ICIW 2019

The Fourteenth International Conference on Internet and Web Applications and Services

ISBN: 978-1-61208-728-3

July 28 – August 2, 2019

Nice, France

ICIW 2019 Editors

Pascal Lorenz, University of Haute-Alsace, France

ICIW 2019

Forward

The Fourteenth International Conference on Internet and Web Applications and Services (ICIW 2019), held between July 28, 2019 and August 02, 2019 in Nice, France, continued a series of co-located events that covered the complementary aspects related to designing and deploying of applications based on IP & Web techniques and mechanisms. It focused on Web technologies, design and development of Web-based applications, and interactions of these applications with other types of systems. Management aspects related to these applications and challenges on specialized domains were also considered. Evaluation techniques and standard positions on different aspects were part of the agenda.

Internet and Web-based technologies led to new frameworks, languages, mechanisms and protocols for Web applications design and development. Interaction between web-based applications and classical applications requires special interfaces and exposes various performance parameters.

Web Services and applications are supported by a myriad of platforms, technologies, and mechanisms for syntax (mostly XML-based) and semantics (Ontology, Semantic Web). Special Web Services based applications such as e-Commerce, e-Business, P2P, multimedia, and GRID enterprise-related, allow design flexibility and easy to develop new services. The challenges consist of service discovery, announcing, monitoring and management; on the other hand, trust, security, performance and scalability are desirable metrics under exploration when designing such applications.

Entertainment systems became one of the most business-oriented and challenging area of distributed real-time software applications' and special devices' industry. Developing entertainment systems and applications for a unique user or multiple users requires special platforms and network capabilities.

Particular traffic, QoS/SLA, reliability and high availability are some of the desired features of such systems. Real-time access raises problems of user identity, customized access, and navigation. Particular services such interactive television, car/train/flight games, music and system distribution, and sport entertainment led to ubiquitous systems. These systems use mobile, wearable devices, and wireless technologies.

Interactive game applications require particular methodologies, frameworks, platforms, tools and languages. State-of-the-art games today can embody the most sophisticated technology and the most fully developed applications of programming capabilities available in the public domain.

The impact on millions of users via the proliferation of peer-to-peer (P2P) file sharing networks such as eDonkey, Kazaa and Gnutella was rapidly increasing and seriously influencing business models (online services, cost control) and user behavior (download profile). An important fraction of the Internet traffic belongs to P2P applications.

P2P applications run in the background of user's PCs and enable individual users to act as downloaders, uploaders, file servers, etc. Designing and implementing P2P applications raise particular requirements. On the one hand, there are aspects of programming, data handling, and intensive computing applications; on the other hand, there are problems of special protocol features and networking, fault tolerance, quality of service, and application adaptability. Additionally, P2P systems require special attention from the security point of view. Trust, reputation, copyrights, and intellectual property are also relevant for P2P applications. On-line communications frameworks and mechanisms allow distribute the workload, share business processes, and handle complex partner profiles. This requires protocols supporting interactivity and real-time metrics.

Collaborative systems based on online communications support collaborative groups and are based on the theory and formalisms for group interactions. Group synergy in cooperative networks includes online gambling, gaming, and children groups, and at a larger scale, B2B and B2P cooperation. Collaborative systems allow social networks to exist; within groups and between groups there are problems of privacy, identity, anonymity, trust, and confidentiality. Additionally, conflict, delegation, group selection, and communications costs in collaborative groups have to be monitored and managed. Building online social networks requires mechanism on popularity context, persuasion, as well as technologies, techniques, and platforms to support all these paradigms.

Also, the age of information and communication has revolutionized the way companies do business, especially in providing competitive and innovative services. Business processes not only integrates departments and subsidiaries of enterprises but also are extended across organizations and to interact with governments. On the other hand, wireless technologies and peer-to-peer networks enable ubiquitous access to services and information systems with scalability. This results in the removal of barriers of market expansion and new business opportunities as well as threats. In this new globalized and ubiquitous environment, it is of increasing importance to consider legal and social aspects in business activities and information systems that will provide some level of certainty. There is a broad spectrum of vertical domains where legal and social issues influence the design and development of information systems, such as web personalization and protection of users privacy in service provision, intellectual property rights protection when designing and implementing virtual works and multiplayer digital games, copyright protection in collaborative environments, automation of contracting and contract monitoring on the web, protection of privacy in location-based computing, etc.

The conference included the following tracks:

- Service computing
- Trends on Internet-based data, applications and services
- Web Services-based Systems and Applications

We take here the opportunity to warmly thank all the members of the ICIW 2019 technical program committee, as well as all the reviewers. The creation of such a high quality conference program would not have been possible without their involvement. We also kindly thank all the authors who dedicated much of their time and effort to contribute to ICIW 2019. We truly believe that, thanks to all these efforts, the final conference program consisted of top quality contributions.

We also thank the members of the ICIW 2019 organizing committee for their help in handling the logistics and for their work that made this professional meeting a success.

We hope that ICIW 2019 was a successful international forum for the exchange of ideas and results between academia and industry and to promote further progress in the area of Web applications and services. We also hope that Nice, France provided a pleasant environment during the conference and everyone saved some time to enjoy the charm of the city.

ICIW 2019 Chairs

ICIW Steering Committee

Stefanos Gritzalis, University of the Aegean, Greece
Sebastien Salva, UCA (Clermont Auvergne University), France
Raj Jain, Washington University in St. Louis, USA
Jian Yu, Auckland University of Technology, New Zealand
Christoph Meinel, Hasso-Plattner-Institut GmbH, Germany

ICIW Industry/Research Advisory Committee

José Luis Izkara, TECNALIA, Spain
Christos J. Bouras, University of Patras, Greece
Alex Ng, Cybersecurity Program | CS&IT | La Trobe University, Australia
Rema Hariharan, eBay, USA

ICIW 2019

Committee

ICIW Steering Committee

Stefanos Gritzalis, University of the Aegean, Greece
Sebastien Salva, UCA (Clermont Auvergne University), France
Raj Jain, Washington University in St. Louis, USA
Jian Yu, Auckland University of Technology, New Zealand
Christoph Meinel, Hasso-Plattner-Institut GmbH, Germany

ICIW Industry/Research Advisory Committee

José Luis Izkara, TECNALIA, Spain
Christos J. Bouras, University of Patras, Greece
Alex Ng, Cybersecurity Program | CS&IT | La Trobe University, Australia
Rema Hariharan, eBay, USA

ICIW 2019 Technical Program Committee

Ado Adamou Abba Ari, University of Maroua, Cameroun
Mohd Helmy Abd Wahab, Universiti Tun Hussein Onn, Malaysia
Witold Abramowicz, Poznan University of Economics and Business, Poland
Mehmet Aktas, Yildiz Technical University, Turkey
Grigore Albeanu, Spiru Haret University - Bucharest, Romania
Markus Aleksy, ABB AG, Germany
Pedro Álvarez, University of Zaragoza, Spain
Leonidas Anthopoulos, University of Applied Science (TEI) of Thessaly, Greece
Filipe Araujo, University of Coimbra, Portugal
Ezzy Ariwa, University of Bedfordshire, UK
Jocelyn Aubert, Luxembourg Institute of Science and Technology (LIST), Luxembourg
Arnim Bleier, GESIS - Leibniz Institute for the Social Sciences, Germany
Masoud Barati, Cardiff University, UK
Andres Baravalle, University of East London, UK
Dan Benta, Agora University of Oradea, Romania
Luis Bernardo, Universidade NOVA de Lisboa, Portugal
Christos J. Bouras, University of Patras, Greece
Mahmoud Brahimi, University of Msila, Algeria
Tharrenos Bratitsis, University of Western Macedonia, Greece
Paulo Caetano da Silva, Salvador University - UNIFACS, Brazil
Jorge C. S. Cardoso, University of Coimbra, Portugal
Dickson Chiu, The University of Hong Kong, Hong Kong
Soon Ae Chun, City University of New York, USA
Marta Cimitile, Unitelma Sapienza University, Italy

Gianpiero Costantino, Institute of Informatics and Telematics (IIT) - National Research Council (CNR), Italy
Giuseppe A. Di Lucca, University of Sannio | RCOST (Research Center on Software Technology), Italy
Ronaldo dos Santos Mello, Federal University of Santa Catarina, Brazil
Luis Espinosa-Anke, Cardiff University, UK
Javier Fabra, Universidad de Zaragoza, Spain
Muhammad Fahad, Universite Lumiere LYON 2, France
Johannes Fähndrich, Technische Universität Berlin, Germany
Stefan Fischer, University of Lübeck, Germany
Karla Donato Fook, IFMA - Maranhao Federal Institute for Education, Science and Technology, Brazil
Panayotis Fouliras, University of Macedonia, Greece
Ivan Ganchev, University of Limerick, Ireland / Plovdiv University "Paisii Hilendarski", Bulgaria
Marcos Garcia, University of Coruña, Galicia, Spain
Mohamed Gharzouli, University of Abdelhamid Mehri Constantine 2, Algeria
Katja Gilly de la Sierra - Llamazares, Universidad Miguel Hernández, Spain
Patrizia Grifoni, National Research Council, Italy
Stefanos Gritzalis, University of the Aegean, Greece
Carlos Guerrero, University of Balearic Islands, Spain
Bidyut Gupta, Southern Illinois University, USA
Ileana Hamburg, Institut Arbeit und Technik, Germany
Rema Hariharan, eBay, USA
Mamoun Abu Helou, Al Istiqlal University, Palestine
Martin Henkel, Stockholm University, Sweden
José Luis Herrero, University of Extremadura, Spain
Tzung-Pei Hong, National University of Kaohsiung, Taiwan
Arthur Huang, Tarleton State University, USA
Emilio Insfran, Universitat Politecnica de Valencia, Spain
Naghmeh Ivaki, University of Coimbra, Portugal
José Luis Izkara, TECNALIA, Spain
Raj Jain, Washington University in St. Louis, USA
Marc Jansen, University of Applied Sciences Ruhr West, Germany
Jehn-Ruey Jiang, National Central University, Taiwan
Sokratis K. Katsikas, Center for Cyber & Information Security | Norwegian University of Science & Technology (NTNU), Norway
Fahim Hasan Khan, Military Institute of Science and Technology (MIST), Bangladesh / University of California Santa Cruz, USA
Selma Khouri, ESI, Algeria / ENSMA, France
Mark Kibanov, University of Kassel, Germany
Suhyun Kim, KIST (Korea Institute of Science and Technology), South Korea
Samad Kolahi, Unitec Institute of Technology, New Zealand
George Koutromanos, National and Kapodistrian University of Athens, Greece
Sarasi Lalithsena, Kno.e.sis Center | Wright State University, USA
José Laurindo Campos dos Santos, Instituto Nacional de Pesquisas da Amazônia, Brazil
Longzhuang Li, Texas A&M University-Corpus Christi, USA
David Lizcano Casas, Open University of Madrid (UDIMA), Spain
Malamati Louta, University of Western Macedonia, Greece
Aastha Madaan, University of Southampton, UK
Khalid Mahmood Malik, Oakland University, USA

Sayed Gholam Hassan Tabatabaei, Isfahan University of Technology, Iran
Andon Tchechmedjiev, Laboratoire d’Informatique, Robotique et Microéléctronique de Montpellier (LIRMM) at the University of Montpellier, France
Michele Tomaiuolo, University of Parma, Italy
Orazio Tomarchio, Universita' di Catania, Italy
Vicente Traver, ITACA - Universitat Politècnica de València, Spain
Christos Troussas, University of Piraeus, Greece
Alexander Troussov, Russian Presidential Academy of National Economy and Public Administration, Russia
Žiga Turk, University of Ljubljana, Slovenia
Hamed Vahdat-Nejad, University of Birjand, Iran
Bert-Jan van Beijnum, University of Twente, The Netherlands
Iván P. Vélez-Ramírez, NeoDeck Holdings Corp., Puerto Rico
Xianzhi Wang, University of Technology Sydney, Australia
Jian Yu, Auckland University of Technology, New Zealand
Sherali Zeadally, University of Kentucky, USA
Yongfeng Zhang, University of Massachusetts Amherst, USA
Martin Zimmermann, Lucerne University of Applied Sciences and Arts, Switzerland

Copyright Information

Table of Contents

Model-Driven Engineering of Fault Tolerant Microservices

Elena Troubitsyna
Åbo Akademi University
Turku, Finland
email: Elena.Troubitsyna@abo.fi

***Abstract*—The microservices architectural style has gained a significant popularity over the last few years. It promotes structuring applications as a composition of independent services of small granularity – microservices. Such an approach supports agile development and continuous integration and deployment. However, it also poses a significant challenge in ensuring the required quality of service and, in particular, fault tolerance. It requires a systematic analysis of possible failure scenarios and the use of structured techniques to implementing fault tolerance mechanisms capable of coping with the various types of failures. In this paper, we propose a structured approach to model-driven engineering of fault tolerant applications developed in the microservice architectural style. We define modelling patterns to facilitate the design of appropriate fault tolerance mechanisms. We also discuss how to integrate fault tolerance into the design of complex applications. We demonstrate how to graphically model the micorservice architectures and augment them with various fault tolerance mechanisms. The proposed approach facilitates a systematic analysis of possible failures, recovery actions and design alternatives. Our approach supports structured guided reasoning about fault tolerance at different levels of abstraction and enables efficient exploration of design space. It allows the designers to evaluate various architectural solutions at the design stage that helps to derive clean architectures and improve fault tolerance of developed applications.**

Keywords— microservices; fault tolerance; architecture; graphical modelling; fault tolerance pattern component.

I. INTRODUCTION

Microservices architectural style [11] has gained a significant attention over the last few years. The style builds on the service-oriented computing paradigm [10]. It supports continuous integration and deployment software engineering approach, which makes it a good fit for ubiquitous cloud-based environments. The microservices style was motivated by the need of small autonomous teams of developers, who do not own the full life-cycle of the application development, to continuously integrate and deliver, provide on-demand virtualisation and infrastructure automation.

Microservices aim at overcoming the drawbacks associated with developing, refactoring and maintaining monolithic applications. While supporting a quick and efficient development, integration and modification, the style introduces the additional complexity caused by the need to correctly orchestrate the distributed microservices as well as ensure the desired degree of Quality of Service (QoS).This is a challenging task because the microservices, in general, are developed using different languages and rely on lightweight communication to implement complex application-level scenarios. Therefore, to ensure the desired high degree of QoS and in particular, reliability, we should create an approach that enables a systematic analysis of different failures that might occur in the microservices architectures. Moreover, we should provide the developers with structured techniques to integrate different fault-tolerance mechanisms into the developed applications and analyse their impact.

In this paper, we propose a structured model-driven approach to modelling fault tolerance in the microservice architecture. We rely on Unified Modelling Language (UML) [9] – a popular graphical modelling language – to define patterns for representing different fault tolerance mechanisms and support their structured integration into the application architecture. We define the modelling patterns for representing fault tolerance mechanisms at different levels of abstraction.

We propose static and dynamic fault tolerance mechanisms. The static mechanisms are the structural solutions, which rely on availability of redundant service providers that can be requested to provide services in the case of failures of the main service providers. This mechanism allows the designers to mask failures of the individual service providers. The dynamic fault tolerance mechanisms rely on different monitoring solutions that enable more efficient handling of microservices and communication failures.

We believe that our approach supports structured guided reasoning about fault tolerance and enables efficient exploration of the design space. It allows the designers to evaluate various architectural solutions at the design stage that helps to derive clean architectures and improve fault tolerance of developed complex services.

The paper is structured as follows: in Section II, we discuss the microservices architectural style. In Section III, we propose several fault tolerance mechanisms suitable for the microservice architectures. In Section IV, we introduce modelling of complex composite patterns. Finally, in Section V, we overview the related work and discuss the proposed approach.

II. MICROSERVICES ARCHITECTURAL STYLE

Microservice architectures [11] have emerged as a new architectural style, which aims at overcoming the problems associated with monolithic architecture. In monolithic applications, all functionality is put together to be distributed as a single file. Monolithic applications are simpler to deploy because they usually run on a single machine. Moreover, they are easier to develop because a programmer does not need to deal with abstractions associated with distributed architectures. However, large monolithic applications are hard to maintain, because even a simple refactoring requires rebuilding and redeploying the entire application. Moreover, since a monolith application is usually tightly coupled, failure of even a small part of it leads to the failure of the entire application. Handling runtime failures is especially cumbersome, because all components run in the same environment.

