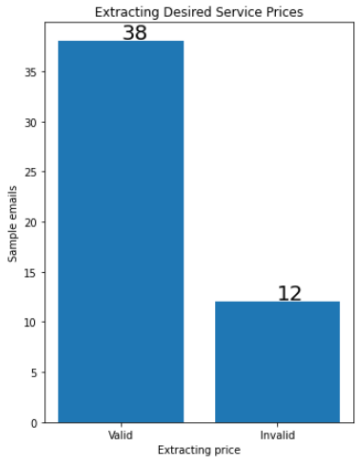

Figure 4. Extracting desired service prices

In our experiment, we use fifty sample emails to extract the prices and fetch the desired service. These email samples were obtained from real responses from the truck companies which contain diverse patterns. These patterns were used to verify if NLTK grammar was correctly designed, so description and the price of truck services were extracted from the emails. After processing and extracting the info from the fifty email samples, we obtained that our system could analyze and extract 38 of the email samples correctly which corresponds to 76% of the samples while 12 samples of the emails were returned incorrectly.

In experiment 1, as we expected, our platform works more efficiently than the human method since the time of extracting and sending the emails to trucking vendors takes about 5 seconds on average while the human method takes about 480 seconds. Also, our platform reduces the amount of mistakes that can occur by sending the quotes to the trucking vendors since there are no intermediates between the trucking vendors and logistics industry employees.

Natural Language Processing is a useful tool to extract sentences that contain part of speech patterns. In experiment 2, after we used regular expressions and NLTK packages, we observed that the success rate of our system was 76%. Even though the description and prices were extracted incorrectly from some of the sample emails since they did not follow the part of speech patterns of our grammar or the system selected the incorrect service price, we could state that the success rate of our system is acceptable.

5. RELATED WORK

Freightos is also an online platform that automates the requests of inquiries [1]. In their work, they compare multiple aspects between the manual process and digital process of logistic workflow. Their production (application), compared to our application, exhibits a more thorough automation for the entire workflow. While it is more narrowly applicable, our application targets a specific client group, employees. This aspect of our work is both a strength and major difference compared to Freightos. As a full-scale automated platform, Freightos sometimes cannot obtain a proper cost for customers, unlike our platform that employs some degree of human resources to better assist customers.

Freightquotes achieves the same function as our application does [2]. One main difference, however, is the source of input to the platform. Freightquotes' target client is customers who will submit their quotes directly to obtain a price from trucking companies. Our application, instead, prompts employees to initiate the submission of prices to trucking companies and then send back costs to customers. Even though it seems to be an unnecessary step, this algorithm can reduce error in the system that may cause more laborious service processes.

In Tung Nguyen's commentary research, one of the paper's points coincides with our future plan for the application [3]. In 3.1.1, it is said that forecasting demand in the industry is "a crucial part of business operation in every sector… to come up with the most reliable data of the upcoming period" (Nguyen 9). I intend to add another algorithm to our platform to analyze the quality and quantity of service provided by each vendor and selectively predict and choose those that are cooperative with us to collaborate for future work. One major difference between our work is that this paper is mostly theorized and researched-based while our project is in a real application scenario.

6. CONCLUSIONS

Overall, to reduce human resources needed for a logistic business when requesting repetitive inquiries to different trucking vendors, I design an automated platform on which employees can enter information to obtain a list of quotes from truckers [15]. After that information is entered, an email will be generated and compiled to send to a number of vendors for reply. For the platform will scan the inbox every once in a while, once replies are back, algorithms such as regular expression and natural language learning model will analyze the content of that email to extract a final price, taking extra fees into the account and calculations. After experimenting with a number of test cases, the system and algorithm proved to be effective for most professional business email. With a few editions on the implementation of regular expression, specifically, the program achieves greater accuracy and efficiency. Since the number of emails and users that will be served on the platform, each user is assigned an individual id. Each piece of email will be labeled with a major quote id and two minor id, indicating from which the email is sent and to which it is sent. In these ways, a well-constructed and mature data structure can be utilized for business purposes. Through experimentation, this data structure has also proved to be effective so far for current and future use in the near future. In conclusion, the current platform is mature and ready to put into real applications. It will alleviate employee's working pressure by reducing repetitions and unnecessary efforts.

Even though its platform is considered mature and functional, it has a few limitations that will hinder it from full scale automation in the future. Since the transportation system in logistics is divided into several regions across the continent, each region has several vendors that carry the service. In what way this system will manage inquiries to be sent to a proper set of vendors that

are responsible for the region is one potential threat to current practicability. If that limitation is overcome, the platform is fundamentally practical across the globe and can better serve every employee in every logistic business.

The limitation above can be solved by setting up a database that stores all trucking companies. In this database, vendors are further divided by states in which they can provide service. Each time the user calls for a service, the platform will send email to the vendors that are responsible for the state. In this way, the platform is perfected.

REFERENCES

[1] Riedl, Jens, et al. "The Digital Imperative In Freight Forwarding." BCG, 2018, https://image-src.bcg.com/Images/BCG-The-Digital-Imperative-in-Freight-Forwarding-Nov-2018_tcm108-207934.pdf.

[2] Le Blanc, Louis, and Laura Valentine. "FREIGHTQUOTE.COM: Value-Added Intermediation in Transportation." FREIGHTQUOTE.COM: VALUE-ADDED INTERMEDIATION IN TRANSPORTATION , 2010, https://ctrf.ca/wp-content/uploads/2014/07/3LeBlancValentineFreightQuote.pdf.

[3] Nguyen, Tung. "Enhancing Logistics and Warehouse Management For a Startup Company: Challenges and Opportunities." Theseus, 1 Jan. 1970, https://www.theseus.fi/handle/10024/172082.

[4] Gohberg, Israel, and Seymour Goldberg. Basic operator theory. Birkhäuser, 2013..

[5] Davidson, Donald. "Quotation." Theory and decision 11.1 (1979): 27.

[6] Tharp, Roland G., and Ronald Gallimore. "Inquiry process in program development." Journal of Community Psychology 10.2 (1982): 103-118.

[7] Murphy, Paul R., and A. Michael Knemeyer. "Contemporary logistics." (2018).

[8] Lambert, Douglas M., and James R. Stock. Strategic logistics management. Vol. 3. Homewood, IL: Irwin, 1993.

[9] Van Der Aalst, Wil, Kees Max Van Hee, and Kees van Hee. Workflow management: models, methods, and systems. MIT press, 2004.

[10] Plag, Ingo. "Productivity." The handbook of English linguistics (2020): 483-499.

[11] Moore, Richard. "Imitation and conventional communication." Biology & Philosophy 28.3 (2013): 481-500.

[12] Parker, Geoffrey, and Marshall W. Van Alstyne. "Platform strategy." (2014).

[13] Stalnaker, Robert. "Inquiry." (1984).

[14] Zeisel, John. "Inquiry by design." Environment/behavior/neuroscience in architecture, interiors, landscape, and planning (2006).

[15] Kleinbaum, David G., et al. Logistic regression. New York: Springer-Verlag, 2002.

© 2022 By AIRCC Publishing Corporation. This article is published under the Creative Commons Attribution (CC BY) license.

PERVASIVE SYSTEMS DEVELOPMENT: A STEPWISE RULE-CENTRIC RIGOROUS SERVICE-ORIENTED ARCHITECTURAL APPROACH

Nasreddine Aoumeur[1] and Kamel barkaoui[2]

[1]Department of Computer Science, University of Leicester, LE1, 7RH, UK
[2]SYS:Equipe Systèmes Sûrs, Cedric/CNAM, France

ABSTRACT

To stay competitive in today's high market volatility and globalization, cross-organizational business information systems and processes are deemed to be knowledge-intensive (e.g. rule-centric), highly adaptive and context-aware, that is explicitly responding to their surrounding environment, user's preferences and sensing devices. Towards achieving these objectives in developing such applications, we put forwards in this paper a stepwise service-oriented approach that exhibits an explicit separation of concerns, that is, we first conceptualize the mandatory functionalities and then separately and explicitly consider the added-values of contextual concerns, which we then integrate at both the fine-grained activity-level and the coarse-grained process-level to reflect their intuitive business semantics. Secondly, the proposed approach is based on business rule-centric architectural techniques, with emphasis on Event-Conditions-Actions (ECA)-driven transient tailored and adaptive architectural connectors. As third benefit, for formal underpinnings towards rapid-prototyping and validation, we semantically interpret the approach into rewriting logic and its true-concurrent and reflective operational semantics governed by the intrinsic practical Maude language.

KEYWORDS

Context-awareness, ECA-Driven Rules, Architectural Connectors, Service-orientation, Adaptability, Maude Validation.

1. INTRODUCTION

Boosted by technological advances in networking, context-sensing and computation and pressed by stiff and global competitiveness, most of organizations are opportunistically joining their know-how into dynamic giant cross-organizational alliances. Striking features of any of such alliances, include: (1) *process-centricity*, that is, they exhibit very complex business processes with composing activities; (2) *high-agility*, where involved business processes and their activities / tasks are often governed by adaptive and evolving business rules [6,9,10,11,16]; (3) *context-dependency*, with user-preferences, adopted sensing devices and surrounding environment conditions as driving forces [4, 5, 28]; (4) *strong-dependability*, where the fatality of mal-functioning and failures may be economical und humanistic disastrous. Precise conceptualizations and formal techniques are thus highly required before investing any final deployment.

David C. Wyld et al. (Eds): NATP, ACSTY, CCCIOT, MLSC, ITCSS - 2022
pp. 115-132, 2022. CS & IT - CSCP 2022 DOI: 10.5121/csit.2022.120110

Towards reliably developing such complex, agile and context-dependent business applications, we are putting forwards in this paper a stepwise rule-based model-driven and context-aware architectural approach with the following software-engineering milestones:

[**Fine-grain Separation of concerns**]: We argue that for taming the complexity and agility of such business applications, involved concerns such as functionalities, context-awareness, security and quality need to be explicitly and separately handled at a first stage, then integrated at the architectural-level to reflect the reality of the application. Moreover, due to the intractable complexity of any business process, we suggest a fine-grained activity-centric handling of such concerns as first-class modelling entities.

[**Rule-centric architectural handling**]: Towards intuitively coping with agility in such volatile business applications, we are capitalizing on the ubiquity of business rules in such applications. Indeed, business rules reflect evolving policies and laws for doing / collaborating business [6, 9, 10, 16]. Furthermore, to enhance agility and bridge the gap to service-orientation, we shift any intuitive business rule towards transient architectural ECA-driven connectors [13]. More specifically, we propose ECA-generic patterns for both functionality- and context-awareness.

[**Rule-centric concurrent formalization**]: Towards soundly interpreting and validating this conceptualization, we further enforce to stay compliant with this rule-centricity. For that we propose Meseguer's rewriting (rule-based) logic [12] and its enabler efficient Maude language [3, 8].

[**Rule-centric service-oriented deployment**]: To preserve all strengths of the "business-foundation" phases, we propose rule-centric web-services for a compliant deployment [22]. However, to enhance readability and simplicity and keep with the space limitations this phase will not be further discussed. Detail about this phase will be addressed in the extended journal version.

The remaining sections are as follows. In the next section we summarize recent related work referring to any of the three milestones of our approach and their interleaving namely: Context-awareness, adaptability and its rule-centricity. The third section illustrates the working architecture of the proposed approach. In the fourth section, through a simplified banking process, we demonstrate how functionalities are captured at the architectural level using a tailored ECA-driven composition. In the fifth section, we present how context-aware knowledge is to be modelled at the architectural, with the introduction of the so-called context-intensive ECA-driven architectural connectors. In the sixth section, to reflect the intended intuitive business semantics of each activity, we present how both concerns require to be brought together around their activities within the concerned business process. We then address the formalization of the approach using rewriting logic and Maude language; Nevertheless, with the aim to boost the smooth readability of the paper we just skip the phase of the translation of ECA-driven connectors to the service-oriented RuleML [22]. We finally wrap up this paper with concluding remarks.

2. RELATED WORK: CONTEXT-AWARENESS, ADAPTABILITY AND RULE-CENTRICITY

As we pointed out in the introduction the innovative stepwise approach we are putting forwards for developing adaptive and context-aware business information systems bring together in a harmonious and architecturally-based manner, the following software-engineering ingredients: (1) Context-awareness; (2) Adaptability; (3) Separation in a rule-based way between context- and

functional concerns at the fine-grained activity-level; (4) Adoption of Service-orientation where the ECA-rules are captured as connectors between different service interfaces (5) Formalisation through the rule-based rewriting logic for rapid-prototyping and validation.

Saying that in the following we restrict ourselves to any related work that integrates at-least two of three of these software-ingredients; otherwise, the related work will transcend by far the space limitation of the paper.

The closest approach to our that integrates at-least the notion of rules and context-awareness in coping with dynamic and adaptive business processes in forwarded in [26]; furthermore, the paper addresses most of the related work in this respect and therefore we invite interested reader to go through these related works. In some detail, the authors propose to distinguish between internal and external contexts, both managed using what they refer to context-engine-, resources- and rule-managers. What seems to be close to our approach, is that the transition from one business activity to another is dynamically governed by the current context of the previous activity and the associated rules. Not mentioned in their paper compared to our is clearly the service-orientation and the formal-underpinnings.

Another recent interesting approach with some similarity to ours appeared in [29] and based on three-level architectural solution to develop context-aware application. The three-level are the *perception layer*, which covers in some sense the technical and technological-layer for sensing and interpreting the context; to mention here that in our approach this layer is assumed given at-priori as the technology is far advanced in this respect (see [14,16,27, 28]) for more detail concerning this technical side). The second level named *interference layer* and concerns the contextual rules, where the author suggests directly adopting the RuleML [22] XML-based low representation and not a business-level ECA-driven rules like our approach. Furthermore, besides the rules, the author proposes an inference-engine and context-brother to compute on these rules. the third-level is named the *application layer*; here the so-called application manager decides depending on the available context whether to run the suitable application. Finally, is to mention that a prototype called KoDA has been developed as a proof-of-concept for that approach.

Other approaches focussing mostly on the context-awareness and its modelling and implementation could be found among others found in [26, 27]. Concepts such as RFID, Ambient systems, Sensors and their classification are among others widely explained context-related ingredients there. Furthermore, the application of the context-awareness is experienced within different areas with the very interesting and critical health systems.

Finally, it is important to point out that our early ideas around separating location- from functionality-concerns, as preliminary work towards this currently more disciplined and stepwise approach for pervasive systems development, have been forwarded in [1]. That is, the present paper extends by far these ideas on several perspectives. Firstly, we are leveraging location-awareness with tailored context-awareness primitives to cope with pervasive systems. Second, we have been putting these first ideas into a throughout stepwise and service-oriented development approach. Third, as further contribution of this paper is the formalisation and validation of the approach using the rule-based rewriting logic and its intrinsic true-concurrent Maude Language. Last but not least, we are aiming to efficiently implement the approach using RuleML [10] and putting the approach into the context of development in particular of business information systems and processes at the fine-grained adaptability activity-level as will be detailed in the remaining sections.