Another approach is taken from the service-oriented architectural style. In Service-Oriented Architecture (SOA) an application consists of independent, interoperable and reusable services, usually implemented as Web services. To facilitate achieving a loose coupling between the components, SOA aims at abstracting of the overall business logic [12].

Each service publishes its description, where it defines its capabilities. A service in SOA typically has one of two main roles – a service provider or service consumer. A service provider is invoked via an external source to provide some services according to its published capabilities. A service consumer (sometimes called service requestor) invokes service providers by sending them corresponding messages. A service can play just a single role in the composed application. It can also play both roles, e.g., if it functions as an intermediary that routes and processes messages or as a service director, which needs to invoke other services to provide a composite service, which is a part of application.

Typically, an application is composed of services that are hosted on different servers. SOA promotes an asynchronous communication to ensure stateless nature of services. Obviously, highly distributed nature of SOA makes development and deployment of services more challenging.

Microservice architecture builds on the concept of SOA. It promotes building an architecture consisting of autonomous and, hence, independently replaceable and upgradable services. Each microservice represents a small component specialising in implementing a certain functionality. Usually, microservices run in their own processes distributed across the network.

Microservice architectures have many benefits including availability, scalability as well as continuous integration and deployment [11]. Next, we discuss a few main characteristics of microservices, which we find particularly useful for achieving QoS of complex applications built in the microservice architectural style.

Single responsibility: the functionality of each microservice is narrowly focused. The main goal is to keep the code base as small as possible and ensure that each microservice can be redeveloped and redeployed in short time. Microservices emphasise the modularity principle, which, nevertheless, allows us to build large applications composed of numerous services.

Autonomy. As mentioned above, in a micorservice architecture, each microservice is typically run on its own process and the processes are distributed across the network. This introduces additional complexity but allows an application to cope with different performance demands and avoid tight coupling.

Heterogeneity. The microservice architecture supports technological independence in implementing each individual microservice, e.g., a programmer is free to choose any programming language to implement a microservice. The API of a microservice should be language-agnostic to ensure that the services can communicate with each other on different platforms.

Scaling. Microservice can be replicated if there is high performance demand or used differently in different parts of the application. Such an approach ensures good scalability of the microservices applications. Microservices have only run-time dependencies on each other and, hence, can be replaced or deployed independently, which further improves scalability and flexibility in developing complex applications.

The growing popularity of the microservices architectural style has led to creating several specialised platforms. Among the most popular platforms are Spring Boot and building on it Spring Cloud, WildFly Swarm, Payara Micro and SilverWare [11]. In addition, there are different libraries, frameworks and application servers, which allow the developers to create the applications in the microservice style.

Since microservices should run in highly distributed environments, their developers should deal with the complexity inherent to all distributed systems, in particular, complexity of achieving fault tolerance and reliable behaviour of the overall developed application. To achieve this goal, we should utilise the knowledge and best practices – design patterns – created in the area of fault tolerant computing and adapt them to the microservices style. In the next section, we focus on discussing various fault tolerance mechanisms and model-driven approach to designing them.

III. Fault-Tolerance In Microservice Architecture

The main goal of introducing fault tolerance in the microservice architecture (and SOA in general) is to prevent a propagation of faults to the application interface level, i.e., to avoid an application failure [7][8]. A fault manifests itself as *error* – an incorrect service state [7][8]. Once an error is detected, an error recovery should be initiated. Error recovery is an attempt to restore a fault-free state or at least to preclude system failure.

Error recovery aims at masking error occurrence or ensuring deterministic failure behaviour if the error cannot be masked. In the former case, upon detection of error, certain actions are executed to restore a fault-free system states and then guarantee normal service provisioning. In the latter case, the service provisioning is aborted and failure response is returned.

In this paper, we focus on the architectural graphical modelling [9] of fault tolerance mechanisms, which can be integrated into the microservice architecture [11]. We demonstrate how to explicitly introduce handling of faulty behaviour into the microservice architecture.

To model a microservice, we should analyse its interactions with the other microservices in the application under development. At the abstract modelling level, we treat a microservice as a black box with the defined logical interfaces. As we mentioned in Section II, in general, each microservice can play one of two roles – service consumer or service provider. Figure 1 and Figure 2 show the patterns for modelling service provider and service consumer correspondingly.

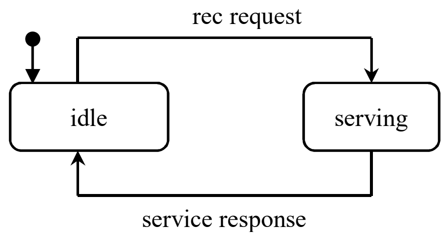


Figure 1. Behavior of service provider

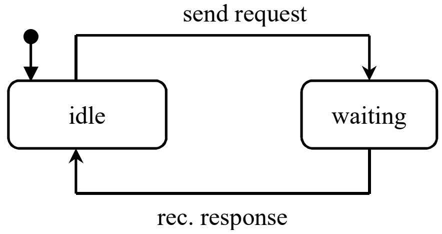


Figure 2. Behavior of service consumer

To model a microservice, we should analyse its interactions with the other microservices in the application under development. At the abstract modelling level, we treat a microservi

A high-level state diagram of the service provider is depicted in Figure 1. The microservice is idle and upon receiving a service request enters the state serving. When the requested computation is completed, the service provides the requested outcome and returns to the state idle.

A high-level state diagram of the service consumer is depicted in Figure 2. Similarly, the data consumer is activated after is issues the service request and upon receiving the requested results returns to the state *idle*.

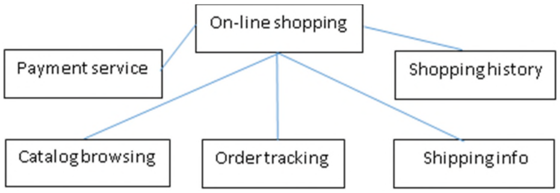


Figure 3. Example of microservice architecture

A generic microservice architecture can be represented by a diagram similar to the one shown in Figure 3. The scenarios to be supported can be modelled as a sequence of service requests and replies. Correspondingly, the microservices involved into an execution of the scenario play the roles of service providers and service consumers.

Let us analyse the possible failures that might occur while executing an application composed of microservices. We can identify three classes of failures:

1. Invalid service request
2. Invalid service response
3. Network failure

The first class of failures caused by an error in the request parameters. This failure can be easily detected during the scenario execution by the explicit response indicating the error. The second class of failures – the invalid service response – is caused by a logical error in implementing a certain microservice. This type of error can be detected by integrating the corresponding functionality into the service consumer that checks the validity of the obtained response.

Finally, the network failures are not caused by the microservices themselves. They can only be detected by integrating the corresponding monitoring mechanisms, for instance, timeouts, into the microservice architectures as well as the appropriate fault tolerance patterns, which we discuss next.

Timeout pattern. This pattern aim at coping with unreliable networks. As we discussed previously, a microservice architectural style promotes loose coupling of services, which results in a highly distributed execution of scenarios. In general, the network is unreliable and hence, connections might fail or become slow. Such network behaviour negatively effects the executions relying on the synchronous remote calls, i.e., can deadlock the scenario execution.

To prevent this, timeouts can be used to bound the waiting time. Despite the fact that timeout mechanisms are actively used at the operating systems level, their use at the application level is less common.

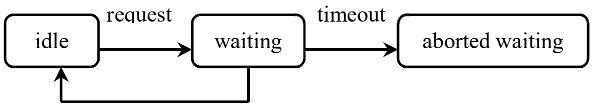


Figure 4. Timeout pattern for service consumer

Figure 4 graphically depicts the timeout pattern with respect to the service consumer. We have decided to single out the state of failed response, since it would allow us to explicitly collect data about failed responses when we study more complex fault tolerance patterns.

The interactions between the service consumer and service provider in the timeout pattern are shown in Figure 5.

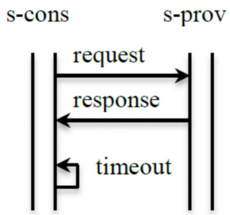


Figure 5. Interaction in timeout pattern

Circuit breaker. Circuit breakers detect excessive current in electric circuits and by failing open the circuit to prevent the connected appliances from damage. When the excessive current is removed, the circuit breaker can be reset and the circuit becomes closed and functioning again.

The idea of circuit breaker in the microservice architecture is similar to the electric one. It is a wrapper, which intercepts the erroneous (and potentially dangerous) calls to make sure that they do not harm the entire application. Hence, the main purpose of the circuit breaker is to monitor the behaviour of the network and services and if the failure rate exceeds certain threshold, make the calls to the remote destinations to fail immediately. After certain timeout, the circuit breaker sends some testing call to the quarantined server to check whether it has recovered. If the calls succeed then the circuit breaker stops failing the calls to this sever, i.e., it "closes" the circuit.

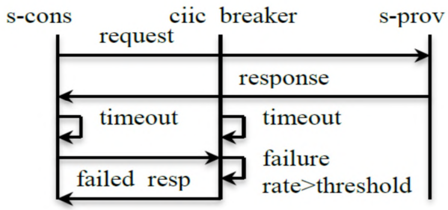


Figure 6. Interactions in the circuit breaker pattern

The graphical model for representing the interactions in the circuit breaker pattern is shown in Figure 6. The state diagram representing the dynamic behaviour is given in Figure 7. The circuit breaker is activated upon receiving a request from a service consumer. It monitors the request execution by the service provider and collects the corresponding data. If request fails, it checks whether the failure rate has exceeded the predefined threshold. If not then the circuit breaker continues to function in the monitoring mode, i.e., does not block the calls to the corresponding service provider.

However, if the failure rate threshold is exceeded, then the circuit breaker changes its mode to block, i.e., fail all the calls to the corresponding service provider. The service provider becomes quarantined. After the quarantine time expires, the circuit breaker sends a test request to the quarantined service provider. If the request is successfully returned the quarantine is removed and the circuit breaker returns to the monitoring mode.

Circuit breakers provide us with an efficient way to prevent cascading failures. They rely on timeout pattern to detect failures and correspondingly, recoveries of the service providers. The circuit breakers collect data about the behaviour of the microservices, which can be used to refactor them and continuously improve the reliability of the microservices architectures.

IV. Composite Fault Tolerance Patterns

In this section, we overview more complex fault tolerance patterns. They built on the patterns introduced in Section III as well as classic fault tolerance techniques.

Proactive fault tolerance. This pattern aims at preventing an execution of complex scenarios that cannot be executed to the completion. Proactive fault tolerance identifies potential failures before the scenario or part of it is executed, signals about the possible deadlock and either proposes an alternative way to execute a scenario or fails it.

The proactive fault tolerance pattern is a composite pattern that relies on timeout and circuit breaker patterns as well as other standard fault tolerance techniques.

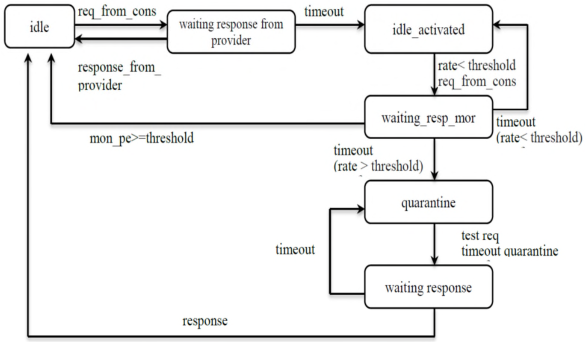


Figure 7. Dynamic behaviour of circuit breaker pattern

The graphical model of proactive fault tolerance pattern is shown in Figure 8.

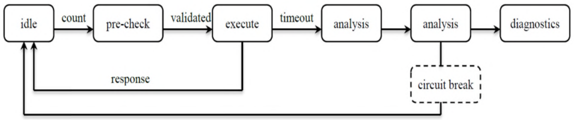


Figure 8. Dynamic behaviour of proactive fault tolerance pattern

Composite services. Some microservices are composite, i.e., to provide a requested service they need to request services from several other microservices, process the responses and finally provide the requested result. Let us note, that any other microservice might also be "composed" of several microservices, i.e., in its turn, the requested microservice execution might be orchestrated by its (sub)service director. Hence, in general, a composite microservice might have several layers of hierarchy.

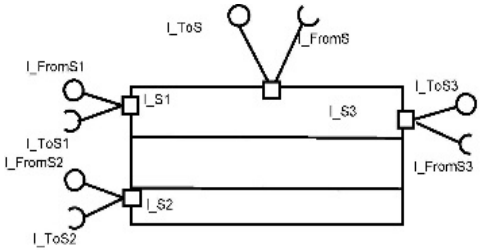


Figure 9. Service director: static view

To model a composite microservice, we introduce the providers of the microservices into the abstract architectural service model. The model includes the external service providers communicating with the microservice director via their service director, as shown in Figure 9.

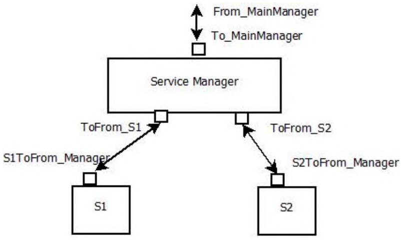


Figure 10. Duplication pattern: static view

Now, let us discuss the patterns that allow us to introduce structural means for fault tolerance using various forms of redundancy.

Duplication pattern. The duplication is a simplest arrangement for structural fault tolerance. It can be introduced if there are two microservice providers, which provide functionally identical microservices. In this case, the request from a service director of a service consumer can be duplicated and microservice providers activated in parallel. An execution is successful if any out of two microservice providers successfully completes the request.

An architectural diagram of the duplication arrangement is given in Figure 10 and the dynamic behavior shown in Figure 11.

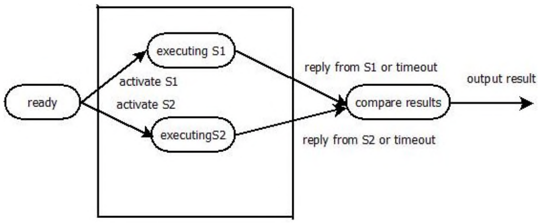


Figure 11. Dynamic behavior of duplication pattern.

Triple modular redundancy pattern. A more complicated scheme for structural redundancy – triple modular redundancy (TMR) is shown in Figure 12. The precondition for implementing it is that we have three microservice providers that provide functionally identical microservices. In this case, a service request from a service consumer or service director should be triplicated. All three microservice providers receive the same service request and work in parallel. The results of the service execution are sent to a voting element.

The voting element is a dedicated microservice that performs comparison of the results and produces the final result. The voting element takes a majority view over the produced results of the successfully executed services and outputs it as the final result of the service execution.

The voting microservice might be implemented in two different ways: it might output the results after receiving the first two replies or it might start to act only after the certain deadline when all non-failed services have replied

The proposed patterns offer suitable solution for achieving fault tolerance in developing applications in the microservice architectural style.

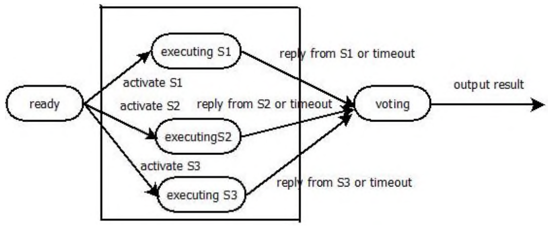


Figure 12. Dynamic behavior of TMR pattern.

V. RELATED WORK AND CONCLUSIONS

While the topic of service orchestration and composition has received significant research attention, the fault tolerance aspect is not so well addressed. Liang [8] proposes a fault-tolerant Web service on SOAP (called FT-SOAP) using the service approach. It extends the standard WSDL by proposing a new element to describe the replicated Web services. The client side SOAP engine searches for the next available backup from the group WSDL and redirects the request to the replica if the primary server failed. It is a rather complex mechanism that hinders interoperability.