3. MULTI-CONCERN AGILE SERVICE-ORIENTED BPS: APPROACH MILESTONES

As depicted in the Figure-1 below, the working general architecture of the approach we are pushing forwards can be highlighted as follows. Above all, we assume as given initial informal requirements such as: business goals and objectives, intentional business rules where specifically context-aware ones have been extracted from the environment sensing devices, actuators and so-on, informal business processes and their composing activities. Given such initial requirements, the gradual development of reliable context-aware service-oriented agile business applications, encompasses the following phases:

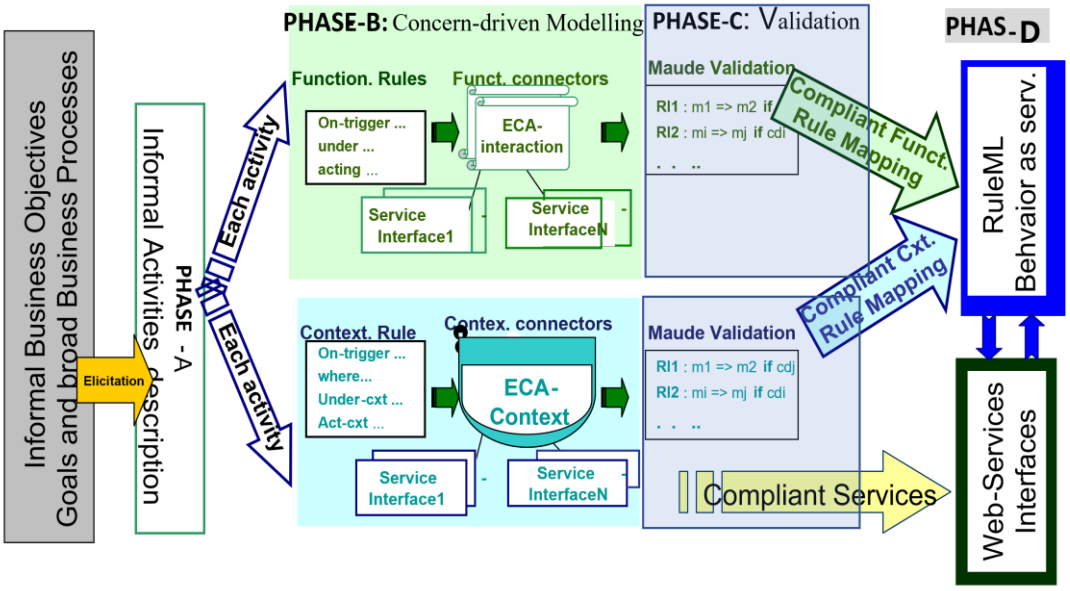

Figure 1. The forwarded Stepwise Architectural Context-aware Approach Milestones

[**Phase-A: Informal look at activities**]: The purpose of this phase is to re-visit the activities that may participate in any business process. For instance, we propose to list and describe them informally as preparation of the "business-conceptual" phase. In particular, we assume at this phase that context-aware information and knowledge have been already sensed at interpreted [27, 28]

[**Phase-B-Separated Rule-centric Modelling of Functionality and Context concerns**]: We consider this as *the most decisive phase* and as a distinguished capability of our approach with respect to the state-of-art. As depicted in Figure-1, this phase is progressive and involves two successive steps:

1. [*ECA-driven behaviour for any activity*]: That is, for each concern (e.g. functionalities, context-awareness), we separately describe the corresponding ECA-driven rules governing any activity. For the context-aware concerns, we propose a set of simple yet tailored business-level primitives to facilitate an intuitive description.
2. [*Architectural conceptualization using ECA-driven connectors*]: Towards deriving a disciplined and agile conceptualization while closing the gap to service-orientation, we propose to shift such informal ECA-driven descriptions towards architectural concepts.

We propose tailored ECA-driven architectural connectors, with roles playing service interfaces and their behavioural glues reflecting the composition logic [1].

[**Phase-C-Validation of Functionality and Context concerns using Rewriting Logic**]: For the formal validation, rapid-prototyping and verification, we propose yet another rule-based logic that completely fits within the proposed ECA-driven rule-based modelling phase. That is, this step within this "business-foundation"-level concerns the formal underpinnings of each concern at the activity-level using the true-concurrent rewriting logic-based semantics [12] supported by its Maude governing language [3,8].

[**Phase-D: RuleML-centric Service-oriented Deployment**]: At this ultimate phase, we propose to deploy the already certified and reliable service-driven application using Web-Services technology [21,25]. To stay compliant with the rule-centricity of the approach and thus preserve all its benefits, we take benefits from rule-based XML languages, specifically reactive RuleML [22], we leverage to service-orientation.

In the following sections, we will detail these phases, by considering a simplified case-study dealing E-banking. More precisely, as illustrated in Figure-2, we consider the following service-oriented business process, where after being identified, a customer can perform any banking action such a withdrawal, acquiring-loan, and so on.

- Customer identification and authentication.
- Customer performing a withdrawal (or deposits, loans, mortgages).

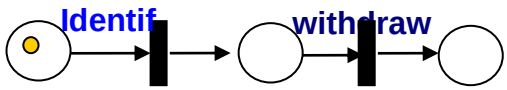

Figure 2. A Simple Petri-Net like Busines Process Model for basic banking Operations

We should already emphasize that most of existing approaches, as detailed in section 2, do not delve into the inside-behaviour of such activities composing a business process. Indeed, even service-oriented proposals, are restricted to only the modelling of message exchanges (e.g. send, receive and invoke) to perform an activity. In terms of Petri nets for instance as shown above, the activities are mostly considered as *black-boxes*, which opens a wide room for the programmers to implement them in ad-hoc and rigid if not in incorrect manner. The main goal of our approach consists thus in bringing to the conceptual level and precisely at the fine-grained activity-level at first stance as much *multi-dimensional knowledge* as possible in a manner that promotes adaptability, composability and dependability.

4. COMPOSITE SERVICE FUNCTIONALITIES AS RULE-CENTRIC INTERACTIONS

As we already point out, to capture the mandatory functionalities of any business activity, in each business process, we first propose to reformulate any governing intentional business rules into operational ECA-centric ones. Moreover, we endeavour describing such ECA-driven business rules at the interaction level, that is, we enforce ourselves to find out which business entities, modelled later as web services requiring from them appropriate interfaces, are involved in the associated ECA-driven rule. Then, the triggering events, the constraints to observe and the actions to perform are to be specified.

Afterwards, we propose to smoothly shift this interaction-driven informal business rules governing any activity into more disciplined architectural interconnections. For the architectural connector behaviour, which should reflect the ECA-driven rule, we propose a tailored ECA-driven generic pattern composed of the tailored primitives as to be described below.

4.1. Functionalities ECA-based Rules Illustrated and Intuitively Clarified

Let us straightaway consider the withdrawal action, indeed the identification-activity presents no functionality at-all as thus completely context-aware one as we will see it later, as a business activity in our banking business process. We propose to externalize at the interaction-level the rule governing the functioning of the withdrawal (i.e. in its simplicity (balance > amount-to-withdraw). Instead of speaking about the "withdrawal method", we are thus speaking about an agreement between the customer and (one of) his/her account(s) while banking. As direct benefit, we can now have different agreements depending on the profile of the customer (e.g. silver, golden) and its account (e.g. running, saving, asset). Moreover, we can address the policies of defining and adapting such agreements on-the-fly, and thereby increasing the competitiveness of any associated financial institution. Last but not least, the notion of triggering event (i.e. the customer wants to perform a withdrawal) is inherently to be understood now as an explicit "invitation" for the account to enter into composition with that customer.

Coincidently, these are the main features in the essence service paradigm SOC [19,25]. Firstly, SOC aims at dynamically composing of different partners (service interfaces) to achieve added-values, impossible to achieve by single partners. Second, SOC is based on service invocation using subscription and notification and dynamic binding.

As shown in Figure-3, for any withdrawal agreement, we require the following information from the two partners: From the customer, we require the triggering event and the fact that (s)he is owning the account; From the account, we require the balance and the debit operation, restricted to just the decreasing of the balance (i.e. NO internal conditions at all). Important requirement for the intended composition logic is the activity, as a composite service itself. We thus describe any activity-behaviour at-first level, based on the ECA-driven interaction puts in place to ensure the underlying business goal.