Artix [2] is IONA's Web services integration product. It provides a WSDL-based naming service by Artix Locator. Multiple instances of the same service can be registered under the same name with an Artix Locator. When service consumers request a service, the Artix Locator selects the service instance based on a load-balancing algorithm from the pool of service instances. It provides useable services for the service consumers. An active UDDI mechanism [4] enables an extension of UDDI’s invocation API to enable fault-tolerant and dynamic service invocation. Its function is similar to the Artix Locator. A dependable Web services framework is proposed in [1]. Once a failure for one specific service occurs, the proxy raises a “WebServiceNotFound” exception and downloads its handler from DeW. The exception handling chooses another location that hosts the same service and re-invoks the method automatically. The main goal of DeW is to realize physical-location-independence. Providing fault-tolerance capability for composite Web service has also been discussed in [3].

A formal approach to introducing fault tolerance to the service architecture in the telecommunication domain has been proposed in [6][7][13][14]. This work extends the set of architectural patterns that can be introduced to achieve fault tolerance as well as propose a systematic support for deriving fault tolerance solutions.

The fault tolerance means are often assessed quantitatively. The techniques for probabilistic assessment of fault tolerance have been proposed in [15]-[18]. These techniques can be applied together with the proposed fault tolerance patterns.

In this paper, we have proposed a systematic model-driven approach to achieving fault tolerance in microservice architectures. We have defined generic modelling patterns, which can be utilised in model-driven engineering in the microservice architectures. Our patterns help to analyse possible failures and propose efficient solutions to cope with them. By integrating the proposed patterns into the architecture of a microservice, we can improve QoS and achieve higher reliability. Our patterns propose both dynamic and static means for achieving fault tolerance. The dynamic patterns rely on run-time monitoring behaviour and activating patterns if certain failure detection conditions occur. The static pattern help the designers to systematically utilise the redundancies present in the provisioning of microservices.

We believe that our approach supports structured guided reasoning about fault tolerance and enables efficient exploration of the design space while developing complex microservice architectures.

REFERENCES

[1] E. Alwagait, S. and Ghandeharizadeh, “A Dependable Web Services Framework” 14th International Workshop on Research Issues on Data Engineering 2004, http://fac.ksu.edu.sa/alwagait/publication/31143 retrieved January 2018.

[2] Artix Technical Brief. http://www.iona.com/artix, retrieved January 2018.

[3] V. Dialani, S. Miles, L.Moreau, D. Roure, and M. Dialani, “Transparent fault tolerance for Web services based architectures”. 8th Europar Conference (EULRO-PAR02), Springer 2002, pp. 889-898. ISBN: 3-540-44049-6

[4] M. Jeckle and B. Zengler, “Active UDDI-An Extension to UDDI for Dynamic and Fault Tolerant Service Invocation” 2nd International Workshop on Web and Databases, Springer 2002, pp. 91-99. ISBN:3-540-00745-8.

[5] L. Laibinis, E. Troubitsyna, and S. Leppänen, “Service-Oriented Development of Fault Tolerant Communicating Systems: Refinement Approach” International Journal on Embedded and Real-Time Communication Systems, vol. 1, pp. 61-85, Oct. 2010, DOI: 10.4018/jertcs.2010040104.

[6] L. Laibinis, E. Troubitsyna, S. Leppänen, J. Lilius, and Q. Malik, “Formal Service-Oriented Development of Fault Tolerant Communicating Systems”, in M. Butler, C. Jones, A. Romanovsky, and E. Troubitsyna (Eds.), Rigorous Development of Complex Fault-Tolerant Systems, LNCS 4157, pp. 261-287, Springer 2006, ISBN 978-3-642-00867-2.

[7] J. C. Laprie. Dependability: Basic Concepts and Terminology. Springer-Verlag, 1991.

[8] D. Liang, C. L. Fang, C. Chen, F. X, Lin. ”Fault-tolerant Web service”. Tenth Asia-Pacific Software Engineering Conference, IEEE Press, Dec. 2003, pp.56-61, ISBN 973-4-642-01867-1

[9] J. Rumbaugh, I. Jakobson, and G .Booch, The Unified Modelling Language Reference Manual. Addison-Wesley, 1998.

[10] Web Services Architecture Requirements http://www.w3.org/TR/wsareqs, retrieved January.2018.

[11] M. Fowler and J. Lewis. Microservices: a definition of this new architectural term. In [Online]. Available https://martinfowler.com/articles/microservices.ml. Accessed: 01-April- 2019.

[12] T. Erl. Serivce-Oriented Architecture (SOA): Concepts, Technology, and Design. Prentice Hall, 2005. ISBN: 978-0131858589.

[13] A. Tarasyuk, E. Troubitsyna, L. Laibinis. Formal Modelling and Verification of Service-Oriented Systems in Probabilistic Event-B. In Proc. of *IFM 2012*, LNCS 7321, pp.237–252, Springer, 2012.

[14] A. Tarasyuk, I. Pereverzeva, E. Troubitsyna, L. Laibinis, Formal Development and Quantitative Assessment of a Resilient Multi-robotic System. In: Proc of *SERENE 2013*, LNCS 8166, pp. 109-124, Springer.

[15] E. Troubitsyna, “Reliability assessment through probabilistic refinement,” Nordic J. of Computing 6 (3), 320-342, 1999.

[16] L. Laibinis, B. Byholm, I. Pereverzeva, E. Troubitsyna, K.E. Tan and I. Porres, Integrating Event-B Modelling and Discrete Event Simulation to Analyse Resilience of Data Stores in the Cloud.The 11th International Conference on Integrated Formal Methods, iFM 2014, LNCS 8739, pp. 103-119, Springer 2014.

[17] I.Pereverzeva, L. Laibinis, E. Troubitsyna, M. Holmberg, M. Pöri, Formal Modelling of Resilient Data Storage in Cloud. In: Lindsay Groves, Jing Sun (Eds.), *Proceedings of 15th International Conference on Formal Engineering Methods*, LNCS 8144, 364–380, Springer-Verlag Berlin Heidelberg, 2013.

[18] A. Tarasyuk, I. Pereverzeva, E. Troubitsyna, L. Laibinis, Formal Development and Quantitative Assessment of a Resilient Multi-robotic System. In: A. Gorbenko, A. Romanovsky, V. Kharchenko (Eds.), Proceedings of the 4th International Workshop on Software Engineering for Resilient Systems (SERENE 2013), Lecture Notes in Computer Science 8166, 109–124, Springer-Verlag Berlin Heidelberg, 2013.

Facial Recognition and Emotion Detection System for Dynamic Advertisement Allocation

Frank Yeong-Sung Lin[1], Evana Szu-Han Fang[1], Chiu-Han Hsiao[2]
Department of Information Management, National Taiwan University[1]
Research Center for Information Technology Innovation, Academia Sinica[2]
Taipei, Taiwan
email: yeongsunglin@gmail.com, evanafang@gmail.com, chiuhanhsiao@citi.sinica.edu.tw

***Abstract*—Advertisements represent a persuasive method of communication for convincing people to change their thoughts or attitudes. Conventional advertisements do not always provide optimal marketing effectiveness because the advertisements are presented uniformly to viewers. To overcome the limitations of traditional methods in advertising research, a dynamic advertisement model is proposed in this paper, and facial expression detection is applied to real-time measurement during media exposure. This is a novel model to recognize viewers' facial expression for emotion regulation and then adjust the decision of content sequence according to their emotions. A decision tree algorithm is used, and each demographic measurement results from a few scenarios. The decision is determined through bottom-up branch searching. Based on the study results, personalized advertising and audience targeting with accurate facial expression analysis can allow marketing and advertising researchers to better understand viewers' emotional valence and behavior and to employ mathematical formulation for establishing the optimal advertising approach.**

Keywords-dynamic advertisement; facial recognition; emotion detection; audience targeting; decision tree.

I. INTRODUCTION

Advertising is a persuasive method of communication for convincing people to change their thoughts or attitudes [1]. This is a type of brand-related stimulus that conveys brand experiences, consisting of subjective and internal customer responses [2]. With time, enduring memories of brand experiences in customers' minds affect customer satisfaction and loyalty [3][4]. Therefore, advertising and marketing companies actively seek an optimal instrument that can recognize true feelings from customers, and they cannot hide their thoughts [5]. Enhancing understanding, evaluation, and advertising effectiveness is of great value both in theory and in practice.

People experience certain emotional responses when viewing an advertisement, which may be positive or negative. This greatly affects the sales of a product, reduces price sensitivity, and creates brand value [6]. Hence, viewers' emotions can be used to predict an advertisement's effectiveness [7]. As a strong connection exists between emotions and facial expression, researchers have been interested in developing methodologies to effectively measure the facial expression and emotions experienced [5][8]-[12]. The facial expression is the clearest method of establishing a person's affective state [13]. Research has indicated that variables correlated to advertising success such as advertisement likability [14], recall [15], and "zapping" [12] can be predicted by facial expressions.

Exposure to an advertising stimulus evokes emotions among people; their attention is subsequently affected, and zapping is also affected by emotion and attention; the extent of these effects varies during exposure to an advertisement [12]. Emotion regulation is a dynamic process. Thus, traditional uniform content advertised to audiences cannot achieve optimal marketing effectiveness, because of different preferences of individuals and their emotions. Brands may squander the opportunity to communicate when targeted customers zap, skip, and zip advertisements [12]. To retain viewers and maximize marketing effectiveness, a novel real-time content adjustment model based on emotion detection is proposed.

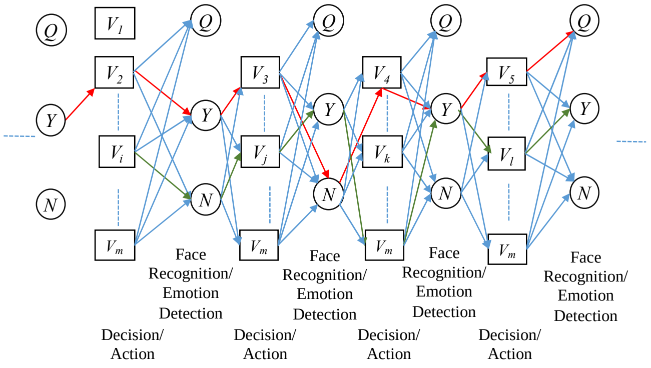


Figure 1. Decision tree for facial recognition and emotion detection.

Decision trees constitute a nonparametric supervised learning method used for classification and regression [16]. They predict the value of a target variable by learning simple decision rules inferred from data features. In this paper, the decision tree learns from data related to facial recognition and emotion detection, as seen in Figure 1, to approximate a set of if-then-else decision rules. If a given situation is observable in a model, the condition is easily explained by Boolean logic. Understanding video clip interpretations is simple, and the selected decision rules are determined completely if viewing promotes positive emotions. In practice, a preferred classification model can be verified through statistical tests and adjusted within a set time interval for real-world applications.

The remainder of this paper is structured as follows. Section II reviews relevant studies on the linkage between facial expression detection, emotion detection, and advertisement content. The bottom-up decision-making method is proposed in Section III. Section IV details the process flow and cases of identifying decisions and rules in practical applications. Finally, conclusions are drawn in Section VI.

II. RELATED WORK

Traditional research methods, such as self-report provide limited understanding of the linkage between emotion and advertisement content as well as of how advertisement effectiveness is measured [17]. Though self-report is cheap, fast, and valid, it cannot capture low-order emotion or the temporal nature of emotion regulation during an advertisement; it may also increase cognitive bias [5][16]. People communicate valence and emotional states powerfully through their faces [18]. Therefore, marketing and advertising researchers can better understand viewers' emotions and behaviors through facial expression analysis and can accordingly establish strategies to improve advertisement effectiveness. Moreover, they may design interactive advertisements to improve viewer experiences [19]. Research has used automatically measured facial expressions to predict emotional valence and advertisement preference during media exposure [14][20][21]. Thus, the use of automated tools augments the feasibility of the approach and exhibits higher predictive capability than self-report does [5][7]. Many studies have initially investigated feigned or acted facial behaviors [16]. Nevertheless, research has progressively emphasized naturalistic and spontaneous behavior [22]-[24] and subtle expressions [25].

Automated facial expression detection combines the fields of psychology, computer vision, and machine learning [18]. In [26], the authors proposed a new nonlinear tensor factorization based on deep variational autoencoders called Factorized Variational Auto-Encoders (FVAE) for modeling movie audiences' facial expressions. The effectiveness of FVAE was determined for a large facial expression dataset extracted from 3179 movie audience members. Even when using only 5% of data for initial observation, FVAE could reconstruct facial reactions more precisely using data from movie audiences than traditional baseline applying entire data. One study [19] focused on predicting user behavior and viewing experiences based on facial expressions during online advertisement viewing. A metric termed Moment-to-Moment Zapping Probability (MMZP) was used to predict user skipping; the preference information extracted from users may be used to enhance advertisement effectiveness. In addition, the authors categorized smiling as a primary facial expression during analysis. Because amusement is a desirable response that advertisers strive to elicit, the entertainment level of an advertisement directly relates to smiling. Sparse reconstruction coefficients were used as features for classifying smiling to make MMZP predictions. The authors in [18] collected spontaneous facial expressions from viewers during the 2012 US presidential debates to predict voter preferences and found an average precision of over 73%. The Facial Action Coding System [21][27][28] was implemented to measure and score facial activity reliably and to distinguish subtle differences in facial expressions [29]. In [30], online video advertisements viewed by Japanese people were analyzed for physiological responses involving facial expressions, heart rate signals, and gaze. The authors integrated each mode's features and evaluated advertisement likability and purchase intent. In [31], the authors proposed an interactive advertisement system with 3D tracking and facial recognition to produce audience profile surveys. They reviewed the conclusions of Lord and Burnkrant [32] regarding the increased attention levels involved when viewers were immersed in highly interactive programs. The psychological sensation of presence could explain this cognitive state. Moreover, [33] indicated that this experience of presence would affect product knowledge, brand attitude, and the purchasing intentions of consumers.

III. PROPOSED METHOD AND PROCESS FLOW

The model depicted in Figure 2 is proposed to maximize viewer experiences during advertisements. A decision tree algorithm of content personalization is employed to implement decisions into fields to achieve objectives. Reward measurement and observation uses facial recognition and emotion detection techniques.

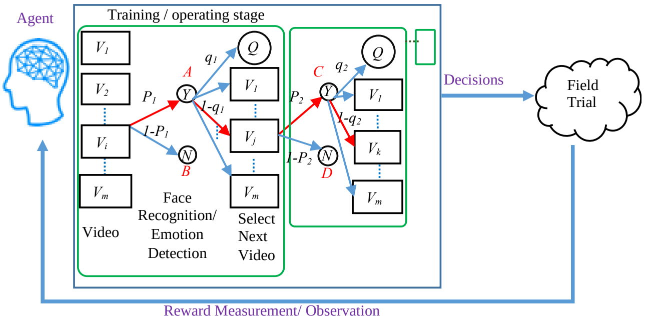


Figure 2. Abstract system architecture.

In Figure 3, the process flow of an advertisement involves several video clips. The decision tree algorithm can be applied to a dynamic scenario mechanism for advertisements based on recognition of viewers' facial expressions. Training and testing involves two stages: probability and value evaluation are determined by facial recognition and emotion detection processes. The bottom-up algorithm is then applied to select video clips forming a complete advertisement tree structure with advertisement video clips. In Figure 3, the constructed path selected in a tree structure represents the sequence of the advertisement video clips for a type of demographic classification. To optimize the effects of the advertisements, a clip is selected on the basis of the viewer's emotion detection results after the end of the preceding clip. The algorithm finally classifies viewers according to the probability and value evaluation determined by facial recognition and emotion detection in the bottom-up backward process.

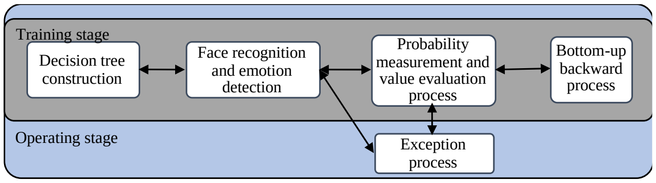

Training stage
Decision tree construction
Face recognition and emotion detection
Probability measurement and value evaluation process
Bottom-up backward process
Operating stage
Exception process

Figure 3. Process flow of proposed stages.

A. *Decision Tree Construction*

The distributions of viewers' demographic characteristics (e.g., age, sex, etc.) are applied to the construction of the optimal trees. Once a viewer is categorized through facial recognition analysis, the tree path of that category is extracted for implementation in the model. Probability and reward values are acquired through measurement.