Example 1 (The ECA-based Functional rule for the Standard Withdrawal): As depicted below, the standard withdraw consists simply in externalizing the usual condition from the account component to the interaction level. The rule says that: *On the occurrence of a withdrawal event (subscription) from the customer, the targeted account balance should be greater than the requested withdrawal amount and in that case a debit message is (asynchronously) sent to that account to debit that account.*

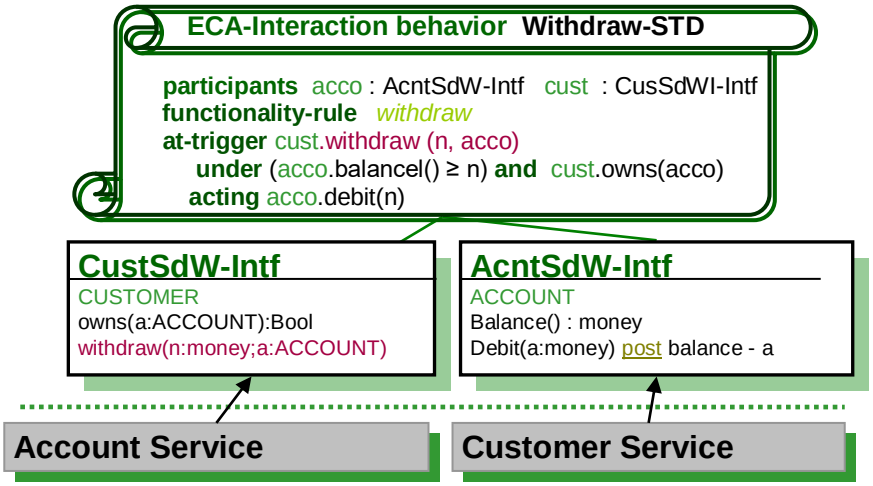

Figure 3. The Standard withdrawal ECA-rule as an Architectural Contract

Example 2 (The ECA-based Functional rule for the VIP Withdrawal): The second possible withdrawal agreement, as illustrated in Figure-4, consists in endowing "privileged" customers with a credit so that they can withdraw below their account balances. The interaction ECA-based rule as an architectural connector takes the following form, where all what changes in respect to the standard case is the condition that becomes more flexible.

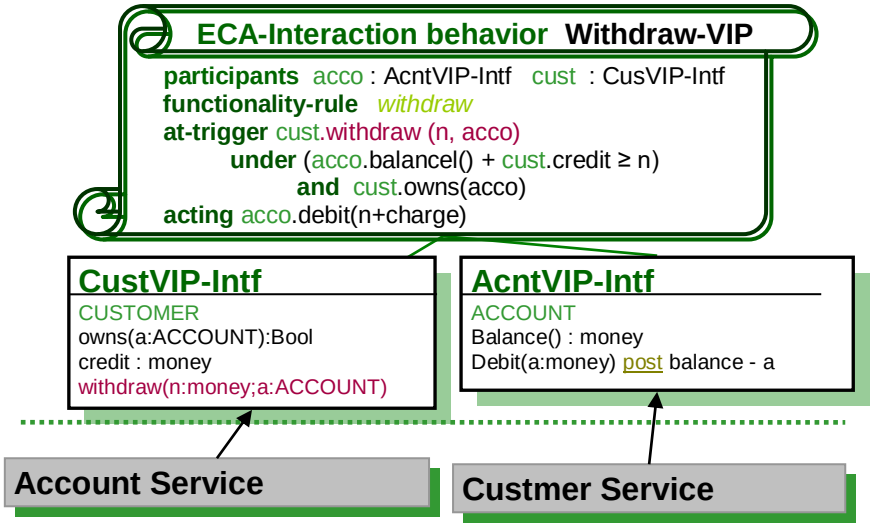

Figure 4. The VIP withdrawal ECA-rule as an Architectural Contract

Notice that a different partner is now required to play the role of the customer: we need a service that offers an operation for obtaining the credit limit currently assigned to the customer. Please note that, participants can be hierarchically organized with, for instance, silver and golden accounts as well as saving, asset accounts. For simplicity, we addressed just the simple flat case.

5. CONTEXT-AWARE CONCERNS AS TAILORED ECA-DRIVEN RULES

As we emphasized, the main objective of this contribution is to first explicitly separately address different concerns while tackling any context-aware service-oriented business applications. In the previous section, we demonstrated how interaction-centric functionality concerns can be conceptualized as transient ECA-driven connectors. In this section, we similarly present how context-aware concerns need to be extracted from any activity and modelled as tailored ECA-driven "contextual" connectors.

Towards forwarding suitable conceptual primitives for context-awareness and in contrast to functionality concerns, we first require contextual predicates [6] for reasoning about the surrounding environment and any involved devices and so on. In this contribution, we restrict ourselves to the role of *locations in defining* the context-aware behavior governing any activity. Nevertheless, as the reader may easily infer, the approach is flexible enough to be extended with further predicates to cope with device resources and user preferences, such as GPS when banking identification is performed using his/her Mobile device and cameras implanted at the ATM for face-recognition and/or fraud detection for instance. More precisely, we propose three context-aware predicates to allow reasoning about context-awareness in terms of ECA-driven rules.

- **The communication status**, reflecting the presence, absence, or quality of the link between locations where given services are performed but require exchange of knowledge (e.g. data, message). This is, captured through the "connect" predicate CNT:set(LOC) → BOOL. Where LOC stands any location-dependent (concrete or abstract) entity.
- The ability to continue the execution of an activity at another location, which requires that the new location is *reachable* from the present one so that the execution context can be moved. The construct *RC:LOC x LOC → BOOL* is proposed. It informs whether a given location is reachable from another one.
- The spatial relationship between two locations so that triggering events may be initiated. We abstract such predicate as: *Near2(LOC,LOC) → BOOL*.

As already mentioned, In the same manner other contextual predicates can be forwarded for informing and reasoning about resources such as devices memory, display, GPS-aware devices, processor-capabilities and cameras among others. We have introduced the three above just for illustration, but we are working on a complete set of primitives for reasoning about different context-aware situations. Similar primitives can be found in [6] and [14] among others.

As for interaction concerns, we propose *ECA-driven context* rules as conceptual architectural connectors to cope with the context-awareness, through the explicit use of the above contextual predicates. More precisely, to be compliant with the followed event-driven paradigm, we let unchanged the triggering primitive, that is, any context-aware rule will start like the interaction rule with the *at-trigger* primitive. To emphasize that now the constraints to be involved for any context-aware involve one or more contextual primitives (to test the current surrounding environment), we propose to split the condition part into two parts. A first part starting with where concerns the testing of the status of surrounding context. The second part concerns the usual conditions, though now dealing with context-awareness issues; we introduce such conditions by starting with under-cxt. Finally, the actions to perform are to be prefixed by act-cxt.

5.1. Context-awareness Concerns graphically illustrated and explained

In the following, we illustrate these context-aware concerns by considering the same activities we addressed at the functionality concern. More precisely, we first consider the withdrawal then the identification activities (which is fully context-dependent).

But before detailing that context concerns, let us again motivate more on the explicit and strict generic separation of these two concerns (i.e. functionalities and context-awareness) while modeling business activities. As depicted in Figure-5 and still with respect to the banking application, we have two strict yet complementary concerns while describing any activity of this application. That is, on the left hand-side, the behavioral functionality issues are to be expressed at the interaction level using (ordinary) business entities such as: (different kinds) customers and accounts.

On the right-side, while we always stick to the modeling at the composition-level to promote adaptability, the entities coming into play are more intrinsic contextual-aware entities: such as ATM, INTERNET, CARD, and so forth. These entities are at-least location-aware, where the context primitives influence by part the interaction.

Finally, it is worth pointing out that these two concerns are to be semantically related to reflect any activity as will be seen later. For instance, we will be speaking about *customer@atm*, *customer@internet*, *account@bank*, *card@atm* and so forth. That is, the business entities will be using or residing in associated context-aware entities. We finally, point out that such separation of entities has been recently reported in [29] as we have already detailed in the related work, although without emphasis on the composition as first-class neither at the fine-grained activity-level.

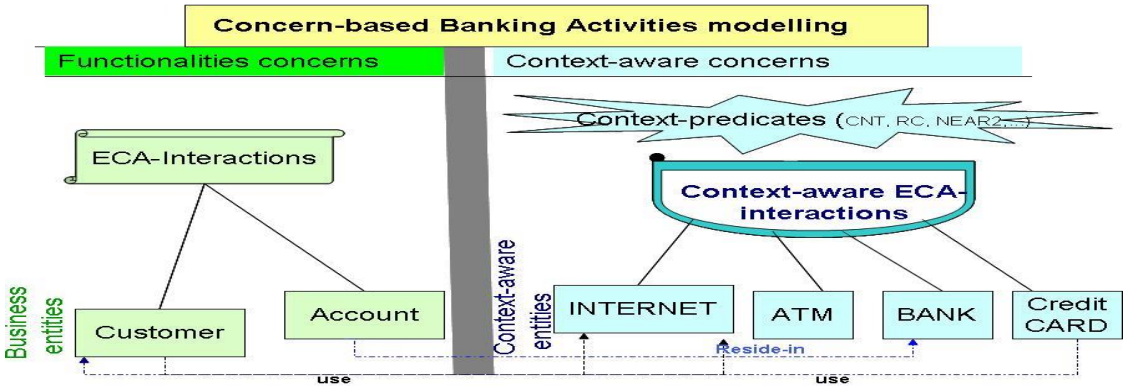

Figure 5. Graphically illustrated how the ECA-based Functionality-Concerns may Interact with the Context Ones for the banking Example

Now coming back to the specific activities in our simplified banking business process, that are, the *identification* and the *withdrawal* activities. Let us start with the withdrawal activity, we already discussed its behavioural functionalities at the interaction-level. Indeed, when we described these ECA-driven functionalities, we did not mention at all what are the contextual situations governing this activity. We were not concerned with the adopted business *channels* as described in the right-hand side of Figure-5 nor with *where* such withdrawal was taking place.

Example 3 (Context concerns for the Withdraw Activity): The conceptualization as shown in Figure-6 details the behavioural added-values and / or restrictions to be observed when baking at an ATM. Indeed, first as involved context-aware entities, we should have the ATM and BANK. From the ATM, we implicitly require its location but also properties such as the available cash and the default amount, both as hidden. An event for triggering the withdrawal is further required; but from the context-aware perspective, we do not care whether it is initiated automatically or from the user by pressing a specific button. Finally, the ATM is to be able to deliver money when the following constraints hold. First, as the first rule details, if the ATM is in a full connection with the corresponding BANK, the withdrawal is performed with the requested amount unless it surpasses the agreed-on maximum to withdraw or no enough cash is available. The second rule in contrast concerns the case where no connection is available (i.e. CNT(ATM, BANK) is false). In this case, only a default amount is allowed using the conditions that such ATM is endowed with off-line reachability to report on the performed transaction later (i.e. RC(ATM, BANK) is true). That is, the transaction is to be moved or migrated later using a banking operation such mv(wdr-op(atm-internal), bank).

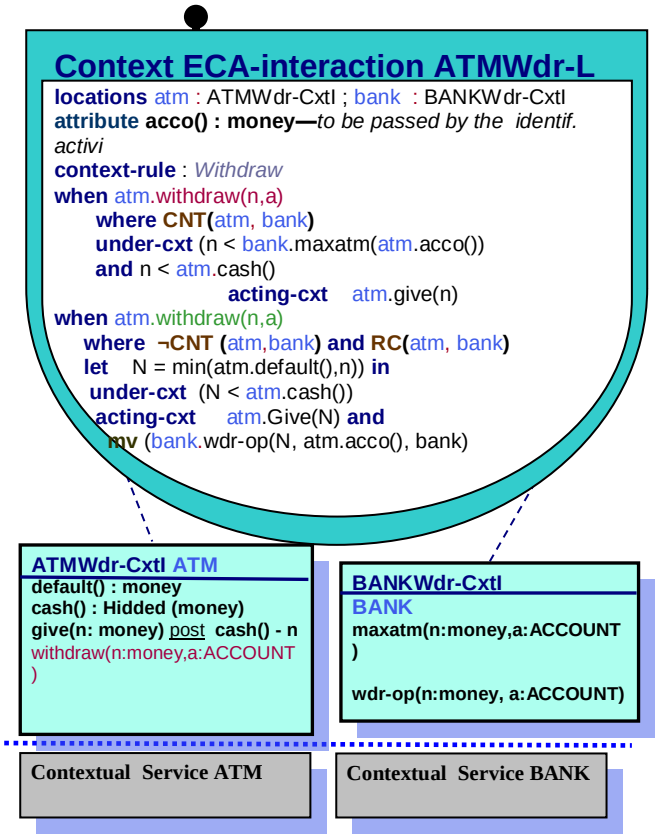

Figure 6: The ECA-based architectural Context-aware Rule for the Withdraw Activity

Example 4 (Context concerns for the Identification Activity): With the aim to illustrate more the crucial importance of context-awareness concerns, let us consider now the identification activity in any banking business process which in our case must precede the withdrawal activity. From the functionality's perspective, we did not skip it, instead there has been *nothing* to functionally describe for this identification activity. In other words, the identification is purely context-aware activity as it concerns the exclusive interaction of context-aware entities such as: INTERNET, PDA, ATM, CARD and where the connections and the other contextual primitives are decisive in defining the associated behaviour. As detailed in Figure-6, for the case of banking

at the ATM, that is, the identification via a Bank-CARD, we require from the ATM the ability to accept /(r)eject Bank-Cards and to enter Pins. We note that the accept message records the account and customer when successful. From the CARD, the hidden coded should present as well as the acceptance and rejection. We also require that the account number and the customer ids to be offered from the Card-magnetic. The corresponding ECA-driven interaction-rule says that we have first to enter the triggering event *EnterPin(Num)*. If the ATM reacts to that CARD, that is either the CARD is inserted or simply it is just near it (via infrared or Bluetooth connection), the entered code is checked with the stored card-code. Then, three attempts are allowed as possible ATM capabilities. That is, even this ATM withdrawal rule could have several variants reflecting specific ATM capabilities and Card (holder) specificities.

We similarly note that, the identification could be done via Internet / PDA. In this case we must enter password under the constraints that a LAN or WLAN connection is available.

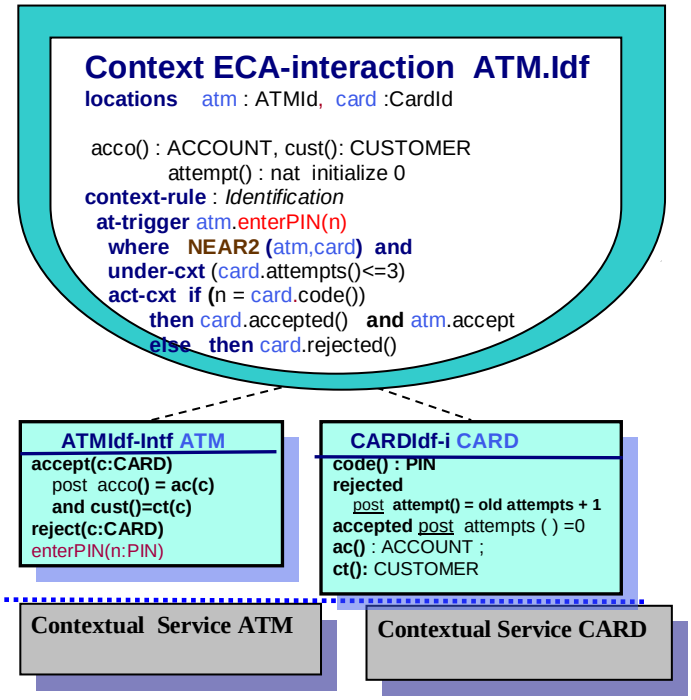

Figure 7. The ECA-based architectural Context-aware Rule for the Identification Activity

6. INTEGRATION OF CONCERNS: ACTIVITIES AND PROCESS MODELLING REFERENCES

In the two previous sections, we separately dealt with functionalities and context-awareness while modeling any activity as highly adaptive and behavioral tailored ECA-driven interactions. In this sense, both concerns can be modeled, evolved and reasoning about completely in separate manner, enhancing thus the mastering the application complexity and evolution.