B. *Facial Recognition and Emotion Detection for Probability Measurement and Value Evaluation*

The parameters P_i represent the probabilities of the next selected video clips. Before finishing the construct of the optimal trees, we have a training and experiment phases with a number of viewers to measure each probability shown in Figure 5. The experiment is implemented and all the viewers see clip V_1 firstly, then some of them may have positive emotion (the node *AY*), the others may have negative emotion (the node *BN*). P_1 and P_2 are measured in this step. After that, assuming the experiment goes to *AY*, half of *AY*'s viewers are aired V_2; half of them are aired V_3 at random. The measured P_3, P_4, P_5 and P_6 correspond to *CY*, *DN*, *EY* and *FN*. We implement the procedures given above to measure the probabilities if the experiment goes to other branches.

A, B, C, D, E,…, which represent marginal effects with facial recognition to expand all possibilities before the optimal decision tree is not coming out yet in the training stage. The model records every viewing experience and establishes the best advertisement editing and composition strategy for each demographic group of viewers in the operating stage. A 5- point or 7-point Likert scale can be used to measure viewers' perceptions and purchase intentions after they view an advertisement. The incentive is provided in accordance with the scale.

C. *Bottom-up Decision-Making*

Binning or discretization is the process of transforming numerical variables into categorical counterparts [34]. Numerical variables are usually discretized in modeling methods based on decision trees, such as *A*, *B*, *C*, *D*, *E*, … representing rewards for positive or negative emotions detected in Figure 5.

In bottom-up decision-making, the reverse approach is applied to top-down decision-making. To ensure that bottom-up decision-making is effective, emotion detection information is used in the predictive model, and outcomes are accordingly predicted. Descriptive modeling is the assignment of observations into decision trees. The rules employed permit associations among observations, and they are based on the entropy using the frequency table of two attributes, which are the expected rewards of emotion detection. The equation used is $E(K,N)=\sum_{k\in K}\sum_{n\in N}P(V_k)E(n)$. $E(n)$ is the emotion detection result. $P(V_k)$ is the probability of selecting the video V_k. For example, the value of (P_3C+P_4D) is the expected reward related to the decision V_2. The value of (P_5E+P_6F) is related to the decision V_3. Entropy and decisions are constructed bottom-up to form the decision tree depicted in Figure 6. The pseudocode is presented in Figure 4, as follows:

```
Training data input in experiment
Generation of Tree (Decisions K, Emotion Detection N)
If stopping_condition(K,N) = true then
leaf = createNode()
leaf.label= Classify(K)
return leaf
root = createNode()
root.test_condition = findBestSplit(K,N)
$E(K,N)=\sum_{k\in K}\sum_{n\in N}P(V_k)E(n)$ : list possible outcome of
root.test_condition
for each value E for branches
Select the maximum E related to decisions K;
Build child = TreeGrowth(K, N) ;
Add child as a descent of root and label the edge
return root
```

Figure 4. Pseudocode of the decision tree algorithm.

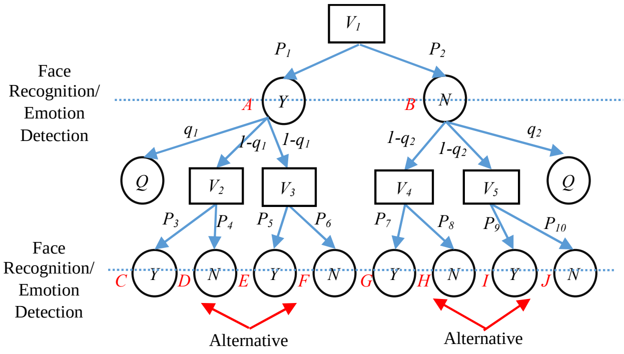


Figure 5. Bottom-up alternative selection.

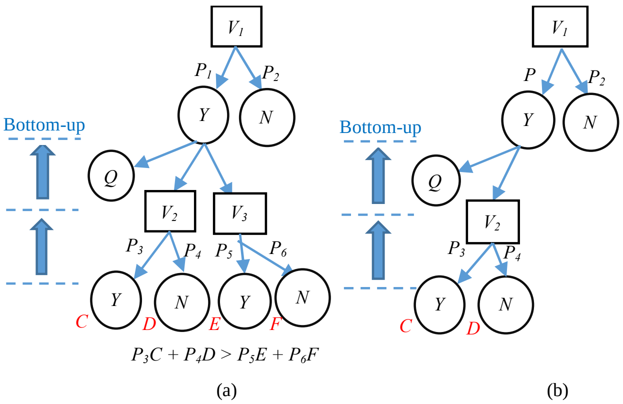


Figure 6. Bottom-up mechanism for the expected value calculation.

To apply the bottom-up mechanism for calculation, the proper video clip is assigned and inserted to continue the sequence of clips by detecting emotions at any level from facial recognition results. For example, after watching V_1, facial expression recognition indicates that the viewer experiences a positive emotion, so AY is subsequently selected, and the model may choose V_2 or V_3 to continue the advertisement: $P_3C + P_4D$ for V_2 or $P_5E + P_6F$ for V_3, depending on which one has the higher expected value for effect. If $P_3C + P_4D$ for V_2 is more favorable, then the branch $P_5E + P_6F$ is deleted.

D. Exception Process

In the case of a negative emotion for V_1, BN is subsequently selected, as depicted in Figure 7, and the aforementioned method can be applied to select V_4 or V_5. If more video clips are available, the bottom-up method can be used to select from the bottom to the top of the model.

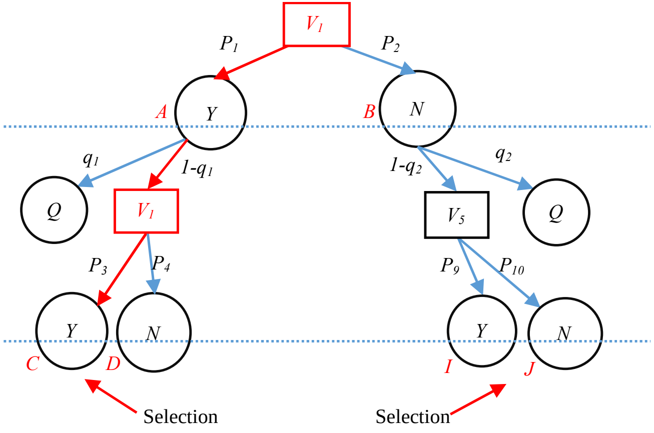


Figure 7. Branch selection for decision-making.

After running the model for a period, as in Figure 8, if the on-record $P_3C + P_4D$ for V_2's expected value is lower than the value of $P_5E + P_6F$, V_2 is switched to V_3. The top branch of V_3 is adjusted accordingly. Continuing to measure the bottom expected value under V_3, if it is lower than the bottom expected value under V_2, then the selection is again switched from V_3 to V_2, and the top branch of V_2 consistently.

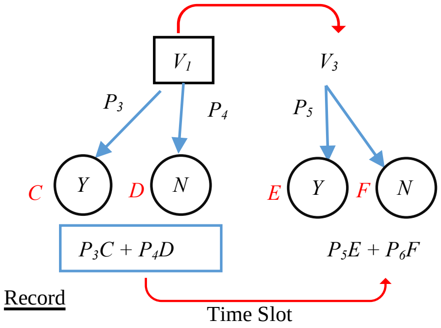


Figure 8. Recording emotion detection values in a period.

A scenario is dynamically created according to the viewers. The structure for selecting proper video clips to suit the advertisement scenario is as follows. As seen in Figure 9, the viewer watches the first clip (V_1), and when it ends, the next clip is selected based on the viewer's emotion detection results.

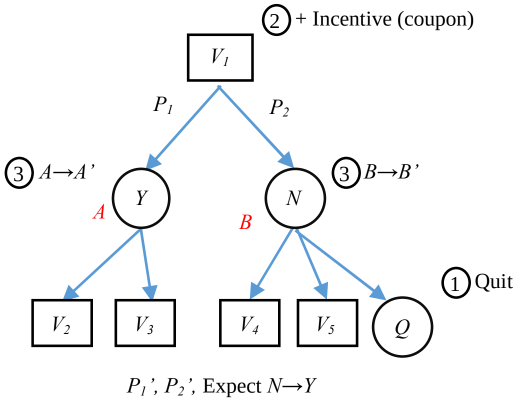


Figure 9. Incentivizing viewers to continue watching videos.

This model enables the timely identification of the user's emotion from the last clip. For example, in Figure 9, if the viewer wants to stop watching additional clips, an incentive can be provided to entice the viewer to continue watching.

IV. Conclusion & Future Work

Although Information Communication Technology is developing rapidly, it is applied in management relatively infrequently, and it has yet to be used to fully establish independent technology. Traditional methods, such as surveys, provide limited understanding of the linkage between emotion and advertisement content as well as of how to measure advertisement effectiveness. Therefore, this paper proposes a dynamic scenario mechanism for advertising based on a decision tree algorithm and on viewers' facial expressions recognized during viewing. Dynamic content changes result from viewer facial expression recognition. The content customization guides the viewer's concentration [35]. Subsequently, this work explores the research topics involved in technology and management issues, and to obtain the results for applying theory to practice, as a suitable reference for the video clip strategies used to the operator or related industry company in optimal advertisement display and well predictive analysis in the future. The future directions are summarized as follows:

- Level of emotion detection:

Emotional responses can be defined and distributed into several categories, such as positive versus negative emotions. They may even be classified according to three types of emotions: happiness or approval, neutral, and disapproval. However, happiness may be further subdivided into extreme delight, surprised excitement, or tears of joy. Due to the complexity of emotions, it can be posited that emotions can be distributed into two types: positive and negative.

- Multiple viewers:

When the camera detects more than one viewer (e.g., three people comprising two male and one female), these people can be distributed into two categories by demographic characteristic. The decisions are determined by the two trees in accordance with these two categories. Then, both trees are extracted, overlapped, and superposition or weighted sum of the probabilities for each branch to form a new tree.

- Level of emotion detection for multiple viewers:

When only one optimal tree exists for a group of viewers, people may leave or start watching halfway through an advertisement; therefore, future research can uncover solutions regarding how this optimal tree can be altered dynamically. Such a study will entail complications, but is definitely worthwhile for developing more customized advertisements for optimizing the viewing experience

References

[1] J. Meyers-Levy and P. Malaviya, “Consumers’ processing of persuasive advertisements: An integrative framework of persuasion theories,” Journal of Marketing, vol. 63, no. 4, Oct. 1999, pp. 45–60.

[2] J. Brakus, B. H. Schmitt, and L. Zarantonello, “Brand experience: What is it? How is it measured? Does it affect loyalty?” Journal of marketing, vol. 73, no. 3, May, 2009, pp. 52–68.

[3] R. L. Oliver, Satisfaction: A Behavioral Perspective on the Consumer, Boston, NY: McGraw-Hill, 1997.

[4] F. F. Reichheld, The Loyalty Effect: The Hidden Force Behind Growth, Profits, and Lasting Value. Boston, NY: Harvard Business School Press, 1996.

[5] P. Lewinski, M. L. Fransen, and E. S. H. Tan, “Predicting advertising effectiveness by facial expressions in response to amusing persuasive stimuli,” Journal of Neuroscience Psychology and Economics, vol. 7, no. 1, Mar. 2014, pp. 1–14.

[6] P. Saraswat, H. Nagar, and S. Khandelwal, “Make it feel: Use of facial imaging technique to analyze the impact of each emotional spot on ad success,” Proc. The 9th International Conference on Complex, Intelligent, and Software Intensive Systems (CISIS 2015), IEEE, Jul. 2015, pp. 502–507.

[7] K. Poels and S. Dewitte, “How to capture the heart? Reviewing 20 years of emotion measurement in advertising,” Journal of Advertising Research, vol. 46, no. 1, Nov. 2006, pp. 18– 37.

[8] P. M. Cole, “Children’s spontaneous control of facial expression,” Child Development, vol. 57, no. 6, Dec. 1986, pp. 1309–1321.

[9] J. J. Gross and R. A. Thompson, “Emotion regulation: Conceptual foundations,” Handbook of Emotion Regulation, pp. 3–26. New York, NY: Guilford Press, 2007.

[10] C. E. Izard, “Facial expressions and the regulation of emotions,” Journal of Personality and Social Psychology, vol. 58, no. 3, Mar. 1990, pp. 487–498.

[11] Darwin and Charles, The Expression of the Emotions in Man and Animals. London: Murray, 1872.

[12] T. Teixeira, M. Wedel, and R. Pieters, “Emotion-induced engagement in Internet video advertisements,” Journal of Marketing Research, vol. 49, no. 2, Apr. 2012, pp. 144–159.

[13] H. P. Mal and P. Swarnalatha, “Facial expression detection using facial expression model,” Proc. International Conference on Energy, Communication, Data Analytics and Soft Computing (ICECDS 2017), IEEE, Aug. 2017, pp. 1259–1262.

[14] D. McDuff, R. Kaliouby, T. Senechal, D. Demirdjian, and R. W. Picard, “Automatic measurement of ad preferences from facial responses gathered over the internet,” Image and Vision Computing, vol. 32, no. 10, Oct. 2014, pp. 630–640.

[15] R. L. Hazlett and S. Y. Hazlett, “Emotional response to television commercials: Facial EMG vs. Self-report,” Journal of Advertising Research, vol. 39, no. 2, Mar. 1999, pp. 7–24.

[16] A. Suarez and J. F. Lutsko, “Globally optimal fuzzy decision trees for classification and regression,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 21, no. 12, Dec. 1999, pp. 1297–1311.

[17] D. McDuff, R. Kaliouby, J. F. Cohn, and R. W. Picard, “Predicting ad liking and purchase intent: Large-scale analysis of facial responses to ads,” IEEE Transactions on Affective Computing, vol. 6, no. 3. Jul.-Sep. 2015, pp. 223–235.

[18] D. McDuff, R. Kaliouby, E. Kodra, and R. W. Picard, “Measuring voter's candidate preference based on affective responses to election debates,” Proc. Humaine Association Conference on Affective Computing and Intelligent Interaction (ACII 2013), IEEE, Sep. 2013, pp. 369–374.

[19] S. Yang and L. An, “Analyzing user behavior in online advertising with facial expressions,” Proc. 23rd International Conference on Pattern Recognition (ICPR 2016), IEEE, Dec. 2016, pp. 4238–4243.

[20] D. McDuff, R. Kaliouby, K. Kassam, and R. W. Picard, “Affect valence inference from facial action unit spectrograms,” Proc. Computer Society Conference on Computer Vision and Pattern Recognition – Workshops (CVPRW 2010), IEEE, Jun. 2010, pp. 17–24.

[21] E. Kodra, T. Senechal, D. McDuff, and R. el Kaliouby, “From dials to facial coding: Automated detection of spontaneous facial expressions for media research,” Proc. 10th International Conference and Workshops on Automatic Face and Gesture Recognition (FG 2013), IEEE, Apr. 2013, pp. 1–6.

[22] D. McDuff, R. Kaliouby, and R. W. Picard, “Crowdsourcing facial responses to online videos,” IEEE Transactions on Affective Computing, vol. 3, no. 4, Oct.-Dec. 2012, pp. 456–468.

[23] M. Pantic, “Machine analysis of facial behaviour: Naturalistic and dynamic behaviour,” Philosophical Transactions of the Royal Society of London B: Biological Sciences, vol. 364, no. 1535, Dec. 2009, pp. 3505–3513.

[24] J. Whitehill, G. Littlewort, I. Fasel, M. Bartlett, and J. Movellan, “Toward practical smile detection,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 31, no. 11, Nov. 2009, pp. 2106–2111.

[25] T. Senechal, J. Turcot, and R. Kaliouby, “Smile or smirk automatic detection of spontaneous asymmetric smiles to understand viewer experience,” Proc. 10th International Conference and Workshops on Automatic Face and Gesture Recognition (FG 2013), IEEE, Apr. 2013, pp. 1–8.

[26] Z. Deng et al. “Factorized variational autoencoders for modeling audience reactions to movies,” Proc. Conference on Computer Vision and Pattern Recognition (CVPR 2017), IEEE, Jul. 2017, pp. 6014–6023.

[27] P. Ekman and W. V. Friesen, Facial Action Coding System: A Technique for the Measurement of Facial Movement. Palo Alto, CA: Consulting Psychologists Press, 1978.