Nevertheless, once such concerns and with respect to any involved business activity, in each service-driven business application are developed, we require bring them together to reflect the complete and intuitive business semantics we aim for any business activity. Being able to model these concerns separately does not thus mean that they are independent. The way a business activity is performed within a process system emerges from the functionalities as well as from the contextual ECA-driven rules that jointly apply to that activity. Indeed, whereas and still at the

conceptual-level the "What" question reflects the functionality concerns, the "Where" with its "How" capabilities (such as sensors, actuators, cameras, etc.) should reflect the context-awareness of any activity.

6.1. The Intrinsic Integration of functionalities and Context-awareness graphically illustrated and explained

To be more illustrative, when banking at ATM for instance, we have on the one side the withdrawal functionalities reflecting the intended interaction of the customer with its customer to perform the right withdrawal. On the other side, we have the added-value of opting for a withdrawal using the ATM, that is, the context-aware interactions while withdrawing.

As illustrated in Figure-7, with respect to this withdrawal activity, it is more logical at the end to speak about *customer@atm* and *account@bank*. That is, we have to bring together the functionalities and context-awareness ECA-driven rules together while running any withdrawal activity using the ATM. As shown in the picture, for different customers and ATMs, we may have different instances running each with the right functionalities and context-aware rules. Since that we were coherently using ECA-driven rules, this integration of concerns around activities is not that much difficult and become very intuitive. Indeed, we have just to join together different clauses as conjunction. More precisely, first the events require to be unified by integrating all their parameters. Then, all conditions in both selected rules have to gathered as conjunction. Finally, all actions in both selected rules have to performed.

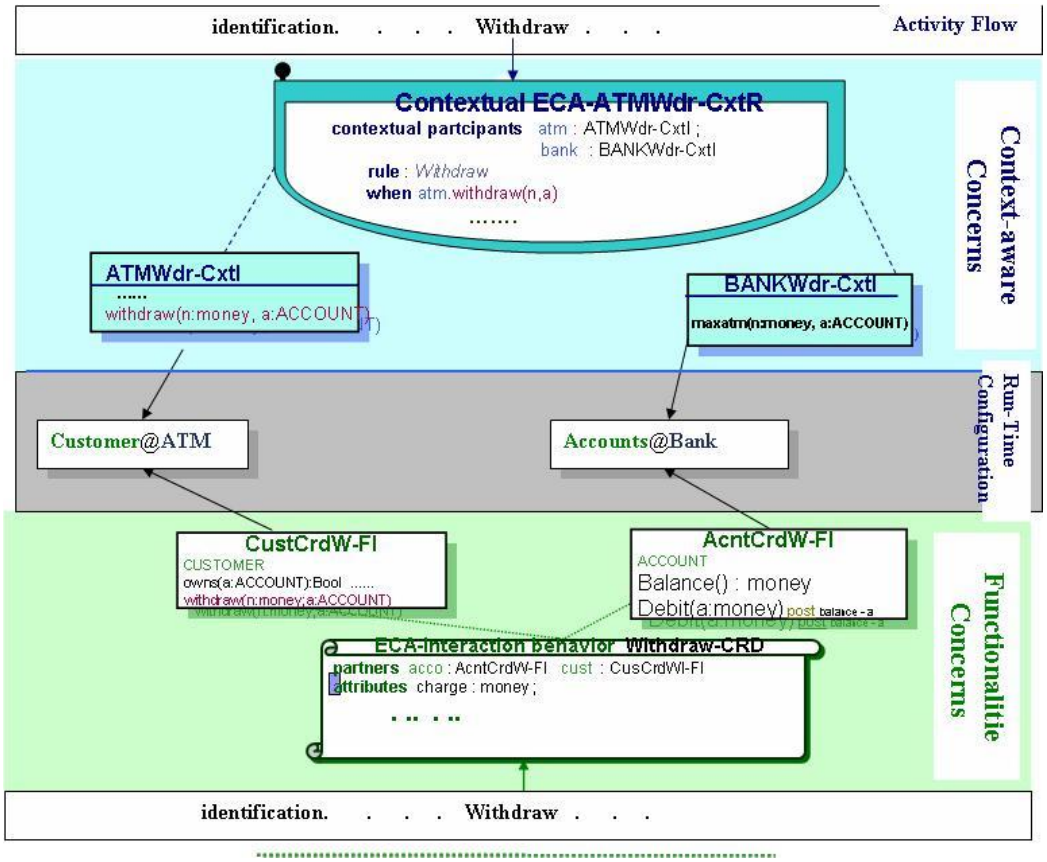

Figure 8. The Architectural Integration of the both Functional and Context-based Concerns Graphically Explained for the Withdrawal activity

Example 5 (Concerns Integration for the Withdraw Activity): The integrated rule of a withdrawal at ATM with customer enjoying a credit withdrawal could thus be represented as detailed in Figure-8. That is, first we have to unify the withdrawal triggering to become cust@atm.withdraw(cs, m). Then, we have to check that a connection between the ATM and corresponding Bank is holding (if not the integration concern the second context-aware rule of the withdrawal as given above). First, we have to check that the functionality rule holds, that is, there is enough money in the account plus the credit and that the customer is owning that account

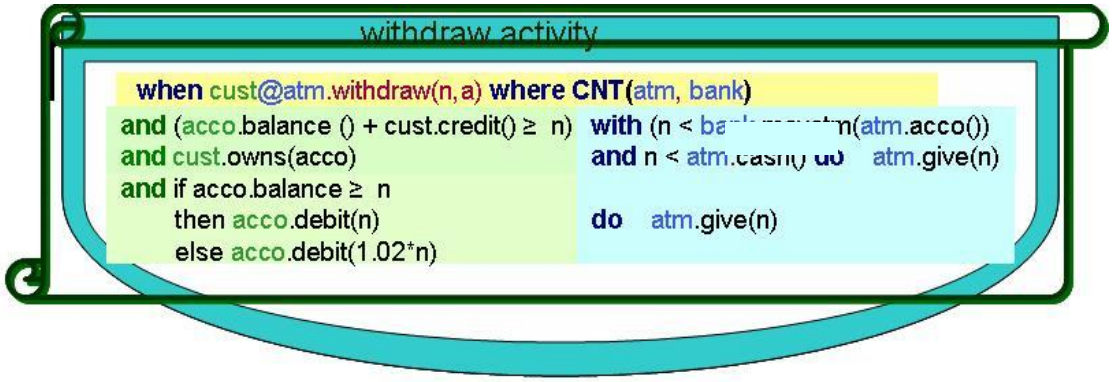

Figure 9. The Integration of the Functional and Context ECA-Driven Rules for the Withdrawal Activity

On the ATM context-aware side, we have to verify that the customer is allowed to withdraw such amount from the ATM and there is enough cash in that ATM. When all these constraints are holding, the account is debited, and the ATM deliver that requested amount.