[28] J. F. Cohn, Z. Ambadar, and P. Ekman, “Observer-based measurement of facial expression with the Facial Action Coding System,” The Handbook of Emotion Elicitation and Assessment. Oxford University Press Series in Affective Science, pp. 203–221. New York, NY: Oxford University Press, 2007.

[29] C. Hjortsjö, Man’s face and mimic language. Lund, Sweden : Studentlitteratur, 1969.

[30] G. Okada, K. Masui, and N. Tsumura, “Advertisement effectiveness estimation based on crowdsourced multimodal affective responses,” Proc. Conference on Computer Vision and Pattern Recognition Workshops (CVPRW 2018), IEEE/CVF, Jun. 2018, pp. 1344–1348.

[31] M. Taspinar, A. T. Naskali, G. Eren, and M. Kurt, “The importance of customized advertisement delivery using 3D tracking and facial recognition,” Proc. 2nd International Conference on Digital Information and Communication Technology and it's Applications (DICTAP 2012), IEEE, May, pp. 520–524.

[32] K. Lord and R. Burnkrant, “Attention versus distraction: the interactive effect of program involvement and attentional devices on commercial processing,” Journal of Advertising, vol. 22, no. 1, Mar. 1993, pp. 47–60.

[33] H. Li, T. Daugherty, and F. Biocca, “Impact of 3-D advertising on product knowledge, brand attitude, and purchase intention: The mediating role of presence,” Journal of Advertising, vol. 31, no. 3, May, 2002, pp. 43–57.

[34] D. Bacciu, A. Micheli, and A. Sperduti, “Compositional generative mapping for tree-structured data—Part I: Bottom-up probabilistic modeling of trees,” IEEE Transactions on Neural Networks and Learning Systems, vol. 23, no. 12, Dec. 2012, pp. 1987–2002.

[35] K. Lee, S. Rho, and E. Hwang, “Scenario based dynamic content management system for e-Learning environment,” Proc, 30th International Conference on Advanced Information Networking and Applications Workshops (WAINA 2016), IEEE, Mar. 2016, pp. 183–186.

Applying a Shared Decision Making Platform to Improve I-131 Patients Medical Care Quality

Jui-Jen Chen
Department of Nuclear Medicine, Chang Gung Memorial Hospital, Kaohsiung Medical Center, Chang Gung University College of Medicine, Taiwan (R.O.C.)
Email:china111@ms57.hinet.net

Yu-Ting Lai
Department of Nuclear Medicine, Chang Gung Memorial Hospital, Kaohsiung Medical Center, Chang Gung University College of Medicine, Taiwan (R.O.C.)
Email:yutinglai@cgmh.org.tw

***Abstract*—This paper provides post-operative thyroid cancer patients with assessment and health education about the radioactive iodine (I-131) treatment and proposes a platform where information technologies are applied to facilitate effective physician-patient communication to ensure optimal treatment for patients. A questionnaire survey was conducted in the Department of Nuclear Medicine of Chang Gung Memorial Hospital, Taiwan, from September 2017 to May 2018. Before implementation of Shared Decision Making (SDM), the 1 to 7 questions mean score was 1.3; after the implementation of SDM, the 1 to 7 questions mean score increased to 4.89. We conclude that applying SDM establishes a standardized and effective communication process for physicians and patients, and offers each patient a personalized and most appropriate treatment method.**

Keywords—Shared Decision Making (SDM); patient safety; medical quality; nuclear medicine.

I. Introduction

Patient safety is the foundation of medical quality and also the common goal of care providers and patients. The Ministry of Health and Welfare's Taiwan Patient Safety Goals for Hospitals for 2018-2019 outlines eight goals. These goals are meant to build consensuses on patient safety among medical institutions and serve as a guideline for Taiwan's medical institutions and personnel. The underlying purpose of these goals is to drive a simultaneous improvement of all institutions rather than to audit them. The eight goals are as follows: (1) Improve communication among health care personnel; (2) Reinforce management of patient safety events; (3) Improve surgical safety; (4) Prevent falls and reduce the degree of injuries in patients; (5) Improve medication safety; (6) Ensure effective control of infections; (7) Improve tubing safety; (8) Encourage patients and patient families to participate in patient safety tasks. As to the goal of encouraging patients and patient families to participate in patient safety tasks, the strategy has been modified from its 2017 predecessor as follows: provide multiple channels for participation in patient safety practices, increase awareness of patient safety, implement Shared Decision Making (SDM), and improve the care knowledge among primary caregivers in hospitalization and post-hospitalization settings [1][2]. The 2016 annual report of Taiwan Patient-safety Reporting System (TPR) provides an analysis of communication-related causes of safety events among 10,749 cases that occurred in 2016 [3]. The analysis shows that 41 out of 100 cases were related to "poor communication between the care team and patients or patient families", making this factor the primary cause. The second cause was "poor communication within the care team", accounting for 29.8 cases per 100 cases. As shown in Figure 1, these statistics highlight the importance of communication.

The term "shared decision making" was first mentioned in a U.S. patient-centered care program in 1982 to promote mutual respect and communication between physicians and patients[4]. In 1997, Cathy Charles proposed an operational definition of SDM, suggesting that SDM is where at least two participants, namely physician and patient, are involved; the physician provides evidence of various treatment options available to the patient; both parties share and discuss information; an agreement is reached on the treatment to implement [5]-[7]. SDM is a patient-centered clinical process characterized by three elements, namely knowledge, communication, and respect. As shown in Figure 2, the objective of SDM is to allow clinicians and the patient to share evidence on treatment outcomes, consider the patient's preferences and values, provide treatment options to the patient, involve both parties in the care process, create a consensus over the care decision, and support the patient's choice of treatment based on personal preference. Evidence-Based Medicine (EBM) is an attempt to confirm the medical outcome of a treatment based on scientifically obtained evidence [8]. EBM is also defined as the conscientious, explicit, and judicious use of best available evidence in making decisions about the care of individual patients [9]. The underlying principles of EBM include: (1) the clinical decision should be made based on the most up-to-date and best available scientific evidence; (2) from which domain to seek scientific evidence depends on the clinical problem; (3) the best available evidence is determined based on epidemiology and biostatistics; (4) the conclusion derived from EBM needs to be implemented in the clinical decision (i.e., it should be able to influence the physician's treatment of the patient); and (5) how it is implemented should be continuously evaluated [10].

In Section 2.1 Patient's and Family's Rights and Responsibilities of 2017 Hospital Accreditation Standards and Evaluation Criteria (for Medical Centers), Article 2.1.2 describes that it is necessary for hospitals to communicate with patients and explain the medical condition, suggested treatment, and therapy to the patients. Hospitals should also have a code of practice and require patients to sign a letter of consent when an invasive examination or treatment is needed. Developing characteristic and effective communication and explanation methods suitable for patients with diverse needs is considered of great importance [11][12]. Article 2.1.3 indicates that hospitals should explain to hospitalized patients and their families about the necessity of hospitalization and the physician's treatment plan, and also have measures to assist and encourage their engagement in the medical process and decision making. The requirements for compliance include providing adequate assistance to patients and allowing their families to access care information and participate in decision making. Developing policies and guidelines, promoting active participation in decision making among patients and patients' families, and building consensuses between physicians and patients are indicators of high performance [13][14]. Article 2.8.15 indicates that all inspection and examination procedures shall be effectively and safely carried out. The requirement for compliance is that each inspection and examination shall be conducted following a standardized operating procedure that applies, and if necessary, before the inspection or examination is due to begin, hospitals should arrange for a patient assessment and explain to the patient or the patient's family about the procedure [15].

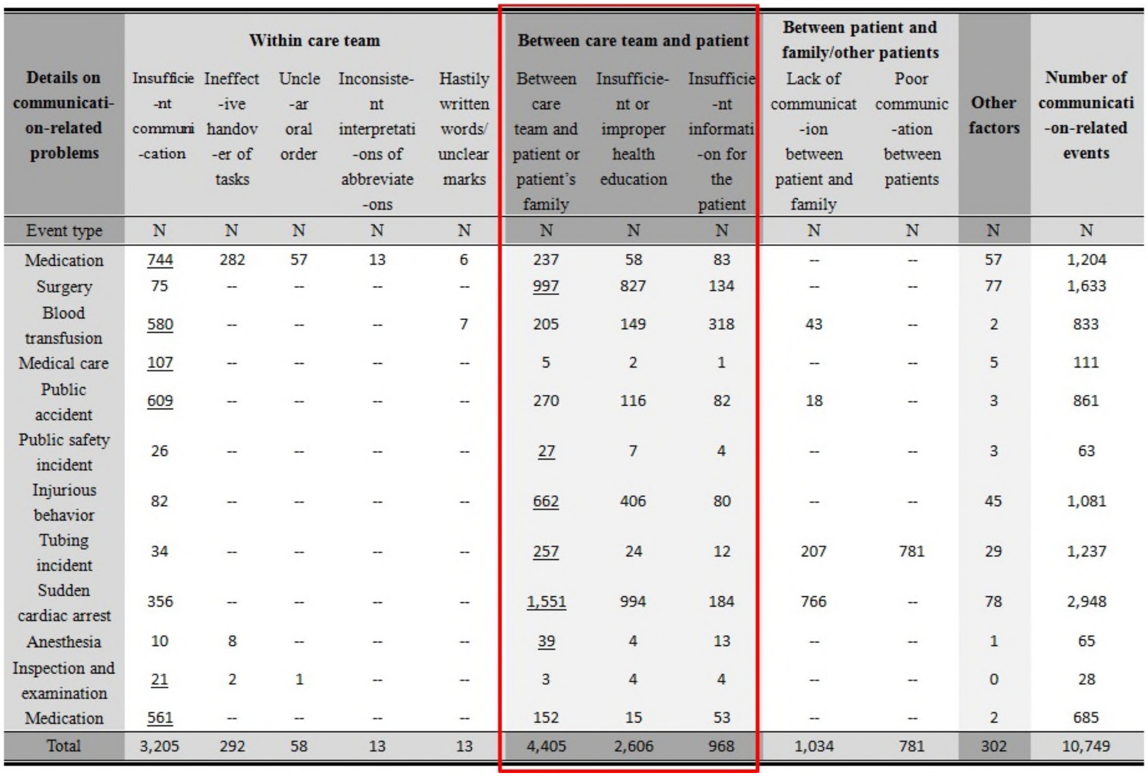
Mayo Clinic is a healthcare system focused on integrated clinical practice, education and research [16]. The Mayo Clinic Shared Decision Making National Resource Center aims to advance patient-centered medical care by promoting shared decision making through the implementation and assessment of patient decision aids and share decision making techniques. The philosophy of the center is "the best interest of the patient is the only interest to be considered". Patients and physicians have different expertise when it comes to making consequential clinical decisions. While physicians know information about the disease and how to treat it, patients know information about their physical conditions, goals for life, and healthcare. Only collaboration in decision making ensures that the ideal of EBM can come true [17].

On Shared Decision Making platform, patients and their families are guided to express their concerns following a structured procedure. Through discussion, physicians and patients can minimize their cognitive gap. Meanwhile, attending physicians can also demonstrate to resident physicians and medical interns how to interact with patients based on a learning by doing approach. This allows resident physicians to learn from the most comprehensive and practical SDM. The rest of the paper is structured as follows. In Section 2, we present the SDM system analysis and design. In Section 3, we have the results and evaluation. Section 4 discusses the process improvement and limitations. We conclude the work in Section 5.

II. METHOD

A. SDM Resources Platform for Nuclear Medicine I-131 Treatment

In this study, we used the nuclear medicine I-131 treatment provided in Kaohsiung Chang Gung Memorial Hospital as an example. Based on EMB and SDM, we



Details on communication-related problems	Within care team					Between care team and patient			Between patient and family/other patients		Other factors	Number of communication-related events
	Insufficient communication	Ineffective handover of tasks	Unclear oral order	Inconsistent interpretations of abbreviations	Hastily written words/ unclear marks	Between care team and patient or patient's family	Insufficient or improper health education	Insufficient information for the patient	Lack of communication between patient and family	Poor communication between patients		
Event type	N	N	N	N	N	N	N	N	N	N	N	N
Medication	744	282	57	13	6	237	58	83	--	--	57	1,204
Surgery	75	--	--	--	--	997	827	134	--	--	77	1,633
Blood transfusion	580	--	--	--	7	205	149	318	43	--	2	833
Medical care	107	--	--	--	--	5	2	1	--	--	5	111
Public accident	609	--	--	--	--	270	116	82	18	--	3	861
Public safety incident	26	--	--	--	--	27	7	4	--	--	3	63
Injurious behavior	82	--	--	--	--	662	406	80	--	--	45	1,081
Tubing incident	34	--	--	--	--	257	24	12	207	781	29	1,237
Sudden cardiac arrest	356	--	--	--	--	1,551	994	184	766	--	78	2,948
Anesthesia	10	8	--	--	--	39	4	13	--	--	1	65
Inspection and examination	21	2	1	--	--	3	4	4	--	--	0	28
Medication	561	--	--	--	--	152	15	53	--	--	2	685
Total	3,205	292	58	13	13	4,405	2,606	968	1,034	781	302	10,749

Figure 1. The 2016 report of the Taiwan Patient-safety Reporting System

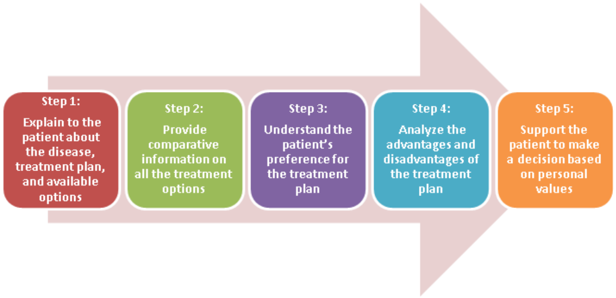

Step 1:
Explain to the patient about the disease, treatment plan, and available options
Step 2:
Provide comparative information on all the treatment options
Step 3:
Understand the patient's preference for the treatment plan
Step 4:
Analyze the advantages and disadvantages of the treatment plan
Step 5:
Support the patient to make a decision based on personal values

Figure 2. The steps for shared decision making proposed by National Health Service Source: Newsletter – June, 2016, Joint Commission of Taiwan

developed an SDM resources platform for nuclear medicine I-131 treatment. This platform consists of five components, including Patient Search System (PSS), Shared Decision Making System (SDMS), Health Education System (HES), Evidence-based Medicine Agent (EBMA), and Data Repository System (DRS). We input physician's explanations about thyroid cancer related diagnoses, goals and effects of the treatment, implementation methods, potential complications, and their occurrence rate and coping methods, success rate and risks from non-treatment, alternative treatments, post treatment reminders, and patients' basic data and health conditions that we had collected into the repository. With this database, the proposed system can directly display information related to the examination result as the physician explains the diagnosis to a patient. The system also takes into account the concerns and questions that patients are likely to have to assist them with making the best decision. The tools used for developing this system include JSP, JSP server Tomcat, MySQL database system, and ontology editor Protégé, all of which are freeware. The system structure is as illustrated in Figure 3. The decision tree is shown in Figure 4 [18]-[20].

B. EBM analysis of I-131 treatment

Figure 5 shows the characteristics according to the American Thyroid Association (ATA) and American Joint Committee on Cancer (AJCC) staging system that may impact post-operative I-131 decision making. The general suggested dose of I-131 is 30~150mCi. Administration of 30mCi I-131 is called a low-dose remnant ablation therapy, while administration of a dose greater than 30mCi is called a high-dose remnant ablation therapy. According to Atomic Energy Council, Executive Yuan, an oral dosage of I-131 greater than 30mCi should be administered only in a hospitalization setting [21][22].

In the current practice, when a patient is not at ATA low risk or not at ATA intermediate risk, but exhibits low risk characteristics or is classified as intermediate risk, the I-131 dose for the patient will be directly determined by the physician. In this study, we integrated patient's data, physical conditions and considerations, examination related information, and points to note in the system analysis to make the decision making mechanism more conscientious and careful, facilitate consensus building over medical decisions, and support the patient to make a decision based on personal preferences. We continuously collected articles addressing the impact of pharmaceutical characteristics on I-131 treatment from various sources, including the hospital's electronic journal systems, the journals and magazines available in the hospital's library, out-of-hospital electronic journal systems, and out-of-hospital journals and magazines. We invited Director Shu-Hua Huang and Dr. Yen-Hsiang Chang of Department of Nuclear Medicine to analyze these articles based on their professional knowledge and expertise. Their analysis results were manually inserted into the database by the research assistant.

C. The principles for I-131 treatment

Because the accuracy and integrity of treatment principles affect the SDM result for the patient, the principles should be established by a domain experts. Thus, we invited Director Shu-Hua Huang and Dr. Yen-Hsiang Chang of Department of Nuclear Medicine to be the co-directors of this study, providing domain knowledge and assisting in analyzing the rules of keywords. For example, for patients who are at ATA low risk or who are at intermediate risk but exhibits low risk characteristics (e.g., metastasis of small-sized central lymph nodes, no residual disease or other abnormal characteristics visible to the naked eye), a low-dose I-131 remnant ablation therapy (approximately 30mCi) is more commonly accepted after they undergo a total thyro-

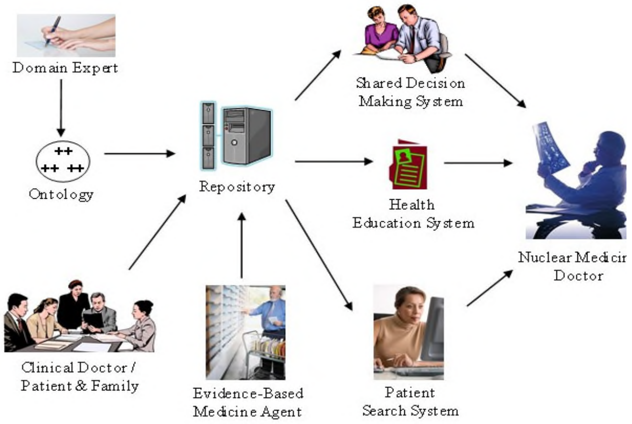


Figure 3. The system structure

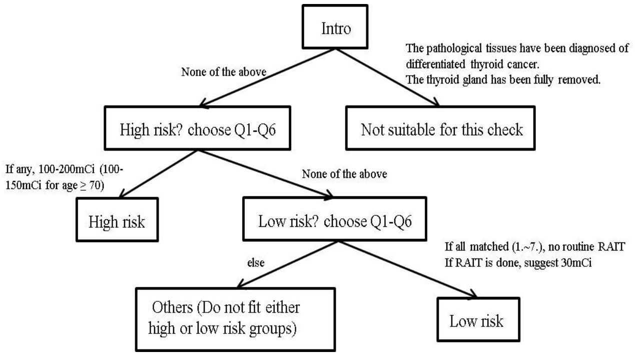
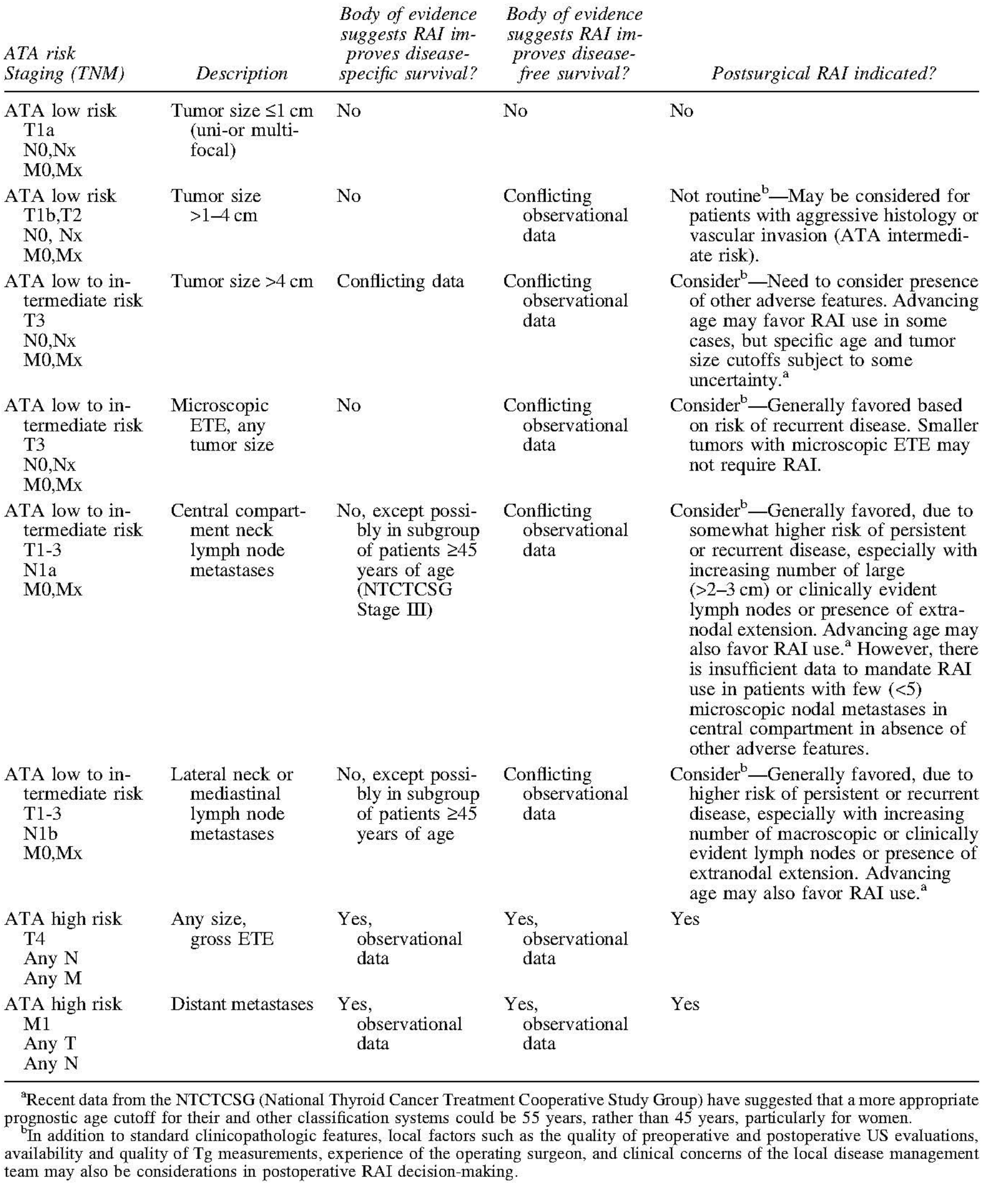

Intro
None of the above
The pathological tissues have been diagnosed of differentiated thyroid cancer.
The thyroid gland has been fully removed.
High risk? choose Q1-Q6
Not suitable for this check
If any, 100-200mCi (100-150mCi for age ≥ 70)
None of the above
High risk
Low risk? choose Q1-Q6
else
If all matched (1.~7.), no routine RAIT
If RAIT is done, suggest 30mCi
Others (Do not fit either high or low risk groups)
Low risk

Figure 4. The decision tree

idectomy. Finally, based on the analysis results, we used the ontology editor Protégé to develop I-131 treatment principles, is shown in Figure 6.

III. Ease of Use

This study was intended to use an information system to assist patients in understanding the examination items involved in the I-131 therapy. The proposed SDM resources platform for I-131 treatment provide a standardized and effective communication procedure for physicians and patients, allowing each patient to receive a personalized and appropriate type of treatment. Besides, through the standardized procedure of this system, attending physicians can also step-by-step guide their resident physicians through interactions with their patients based on the concept of learning by doing. This allows resident physicians to learn from the most comprehensive and practical SDM. The original procedure was that the patient had completed thyroidectomy in the general surgery to the metabolic department to receive radioactive iodine 131 treatment. But in the outpatient clinic, if you encounter health education or hospitalization related questions, you need to answer the nuclear medicine department. After the implementation of SDM, it will be changed to SDM for the first time to complete the understanding of the iodine 131 treatment process, and then to the metabolic department; the patient does not need to go back and forth between two subjects. This is shown in Figure 7.



ATA risk Staging (TNM)	*Description*	*Body of evidence suggests RAI improves disease-specific survival?*	*Body of evidence suggests RAI improves disease-free survival?*	*Postsurgical RAI indicated?*
ATA low risk T1a N0,Nx M0,Mx	Tumor size ≤1 cm (uni-or multi-focal)	No	No	No
ATA low risk T1b,T2 N0, Nx M0,Mx	Tumor size >1–4 cm	No	Conflicting observational data	Not routine[b]—May be considered for patients with aggressive histology or vascular invasion (ATA intermediate risk).
ATA low to intermediate risk T3 N0,Nx M0,Mx	Tumor size >4 cm	Conflicting data	Conflicting observational data	Consider[b]—Need to consider presence of other adverse features. Advancing age may favor RAI use in some cases, but specific age and tumor size cutoffs subject to some uncertainty.[a]
ATA low to intermediate risk T3 N0,Nx M0,Mx	Microscopic ETE, any tumor size	No	Conflicting observational data	Consider[b]—Generally favored based on risk of recurrent disease. Smaller tumors with microscopic ETE may not require RAI.
ATA low to intermediate risk T1-3 N1a M0,Mx	Central compartment neck lymph node metastases	No, except possibly in subgroup of patients ≥45 years of age (NTCTCSG Stage III)	Conflicting observational data	Consider[b]—Generally favored, due to somewhat higher risk of persistent or recurrent disease, especially with increasing number of large (>2–3 cm) or clinically evident lymph nodes or presence of extranodal extension. Advancing age may also favor RAI use.[a] However, there is insufficient data to mandate RAI use in patients with few (<5) microscopic nodal metastases in central compartment in absence of other adverse features.
ATA low to intermediate risk T1-3 N1b M0,Mx	Lateral neck or mediastinal lymph node metastases	No, except possibly in subgroup of patients ≥45 years of age	Conflicting observational data	Consider[b]—Generally favored, due to higher risk of persistent or recurrent disease, especially with increasing number of macroscopic or clinically evident lymph nodes or presence of extranodal extension. Advancing age may also favor RAI use.[a]
ATA high risk T4 Any N Any M	Any size, gross ETE	Yes, observational data	Yes, observational data	Yes
ATA high risk M1 Any T Any N	Distant metastases	Yes, observational data	Yes, observational data	Yes

[a] Recent data from the NTCTCSG (National Thyroid Cancer Treatment Cooperative Study Group) have suggested that a more appropriate prognostic age cutoff for their and other classification systems could be 55 years, rather than 45 years, particularly for women.

[b] In addition to standard clinicopathologic features, local factors such as the quality of preoperative and postoperative US evaluations, availability and quality of Tg measurements, experience of the operating surgeon, and clinical concerns of the local disease management team may also be considerations in postoperative RAI decision-making.

Figure 5. The characteristics that may impact post-operative I-131 decision-making
Source: 2015 American Thyroid Association Management Guidelines

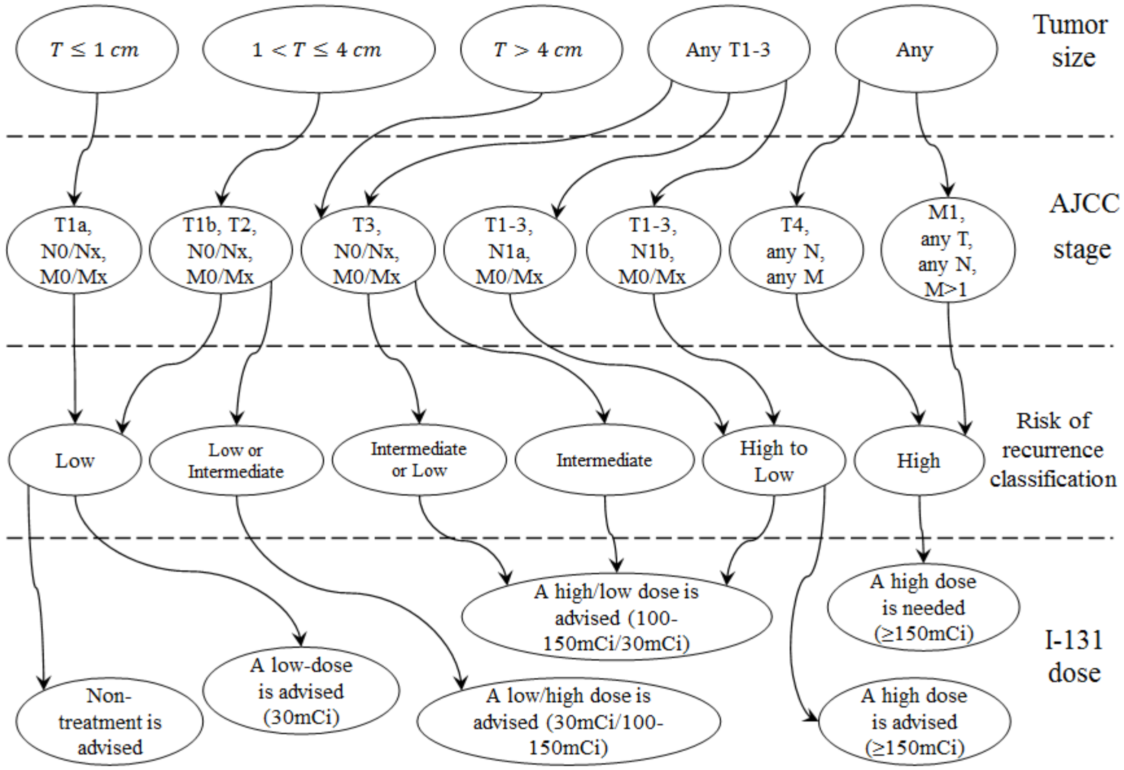
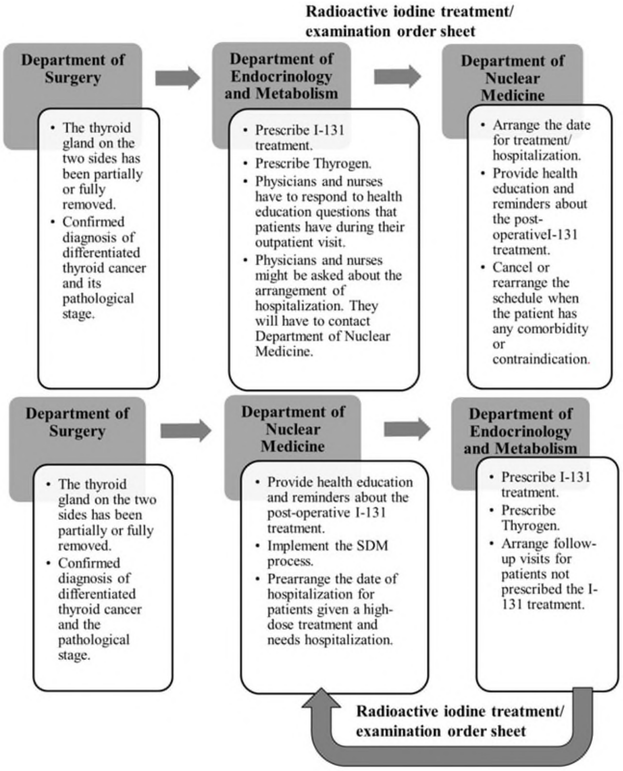


Figure 6. The principals for I-131 treatment


Radioactive iodine treatment/ examination order sheet

Department of Surgery

- The thyroid gland on the two sides has been partially or fully removed.
- Confirmed diagnosis of differentiated thyroid cancer and its pathological stage.

Department of Endocrinology and Metabolism

- Prescribe I-131 treatment.
- Prescribe Thyrogen.
- Physicians and nurses have to respond to health education questions that patients have during their outpatient visit.
- Physicians and nurses might be asked about the arrangement of hospitalization. They will have to contact Department of Nuclear Medicine.

Department of Nuclear Medicine

- Arrange the date for treatment/ hospitalization.
- Provide health education and reminders about the post-operativeI-131 treatment.
- Cancel or rearrange the schedule when the patient has any comorbidity or contraindication.

Department of Surgery

- The thyroid gland on the two sides has been partially or fully removed.
- Confirmed diagnosis of differentiated thyroid cancer and the pathological stage.

Department of Nuclear Medicine

- Provide health education and reminders about the post-operative I-131 treatment.
- Implement the SDM process.
- Prearrange the date of hospitalization for patients given a high-dose treatment and needs hospitalization.