Example 6 (Concerns Integration for the Identification Activity): We stress again that the integration of the identification activity of both concerns remains purely context-aware as no functionalities are bounded to that activity, nevertheless, we have to adapt the previous context-aware ECA-rule so that, for instance, the trigger atm.enterPIN(n) has to be changed to cust@atm.enterPIN(n), the condition from **NEAR2 (**atm,card**)** to **NEAR2 (cust@**atm,card@atm**)** and so on. That is the identification activity should look like as reflected in Figure-9 below.

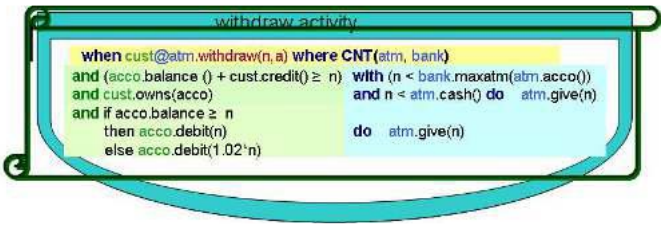

Figure 10. The Integration of the (Functional and) Context ECA-Driven Rules for the Identification Activity

7. FORMAL VALIDATION USING MAUDE AND REWRITING LOGIC

The semantical framework we are proposing for this new concern-based ECA-compliant service oriented architectural conceptualization is based on rewriting logic [Mes92], which has been proved very appropriate for dealing with concurrent systems. Further strengths making this logic very practical is the current implementation of the MAUDE language [3, 8]. In MAUDE object states in are conceived as terms — precisely as tuples— of the form *<Id : C|at1 : v1, .., atk :*

vki>. In this tuple : Id stands for object identity; C identifies an object class; and *at1, ..., atk* denote attribute identifiers with *v1, ..., vk* as current values. Messages (i.e. method invocation) are regarded as operations sent or received by objects, and their generic sort is denoted Msg. Object and message instances flow together in the so-called configuration, which is a multiset, w.r.t. an associative commutative operator denoted by ' ', of messages and (a set of) objects. The effect of messages on objects is captured by appropriate rewrite rules.

Example 7 (The validation of the Withdrawal Rules Using Maude): Without delving into detail about how to we systematically allow deriving the rewrite rules from the ECA-driven interactions of both concerns, we sketch here directly the Maude code corresponding to the withdrawal case, which should look like this module as illustrated in Figure-9.

```
omod WithdrawalAway is
  protecting Money .
  sub-sorts CustId   AcntId < OID .
  ***** participants
  class Account | Bal :  Money.
  class Customer | Own :  CustId .
  event Withdraw :  CustId  Money → Events
  msg-loc Debit :  AcntId  Money → Msg .
  Vars M, Max.  , Charge :  Money .
```

$$[\texttt{WdrAwy}]: Withdraw(Cs, M)\ \langle Ac : AcntId | Bal(Ac) \rangle\ \langle Cs : CustId | Own(Cs, Ac) : True \rangle \Rightarrow \langle Ac : AcntId | Bal(AcntId) \rangle \langle Cs : CustId | Own(Cs, Ac) : True \rangle debit(Ac, M + Charge)\ \texttt{if}\ (M > Max.) \wedge (Bal(Ac) > M)$$

Figure 9. The Validation of the Withdrawal Rules using the Maude Language

Then using the current implementation and environment of the Maude language we can run this specification with respect to concrete agreements between different customers and their respective accounts. The first aim is to check for ambiguity and misconception. Then, as second level we have to tackle inconsistency and conflict between different rules. As third aim, and because Maude is endowed with model-checking properties can be verified. We have to do that independently with respect to both functionalities and contextual concerns. Finally, we have to tackle the integration as we informally described and check again the above issues such misconception, conflict and crucial properties at that formal level.

A detailed specification and validation of the functionalities ECA-driven rules we implemented using the Windows Maude Workstation are shown in Appendix-A.

In order to dynamically integrate different functional and context-aware rules, we take benefits of the reflection of rewrite logic and its implementation as so-called strategies in the Maude language. The following Figure depicts an illustration how different rules can be combined.

The semantical framework we are proposing for this new concern-based ECA-compliant service oriented architectural conceptualization is based on rewriting logic [Mes92], which has been proved very appropriate for dealing with concurrent systems. Further strengths making this logic very practical is the current implementation of the MAUDE language [3, 8]. In MAUDE object states in are conceived as terms — precisely as tuples— of the form <Id : C|at1 : v1, .., atk : vki>. In this tuple : Id stands for object identity; C identifies an object class; and at1, ..., atk denote attribute identifiers with v1, ..., vk as current values. Messages (i.e. method invocation) are regarded as operations sent or received by objects, and their generic sort is denoted Msg.

Object and message instances flow together in the so-called configuration, which is a multiset, w.r.t. an associative commutative operator denoted by ' ', of messages and (a set of) objects. The effect of messages on objects is captured by appropriate rewrite rules.

Without delving into detail about how to we systematically allow deriving the rewrite rules from the ECA-driven interactions of both concerns, we sketch here directly the Maude code corresponding to the withdrawal case, which should look like this module.

Then using the current implementation and environment of the Maude language we can run this specification with respect to concrete agreements between different customers and their respective accounts. The first aim is to check for ambiguity and misconception. Then, as second level we have to tackle inconsistency and conflict between different rules. As third aim, and because Maude is endowed with model-checking properties can be verified. We have to do that independently with respect to both functionalities and contextual concerns. Finally, we have to tackle the integration as we informally described and check again the above issues such misconception, conflict and crucial properties at that formal level.

A detailed specification and validation of the functionalities ECA-driven rules have implemented as depicted in the following Figure A-1 (in Appendix-A). We should further point out that the execution of these rules can be dynamically controlled using the so-called strategy as a reflection-based manner to control the order in which different rules can be executed. An illustration of such strategy for the withdrawal Activity is depicted in Figure A-2 in the Appendix.

8. CONCLUSIONS

In this paper, we put forwards a service-oriented architectural-based approach that addresses current challenges in modern business process modelling for reflecting dynamic cross- and intra-organisational interactions as well as context-aware dependencies. We proposed ECA-driven semantics primitives to separately model, evolve and validate both concerns at the activity-level. We further explained how these concerns to-be integrated to reflect the intuitive business semantics of any business activity. Rewriting logic and its Maude language have been proposed for the formal validation and verification of both concerns. Furthermore, in order to dynamically integrate different functional and context-aware rules, we have taken benefits of the reflection of rewrite logic and its implementation so-called strategies in the Maude language.

To further consolidate and validate this service-oriented architectural approach we are working on more case studies. Among the most interesting and practical fields that we are working on is the healthcare, where context-ware beds, specific heart-devices and other context-intensive medical tools are becoming nowadays more that ubiquitous [5, 27]. We are also implementing the different phases of the approach. One of our main goals is to develop a deeper understanding and classification of business rules so that semi-automatic derivation of functionalities and context-aware architectural ECA-driven connectors can be ultimately achieved.

REFERENCES

[1] R.Allen and D.Garlan, "A Formal Basis for Architectural Connectors", ACM TOSEM, 6(3), 1997, 213-249.

[2] N.Aoumeur, J.Fiadeiro and C.Oliveira, "Distribution concerns in service-oriented modelling" Int. J. Internet Protocol Technology, Vol. 1, No. 3, 2006.

[3] M. Clavel, F. Duran, S. Eker, P. Lincoln, N. Marti-Oliet, J. Meseguer, and C.L: Talcott. All About Maude - A High-Performance Logical Framework, How to Specify, Program and Verify Systems in Rewriting Logic. Lecture Notes in Computer Science (springer), 4350, 2007.

[4] P.Cong Vinh, N.Tat Thanh and H.Chi Minh (eds.), "Context-aware Systems and Applications, and Nature of Composition and Communication", 7th EAI International Conference, iccasa 2018 and 4th EAI International Conference, ICTCC 2018 Proceedings, Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, Springer

[5] P.Cipresso, s.serino and D.Villani (Eds). "Computing Paradigms for Mental Health.", 9th International Conference, MindCare 2019, Buenos Aires, Argentina, April 23–24, 2019 Proceedings, Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, 2019.