Department of Endocrinology and Metabolism

- Prescribe I-131 treatment.
- Prescribe Thyrogen.
- Arrange follow-up visits for patients not prescribed the I-131 treatment.

Radioactive iodine treatment/ examination order sheet

Figure 7. Analysis of the SDM procedure

A. Analysis of the SDM procedure

Since 2016, SDM has been listed by the Joint Commission of Taiwan (JCT) as one of the strategies for "encouraging patients and patient families to participate in patient safety tasks". To comply with JCT's policy, our hospital has also set SDM as a focus of our medical quality and patient safety activities for 2017. In collaboration with the Department of Endocrinology and Metabolism, our department analyzed the procedure for implementing SDM for thyroid cancer patients who are suggested to receive the I-131 treatment after a thyroidectomy. In the original procedure, patients would undergo the thyroidectomy in the Department of Surgery first and then be transferred to the Department of Endocrinology and Metabolism for the I-131 treatment. The physicians of the Department of Endocrinology and Metabolism would directly prescribe a low-dose or a high-dose I-131 treatment for the patients. Later, the patients would have to visit our department for an arrangement of the dates of hospitalization and treatment. However, the clinical efficiency in our department might be affected when our outpatients need health education or have questions about hospitalization. Besides, cancellation or rearrangement of hospitalization or treatment date due to comorbidity or contraindication is also common in our hospital. Changes of this kind would cause trouble to both patients and physicians. With the implementation of SDM, the procedure has been modified as follows: patients visit our department first to understand the full procedure of the I-131 treatment, and then make a treatment decision that best suits their preferences and conditions through shared decision making. Afterwards, they visit the Department of Endocrinology and Metabolism for the prescription. This procedure can save patients time and the hassle of visiting two departments more than once.

B. *Design of the SDM system*

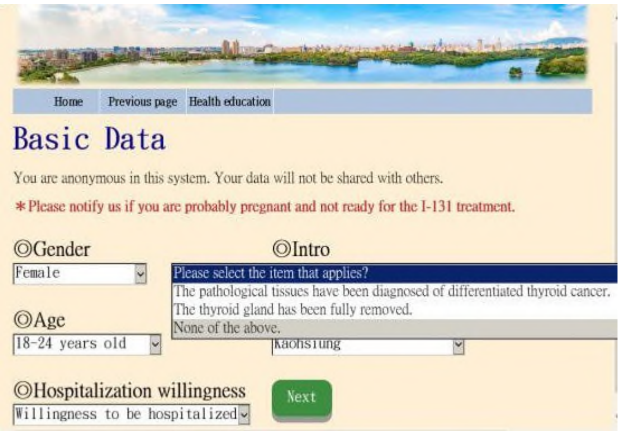
Patient privacy is highly respected in our department. Hence, patients are kept anonymous when using the system. Based on the design concepts for decision-making support tools introduced by Mayo Clinic Shared Decision Making National Resource Center, we adopted a streamlined and simple layout for our system interface. This system has been designed to provide messages in a clear and straightforward manner and guide patients through the questionnaire question by question, based on their post-operative status. This is shown in Figure 8. By applying the principles for I-131 treatment, the system would carry out a decision-tree analysis to infer the risk type (high, intermediate, and low) for treatment and non-treatment, overall mortality, the treatment's potential side effects and impact on daily life, and alternative treatment options for each patient. As most patients who need the I-131 treatment are elderly people, we considered the relatively higher difficulty of text reading in elders in the design of our system interface. We adopted a colorful layout and larger fonts for text messages. Moreover, we also designed the system to visualize percentage data using pictogram charts, which pop up only when the mouse moves over them to keep the display simple and clear. If patients still have difficulty reading the text on the screen, they can click the "Show in a new window" button. The content will then be enlarged in a new window, as shown in Figure 9.

C. *Integrated health education and consultation services*

Health education is an indispensable step in the SDM process. Without appropriate health education aids or a pre-established health education process, patients may not have sufficient understanding of the side effects of the treatment, necessary care at home, and low-iodine diet. In the present, our department uses health education leaflets to provide health information to patients. Although the leaflets contain rich information, including an introduction to the I-131 treatment, indications, points to note, side effects, care at home, and low-iodine diet, physicians or nurses have to use highlights or add notes to remind patients of important information. After implementation of the SDM platform's integrated health education and consultation services, physicians or nurses can simply rely on the standardized health education procedure and the hierarchically structured webpages to explain the instructional materials to patients page after page. As all the important messages are presented in larger fonts and different colors, patients gain sufficient understanding of the treatment after the health education and are able to directly enter the SDM process, where they will have effective communications with their physicians and determine the treatment method that best suits them. The integrated health education and consultation services are as shown in Figure 10.

D. *Effectiveness evaluation*

During September 2017~May 2018, we conducted a questionnaire survey on the proposed SDM platform among post-operative patients who underwent the I-131 treatment.



Basic Data

You are anonymous in this system. Your data will not be shared with others.

＊Please notify us if you are probably pregnant and not ready for the I-131 treatment.

◎Gender

Female

◎Intro

Please select the item that applies?
The pathological tissues have been diagnosed of differentiated thyroid cancer.
The thyroid gland has been fully removed.
None of the above.

◎Age

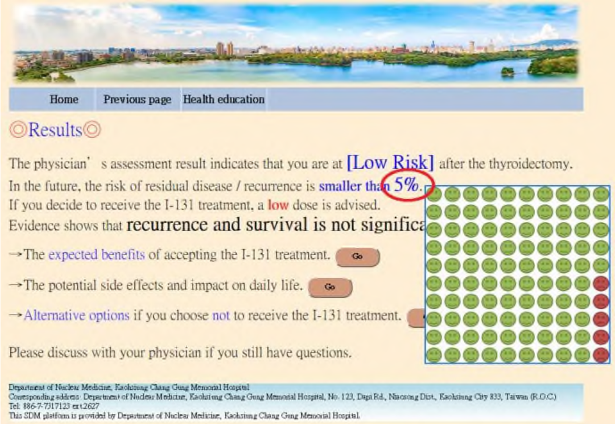
18-24 years old

Kaohsiung

◎Hospitalization willingness

Willingness to be hospitalized

Figure 8. The anonymous user interface



◎Results◎

The physician' s assessment result indicates that you are at [Low Risk] after the thyroidectomy.
In the future, the risk of residual disease / recurrence is smaller than 5%.
If you decide to receive the I-131 treatment, a low dose is advised.
Evidence shows that recurrence and survival is not significa

→The expected benefits of accepting the I-131 treatment.

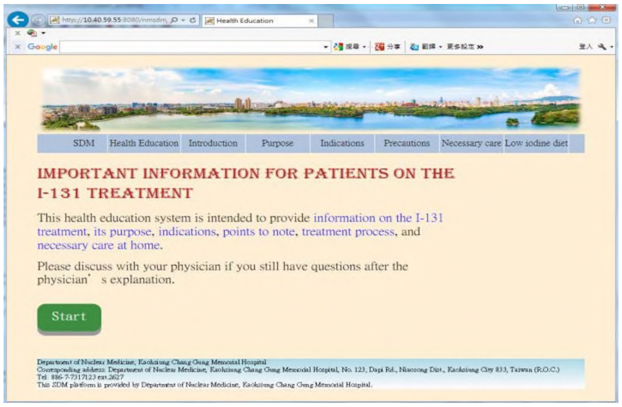
→The potential side effects and impact on daily life.

→Alternative options if you choose not to receive the I-131 treatment.

Please discuss with your physician if you still have questions.

Department of Nuclear Medicine, Kaohsiung Chang Gung Memorial Hospital
Corresponding address: Department of Nuclear Medicine, Kaohsiung Chang Gung Memorial Hospital, No. 123, Dapi Rd., Niaosong Dist., Kaohsiung City 833, Taiwan (R.O.C.)
Tel: 886-7-7317123 ext.2627
This SDM platform is provided by Department of Nuclear Medicine, Kaohsiung Chang Gung Memorial Hospital.

Figure 9. Using a pictogram chart to visualize percentage data



IMPORTANT INFORMATION FOR PATIENTS ON THE I-131 TREATMENT

This health education system is intended to provide information on the I-131 treatment, its purpose, indications, points to note, treatment process, and necessary care at home.

Please discuss with your physician if you still have questions after the physician' s explanation.

Department of Nuclear Medicine, Kaohsiung Chang Gung Memorial Hospital
Corresponding address: Department of Nuclear Medicine, Kaohsiung Chang Gung Memorial Hospital, No. 123, Dapi Rd., Niaosong Dist., Kaohsiung City 833, Taiwan (R.O.C.)
Tel: 886-7-7317123 ext.2627
This SDM platform is provided by Department of Nuclear Medicine, Kaohsiung Chang Gung Memorial Hospital.

Figure 10. Integrated health education and consultation service

A total of 64 copies of the questionnaire were distributed, and 64 responses were returned (100% response rate). Participants were classified by whether they have used the SDM platform, and the questions were intended to measure patients' understanding of SDM on a Likert scale ranging from 1 to 5. The questionnaire of SDM for the I-131 treatment consisted of 8 questions. The analysis of the responses showed that the average score among non-SDM users was 1.3 and that among users reached as high as 4.89. The increased understanding of the treatment suggests that the proposed platform has helped reduce the extent of worry and fear about the I-131 treatment in patients. Because of better understanding, certain concerns would still reside in patients. Therefore, these concerns were excluded in the calculation of mean score. To assess patients' conditions after using the platform, we added two questions in the post-test questionnaire: satisfaction with the I-131 treatment decision and trust for the physician. The mean was 4.97, suggesting that SDM has helped improved physician-patient relationship. The survey results are shown in Figure 11.

IV. DISCUSSION

A. Implementing SDM to improve the I-131 treatment procedure

In recent years, JCT has been very active in promoting SDM. No matter in annual goals of hospital medical quality and patient safety or in hospital accreditation standards and evaluation criteria, SDM is considered an important strategy. According to Taiwan Patient-safety Reporting System, communication problems between care team and patients and patients' families have always constituted a high percentage. The main causes of these problems include: insufficient information is given to the patient during the diagnostic process; the physician has many patients to take care of but limited time for communication; there is a wide knowledge gap between the physician and the patient; and the use of medical terminologies has made it hard to understand the physician's explanation, resulting in misunderstanding and even dispute. The existing journal and research articles on application of SDM have focused mainly on application of SDM in invasive examinations, treatments of high-risk diseases or paper work operation. Research integrating nuclear medicine radioactive examinations into a SDM platform is rare. In our survey, the mean scores for all the questions were higher in the post-test, confirming the effectiveness of the proposed SDM platform for patients and information systems to assist the less.

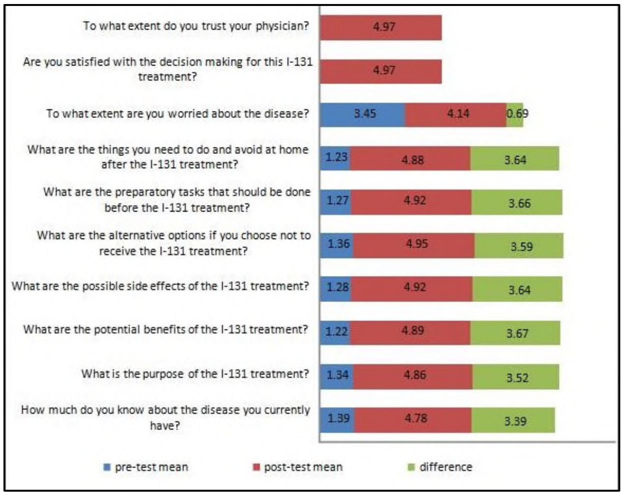


Figure 11. Patients' understanding of the I-131 treatment

B. SDM implementation strategy

Under the department director's support, we implemented the proposed system as a research project. The engineers of the department worked in collaboration with the attending physicians. After several meetings, we finally submitted our project in March 2017 to Department of Medical Research for review. The project was approved in June 2017. For physicians, the time they have to spend on SDM is greater than the time they used to spend on health education, but the SDM platform has alleviated much of their effort. Besides, physicians' communication skills and attitude also affect the success of SDM. If given training on communication skills, physicians might be more able to lead Asian patients to express their opinions during SDM.

V. CONCLUSION

In this study, we used computer programming to develop a SDM platform that guides patients and their families to express their main concerns following a structured procedure and facilitates discussion between physicians and patients to reduce the cognitive gap between the two parties.

The proposed system conforms to the three elements advocated by JCT, namely knowledge, communication, and respect. Results confirmed that it can drastically improve the medical quality, patient safety, and satisfaction of thyroid cancer patients. Besides, the system provides a standardized communication procedure, allowing attending physicians to guide their resident physicians through interactions with patients based on the approach of learning by doing. Their resident physicians thus have an opportunity to learn from the most comprehensive and practical SDM procedure.

ACKNOWLEDGMENT

This study was successfully carried out under the support of Grant CMRPG8G0211 from Chang Gung Memorial Hospital, Kaohsiung, Taiwan.

REFERENCES

[1] Taiwan Patient Safety Net, Taiwan Patient Safety Goals for Hospitals, Ministry of Health and Welfare, 2019.

[2] L. Y. Tsai and C. C. Lin, "Using Shared Decision-Making on a Patient with Renal Cell Carcinoma and Subcutaneous Metastasis: A Care Experience," The Journal of Nursing, vol.62, pp.89-94, 2015, doi: 10.6224/jn.62.3.89.

[3] Taiwan Patient Safety Net, Annual Report 2016, Ministry of Health and Welfare, 2019.

[4] United States, Making health care decisions: a report on the ethical and legal implications of informed consent in the patient-practitioner relationship, President's Commission for the Study of Ethical Problems in Medicine and Biomedical and Behavioral Research, 1982.

[5] C. Charles, A. Gafni, and T. Whelan, “Shared decision-making in the medical encounter: What does it mean?,” Social Science & Medicine, vol.44, pp.681-692, 1997, doi: 10.1016/S0277-9536(96)00221-3.

[6] W. P. Hsu, R. Y. Chang, M. C. Lu, M. C. Chou, and P. C. Hsiao, “Shared Decision Making in Clinical Practice,” Cheng Ching Medical Journal, vol.11, pp.24-29, 2015.

[7] Ministry of Health and Welfare Platform for Shared Decision Making, Shared Decision Making, 2019, [Online]. Available from:http://sdm.patientsafety.mohw.gov.tw/Public/Detail?sn=32&id=1035

[8] Wikipedia, Evidence-based medicine, 2019, [Online]. Available from: https://en.wikipedia.org/wiki/Evidence-ba sed _medicine

[9] Chang Shan Medical University Hospital, Getting to Know Evidence-based Medicine, 2019, [Online]. Available from: http://www.c- sh.org.tw/into/medline/main02.htm

[10] Department of Health, Taipei City Government, An Introduction to Evidence-based Medicine, 2019, [Online]. Available from: http://health.gov.taipei/Default.aspx?tabid=4 17&mid=537&itemid=15639

[11] M. Coylewright, E. S. O'Neill, S. Dick, and S. W. Grande, “PCI Choice: Cardiovascular clinicians' perceptions of shared decision making in stable coronary artery disease,” Patient Educ Couns, vol.100, pp.1136-1143, 2017, doi: 10.1016/ j.pec.2017.01.010.

[12] S. B. Wilton and K. J. Lyons, “The Challenge of Measuring Adherence With Implantable Cardioverter Defibrillator Referral Guidelines in the Era of Shared Decision-Making,” Canadian Journal of Cardiology, vol.33, pp.420-421, 2017, doi: 10.1016/j.cjca.2016.11.008.

[13] F. Bunn et al., “Supporting shared decision-making for older people with multiple health and social care needs: a protocol for a realist synthesis to inform integrated care models,” BMJ Open, vol.7, 2017, doi:10.1136/bmjopen-2016-014026.

[14] J. L. J. Yek et al. , “Defining reasonable patient standard and preference for shared decision making among patients undergoing anaesthesia in Singapore.” BMC Medical Ethics, vol.18, 2017 doi: 10.1186/s12910-017-0172-2.