[6] P.C Dockhorn, P.A. Almeida, L.F. Pires, and M. van Sinderen. "Situation Specification and Realization in Rule-Based Context-Aware Applications." In Proc. of the Int. Conference DAIS'07, page 3247. LNCS, Volume 4531, 2007.

[7] C.Dobre and F.Xhafa. "Pervasive Computing Next Generation Platforms for Intelligent Data Collection." Academic Press is an imprint of Elsevier. Elsevier, 2016.

[8] F. Durán, S. Eker, S. Escobar, N. Martí-Oliet, J. Meseguer, R. Rubio, C. L. Talcott: Programming and symbolic computation in Maude. J. Log. Algebraic Methods Program. 110, 2020.

[9] A.Juan Fuente, , B. López Pérez, G. Infante Hernández and L.J.Cases Fernández, "Using rules to adapt applications for business models with high evolutionary rates.", International Journal of Artificial Intelligence and Interactive Multimedia, Vol. 2, Nº 2, 2013.

[10] Business Rules Group. Defining Business Rules - What Are They Really? In www.businessrulesgroup.org, 2005.

[11] P.Kardasis and P.Loucopoulos, "Expressing and Organising Business Rules", Information and Software Technology, 2006.

[12] J. Meseguer. Conditional rewriting logic as a unified model for concurrency. Theoretical Computer Science, 96:73–155, 1992.

[13] J.Magee and J.Kramer, "Dynamic Structure in Software Architectures", 4th Symp. on Foundations of Software Engineering, ACM Press 1996, 3-14.

[14] E.Marius Oprea, M.Alexandru Moisescu and S.Caramihai, "Context Awareness in Enterprise Systems Design, May 2021." In 23rd International Conference on Control Systems and Computer Science (CSCS), 2021, DOI: 10.1109/CSCS52396.2021.00053.

[15] L.Mutanu and G.Kotonya, "State of runtime adaptation in service-oriented systems: what, where, when, how and right.", Special Issue: Adaptive and Reconfigurable Service-Oriented, Cloud and Virtualised Architectures, IET Software, Vol. 13 Iss. 1, pp. 14-24.

[16] G.J. Nalepa and S.Bobek "Rule-Based Solution for Context-Aware Reasoning on Mobile Devices.", Computer Science and Information Systems 11(1):171–193, 2013.

[17] B.Orrinsi, J.Yang, and M.Papazoglou, "A Framework for Business Rule Driven Web Service Composition", in Proc. of Conceptual Modeling for Novel Application Domains, LNCS 2814 Springer 2003, 52-64.

[18] J.Oukharijane, I.Ben Said, M.Chaâbane, R.Bouaziz, Rafik and E.Andonoff, Eric.. "A Survey of Self-Adaptive Business Processes", In 32nd International Business Information Management Association Conference (IBIMA).", Seville, Spain, Feb 2019.

[19] M.Papazoglou and D.Georgakopoulos (guest editors), Special Issue on Service-Oriented Computing, Communications of the ACM 46(10), 2003.

[20] D.Rosca and C.Wild, "Towards a Flexible Deployment of Business Rules", Expert Systems with Applications 23:385--394, 2002.

[21] M.P. Papazoglou. Web Service: Principles and Technology. Prentice-Hall, Englewood Cliffs,2007.

[22] RuleML: "The Rule Markup Initiative." http://wiki.ruleml.org /index.php /RuleML_Home. 2021

[23] O.Vasilecas, D.Kalibatiene and D.Lavbič, "Rule- and context-based dynamic business process modelling and simulation", Journal of Systems and Softwarem 122, 2016.

[24] W.Wan-Kadir and P.Loucopoulos, "Relating Evolving Business Rules to Software Design", Journal of Systems Architecture, 2003.

[25] T.Elr, "Service-Oriented Architecture: Analysis and Design for Services and Microservices.", Printice-Hall, 2016.

[26] O.Vasilecas, D.Kalibatiene and D.Lavbič, "Rule- and context-based dynamic business process modelling and simulation", Journal of Systems and Software, Volume 122, December 2016, Pages 1-15.

[27] J.Symonds, " Ubiquitous and Pervasive Computing: Concepts, Methodologies, Tools, and Applications", Copyright © 2010 by IGI Global, 2010.

[28] M.Guo, J.Zhou, F.Tang, and Y.Shen, "Pervasive computing : concepts, technologies and applications", Taylor & Francis Group, LLC, 2017.
[29] D.Lupiana, "Architectural Solutions for Context-Aware Applications: KoDA Prototype", International Journal for Information Security Research (IJISR), Volume 9, Issue 1, March 2019

APPENDIX-A

ACNT_CMP_GNR.maude

```
1.  mod ACNT_CMP is
2.     protecting INT .
3.     inc CMP_GNR .
4.     sorts CRDT DBT CHGL TRS His HisL AcntCf AcntId .
5.     subsorts CRDT DBT TRS < obs_Msg .
6.     subsort  CHGL < loc_Msg .
7.     subsorts His < HisL < loc_Msg .
8.     subsort AcntCf < ConfCMP .
9.     subsort AcntId < CMPid .
10.    op Crd(_, _) : AcntId Int -> CRDT [ctor] .
11.    op Db(_, _) : AcntId Int -> DBT [ctor] .
12.    op ChgL(_, _) : AcntId Int -> CHGL [ctor] .
13.    op Trs(_, _, _) : AcntId AcntId Int -> TRS .
14.    op bal: _ : Int -> obs_Prop [ctor gather (&)] .
15.    op limt: _ : Int -> loc_Prop [ctor gather (&)] .
16.    op [] : -> His .
17.    op [_, _] : Int Nat -> His .
18.    op _, _ : His HisL -> HisL .

19.    vars A A1 : AcntId .
20.    vars B B1 L L1 : Int .
21.    var M : Nat .
22.    rl [credit] : Crd ( A, M ) < A | bal: B > => < A | bal: B + M > .
23.    rl [debit] : Db (A, M) < A | bal: B > => < A | bal: B - M > .
24.    rl [chgl] : ChgL ( A, L1 ) < A | limt: L > => < A | limt: L1 > .
25.    rl [transfer] : Trs(A, A1, M) < A | bal: B > < A1 | bal: B1 >
26.                 => Db (A, M) < A | bal: B > Crd ( A1, M ) < A1 | bal: B1 > .
27.  endm
```

Figure A-1. The Implementation of the Withdrawal Rules using the Windows Maude Workstation

ACNT_STR.maude

```
mod ACNT_STR is
inc ACNT_CONF .
protecting META-LEVEL .
vars withdraw? deposit? transfer? : [Result4Tuple] .
var T : Term .
op Compute : Term -> Term .

ceq Compute(T)
 = (if(deposit? :: Result4Tuple)
   then Compute(getTerm(deposit?))
   else if(transfer? :: Result4Tuple)
        then Compute(getTerm(transfer?))
        else if(withdraw? :: Result4Tuple)
             then Compute(getTerm(withdraw?))
             else T
        fi fi fi)

  if withdraw? := metaXapply(upModule('ACNT_CONF,false),
              T,'withdraw,none,0,unbounded,0)
    /\ deposit? := metaXapply(upModule('ACNT_CONF,false),
              T,'deposit,none,0,unbounded,0)
    /\ transfer? := metaXapply(upModule('ACNT_CONF,false),
              T,'transfer,none,0,unbounded,0) .

eq Compute(T) = T [owise] .
endm
```

Figure A-2. A Strategy-based Implementation of The Withdrawal Rules using the Strategy-Module of the Windows Maude Workstation

© 2022 By AIRCC Publishing Corporation. This article is published under the Creative Commons Attribution (CC BY) license.

AUTHOR INDEX