[15] Joint Commission of Taiwan, 2017 Hospital Accreditation Standards and Evaluation Criteria –Trial Version for Medical Centers, 2019, [Online]. Available from: https://www.jct.org.tw/mp-2.html

[16] Mayo Clinic, 2019, [Online]. Available from: https://www.mayoclinic.org/

[17] MayoClinic, Mayo Clinic Shared Decision Making National Resource Center, 2019, [Online]. Available from https://share ddecisions.mayoclinic.org

[18] Apache Tomcat, Apache Software Foundation, 2019, [Online]. Available from: https://tomcat.apache.org/download-70.cgi

[19] MySQL, Oracle, 2019, [Online]. Available from: https://www.mys ql.com/

[20] Protégé, Stanford University, 2019, [Online]. Available from: https://protege.stanford.edu/

[21] B. R. Haugen et al., “2015 American Thyroid Association Management Guidelines for Adult Patients with Thyroid Nodules and Differentiated Thyroid Cancer,” Thyroid, vol. 26, 2016 doi: 10.1089/thy.2015.0020

[22] Administrative Regulations for Radioactive Material and Equipment Capable of Producing Ionizing Radiation and Associated Practice, Atomic Energy Council, 2019, [Online]. Available from: https://erss.aec.gov.tw/law/EngLawContent.aspx?lan=E&id=77

Collaborative Filtering Based Recommender System Design For E-Commerce: A Case Study

Merve Artukarslan
Galatasaray University
Department of Industrial Engineering
Istanbul, Turkey
email: merveartukarslan@gmail.com

S. Emre Alptekin
Galatasaray University
Department of Industrial Engineering
Istanbul, Turkey
email: ealptekin@gsu.edu.tr

Abstract— **Recommender Systems (RS) are one of the core engagement functions for e-commerce industry. In a typical recommender system, customer and product data is analyzed and a prediction model is generated, which evaluates products for prospective customers. In terms of business value, it helps individuals identify their interest among an overwhelming variety of products. In this paper, a collaborative filtering based recommender system framework is proposed for Turkey's leading e-commerce platform hepsiburada. First of all, implicit feedback and customer-product prediction pairs are prepared from collected data. Second, a regularized Singular Value Decomposition (SVD) based matrix factorization model is established for Collaborative Filtering (CF). Customers and products are represented with latent factor vectors. This model is trained with implicit feedback, as the SVD problem is solved with Alternating Least Squares (ALS). Third, predictions are gathered from the CF model. Then, predictions are limited to ten-product recommendation sets. Finally, recommendations are evaluated by behavioral data generated by prospective customers. The initial results show that 19% of recommendations match customers' interests.**

***Keywords-** Collaborative Filtering; Singular Value Decomposition; Alternating Least Squares; Recommender Systems; Matrix Factorization.*

I. INTRODUCTION

Nowadays, e-commerce platforms accommodate a high diversity of choices for a vast number of visitors. Under these circumstances, there are two anticipated challenges. One of them is to expedite the decision making process for the prospective customers and the other one is scaling the technological solutions for high demand.

People usually consider the recommendations and mentions of their peers for purchase decisions. Recommender systems in this context are intelligent pieces of software, which interpret the digital footprint of users and products and then predict users' future behavior or requirements. They count as one of the core engagement functions of modern online retail businesses. These tools serve users with personalized recommendations suiting their unique desires and tastes via different feedback mechanisms.

Explicit feedback is defined as the categorical assessment of the customer for a product regarding their interest such as star ratings. Implicit feedback is the customer behavior inferring the user's preferences. Purchase history, browsing history, search patterns or page view period are examples of implicit feedback. Explicit feedback is preferred as it leads to a pure classified information. However, implicit feedback is less limited in terms of data collection effort. The numerical value of explicit feedback represents the customer preference, while implicit feedback supports its confidence [1].

This study aims to build a comprehensive recommender system for an e-commerce platform by targeting the prospective customers based on behavior. Recommender systems are introduced in Section 2. Value proposition of implicit feedback and collaborative filtering are analyzed to introduce the essence of customer preference notion. Section 3 consists of the proposed methodology. The purpose behind using matrix factorization and ALS is explained. The utilization of ALS with weighted λ regularization is detailed. Then, implicit feedback data with confidence level approach is introduced. Section 4 presents a business case implementation. We fine-tune the model parameters and hyper parameters of the SVD problem. We retrieve predictions from the model for the predefined customer-product pairs. Finally, prediction and recall metrics are calculated and analyzed. Future work and improvements are also discussed in Section 5.

II. RECOMMENDATION SYSTEMS

CF is a recommendation algorithm based on selecting and aggregating other users' behavior and ratings. It was first articulated by Goldberg in 1992 as a collaboration of people to aid one another to execute document filtering by interpreting readers' reaction to documents they read [2]. A user's preference is predicted, in a way, by interpreting others opinions. If users agree about the relevance of certain items, they will likely agree about others. CF highlights the serendipity of recommendations [3].

Besides CF, there are also other recommendation methodologies. Content based recommendations are based on user and product profiles which require external information and strategy. User content and product content are associated for recommendations. However, CF is based on users' behavioral history. Users' analytical judgments for the products they use are shared for better decisions. A personalized recommendation set is the deliverable of a CF based model.

Comparing to content based filtering, collaborative filtering has several major advantages. First of all, in CF, the only information needed is a user showing an intention to a product. Second, CF engages as a product satisfies a user's wishes. Hence, it aims to be more than a mere content analysis. Quality or taste as woven within human decisions is incorporated into the recommendation. Third, people may make desirable decisions by accident. The CF technique can generate serendipitous recommendations, which are valuable to the user but not expected according to the content of the product or user [4]. Despite the advantages, there are also disadvantages of collaborative filtering. Regarding the sparsity of user preference, it is not easy to find users with similar intentions. Recommendations may include many similar products and outliers may bias the model.

There are two frequently used approaches for CF. The first one is based on neighborhood methods discovering the relationships within the users or products. The second approach is based on latent factors models. A simple recommender system models the similarities between people or products. A latent factor model tackles the problem with a more sophisticated approach by converting data into a theme space. Then, the similarities in this theme space are explored. Latent factor models are preferred as latent space explains ratings by characterizing both users and products as factors inferred from implicit feedback [5].

In an user to user CF, one's predicted preference is based on the similarity with other users in terms of common ratings [6]. There are several methods for calculating user similarities. Pearson correlation [4] finds the statistical correlation between two users' common ratings to determine similarity. One pitfall of this method is that it may result in a high similarity between users that have few ratings in common. Constrained Pearson correlation scales the ratings in a like-dislike range [7]. Spearman rank correlation coefficient [4] is another method. It is derived from Pearson, except the ratings are replaced by ranks. Cosine similarity is a vector-space approach where users are represented by item rating vectors and similarity is measured by cosine distance between item rating vectors. The cosine distance is calculated by dividing the dot product of two vectors by the product of their Euclidean norms [8]. Item to item CF, on the other hand, uses the similarities between the rating patterns of items. The similarity of items can be calculated by the methods mentioned for user similarity.

Schafer et al. [15] explained how recommender systems bring value to e-commerce systems and analyzed different e-commerce platforms. RS enhance e-commerce services by converting browsers into buyers by utilizing sorting tasks for users. As mentioned in [15], Amazon book recommendations are based on 'customers who bought' strategy which refers to user similarity. They also recommend authors other than items. EBay builds a feedback profile feature allowing buyers and merchants to provide satisfaction rating. Feedback is used for merchant recommendation for users. They also have a personal shopper feature, which allows users to flag items they are interested in.

User's preference for an item in a recommender system is represented by the combination of the user's interest in the topic and an item's relevance to the topic. Hence, user-item ratings are established using a vector-space, which is likely to be high dimensional. This high dimensional model representation constructs a purchase history vector for each customer by producing one output for a set of inputs at a time. Prospective customers are targeted and similar customers are identified with cosine similarity based on the purchase history vectors of customers [9]. To increase robustness of the model and simplify model training, dimensionality reduction of rating space by dropping the singular values is recommended. This conversion helps to reduce noise in data and results in higher quality recommendations, as strong trends in the model are kept [10].

Similarity evaluations in CF approaches have to deal with the sparsity of data and dependency to common rated items. Solutions have been proposed to deal with these issues [11]. Wang et al. [11] developed an extend Proximity–Significance–Singularity model combined with item similarity. Their approach was tested in various sparse data sets and their results promise flexibility and break the constraint of common rated items. Furthermore, CF literature makes use of matrix factorization methodology to increase the level of accuracy and scalability. Single value decomposition technique as part of matrix factorization is applied to identify latent semantic factors. The latent space characterizes products and users on factors which are a form of user feedback and demonstrate ratings. The user and item latent factors are calculated using the alternating least squares technique which solves the optimization problem by fixing in each iteration either user latent factors or item latent factors and solving for the other and iterating until conversion [13].

In order to deal with limited data availability in recommender systems, implicit feedback based recommender systems have been developed [12]. The model is optimized by minimizing a ranking objective problem instead of the conventional mean square error. The key components of this model are a matrix factorization model, a ranking based objective function and an optimizer [12].

In real life scenarios, a vast number of user-item pairs complicates the optimization process. Methodologies demonstrated as in [1] make use of SVD for implicit feedback dataset-based collaborative filtering applications. Since the cost function of an SVD contains a vast number of user-item pairs, this minimization problem cannot be solved by a conventional technique, such as stochastic gradient descent. Hence, the quadratic nature of the cost function of ALS methodology proves useful as its complexity linearly increases in data size [1]. Moreover, as proposed in [1], the data sparsity and dense cost function could be dealt with confidence levels implementations. New factor models are also proposed by dividing ratings into confidence level and prediction [1].

III. Proposed Methodology

User and item similarity based recommender systems are implemented for e-commerce systems. However, they require manual effort for content profiling and are based on unchanged customer behavior, contrary to real life. User similarity based CF is preferred, as it is close to a persistent and automated process, yet it assumes the sessions are long enough to learn about customer patterns [15]. CF with implicit feedback data has major advantages such as behavioral data dependency, focusing on customer preference, serendipity of recommendations and efficient responsiveness for cold starters. As mentioned in [14], CF techniques are commonly utilized for suggesting products to users. Considering the penetration and traffic of digital retail services, scalability and user profile sparseness problems eventually arises. [14] proposed an ALS algorithm with weighted λ regularization, which tackles these problems.

A. *Alternating Least Squares for Optimizing Singular Value Decompostion Problem*

Matrix factorization is a method used for latent factor models. It characterizes users and items as vectors of factors inferred from item rating patterns. A valuable recommendation carries high correspondence within user and item factors. This method is preferred for two reasons, it is scalable and accurate in predictions. Matrix factorization models fit users and items into a latent factor space of dimensionality. User-item interactions are considered as inner products in this space.

Singular value decomposition is a technique for identifying latent semantic factors in information retrieval. In collaborative filtering, it is applied to the user-item rating matrix. To learn the factor vectors, the model minimizes the regularized squared error on the set of known ratings. The learning model is generated by fitting the previous quantitative implicit feedback data in terms of ratings. The overall goal of a model is to reuse the model for unknown rating predictions [13].

Alternating Least Squares factorizes a rating matrix into two factors, user and item matrices, having the number of latent factors as row dimension. By fixing one of the matrices, the problem becomes quadratic, which can be solved directly. Alternately, this step is applied to user and item matrices, and the matrix factorization problem is iteratively improved.

B. *Tackling Overfitting with Weighted λ Regularization*

Overfitting is considered as overtraining a model by feeding it noisy and inaccurate data. Therefore, we may end up with an unrealistic model. Regularization is implemented to reduce the variance of the model without increasing the bias. Bias is considered as the error of the model. Variance is the change in predictions observed with different training models. Therefore, a tuning parameter λ is added to the model to deal with variance. As the tuning parameter increases, it reduces the value of coefficients and variance.

An ALS with weighted λ regularization model is proposed in [14] for large scale collaborative filtering. The main purpose of weighted regularization is to ensure the model does not overfit with increased number of features (latent factors) or iterations. Besides that, the authors mentioned that only about 1% of the user-movie matrix has been observed, with the majority of ratings missing, which is a challenge for the training data. Our study applies the method of implicit data feedback, unlike explicit movie ratings, which is used for the Netflix Prize competition referred in [14]. Equation (1) below is used to calculate the objective function of the model.

$$F(u,i) = \sum_{(u,i)\in\mathcal{K}}(r_{ui} - q_i^T p_u)^2 + \lambda\big(\sum_i n_{q_i}\|q_i\|^2 + \sum_u n_{p_u}\|p_u\|^2\big) \qquad (1)$$

where

$u: user$

$i: item$

$\mathcal{K}: user - item\ pairs\ in\ training\ data$

$r_{ui}: rating\ of\ user\ for\ an\ item$

$\hat{r}_{ui}: estimated\ rating\ of\ user\ for\ an\ item, q_i^T p_u$

$q_i: latent\ factor\ vector\ of\ item\ i, q_i \in R^f$

$p_u: latent\ factor\ vector\ of\ user\ u, p_u \in R^f$

$\lambda: regularization\ factor$

$n_i: number\ of\ items$

$n_u: number\ of\ users$

$n_{q_i}: number\ of\ ratings\ of\ item\ i$

$n_{p_u}: number\ of\ ratings\ of\ user\ u$

$I_u: Set\ of\ items\ that\ user\ u\ rated$

$I_i: Set\ of\ users\ that\ rated\ item\ i$

$Q: item\ feature\ matrix$

$P: user\ feature\ matrix$

$R: user - item\ matrix, \{r_{ui}\}_{n_u \times n_i}$

$Q_{I_u}: Submatrix\ of\ Q,\ i \in I_u$

$R(u, I_u): ratings\ row\ vector\ of\ items\ that\ user\ u\ rated$

$P_{I_U}: Submatrix\ of\ P,\ u \in I_i$

$R(I_i, i): ratings\ column\ vector\ of\ users\ that\ rated\ item\ i$

$n_f: feature\ dimension\ space$

$E: n_f \times n_f\ identity\ matrix$

$R = \{r_{ui}\}_{n_u \times n_i}$ represents the user-item matrix. Each element $\{r_{ui}\}$ represents the implicit feedback from customer u for item i. There is a user and item feature vector corresponding to each and every user and item, denoted by q_i and p_u, respectively. Each given and estimated rating, or implicit feedback, is the inner product of the corresponding latent factor vectors. The authors of [14] suggested to minimize the summation of loss of user and item feature matrices of known ratings, P and Q. The loss function is regularized for handling the overfitting of sparse data set.

Since the SVD algorithm is not able to find P and Q with a large number of missing ratings, ALS is applied. The minimization problem has two sets of decision variables as part of the optimization goal. Therefore, as one of the decision variables set is fixed to solve the problem for the remaining set, the problem is solved. As mentioned in the previous section, ALS rotates the problem by fixing item latent factors and user latent factors sequentially. The least squares computation problem is solved and the regularized squared error is decreased until convergence. ALS is preferred over gradient descent as it can use parallelization. ALS with weighted λ regularization will also address the scalability limitations related to the number of latent factors and the number of ALS epochs.

Matrix $Q = [q_i]$ is initialized by assigning the average rating for an item as the first row, and small random numbers for the remaining entries. Then, Q is fixed and $P = [p_u]$ is solved by minimizing the sum of squared error in the objective function. Then, P is fixed and Q is solved similarly. This rotation is repeated until the mean squared error converges. A given column of P, which latent factor vector of user u denoted as p_u, is determined by solving a regularized linear least squares problem involving the known ratings of user u and feature vectors q_i of the items that user u rated. p_u becomes an expression of (3) and (4), which is given in (2).

$$\frac{1}{2}\frac{\partial f}{\partial i_{kj}} = 0,\ \forall u,k$$
$$\Rightarrow \textstyle\sum_{i\in I_u}(p_u^T q_i - r_{ui})q_{ki} + \lambda n_{p_u} p_{ku} = 0,\ \forall u,k$$
$$\Rightarrow \textstyle\sum_{i\in I_u} q_{ki} p_u^T q_i + \lambda n_{p_u} p_{ku} = \sum_{i\in I_u} r_{ui} q_{ki}\,,\ \forall u,k$$
$$\Rightarrow \left(Q_{I_u} Q_{I_u}^T + \lambda n_{p_u} E\right) p_u = Q_{I_u} R^T(u, I_u),\ \forall u$$

$$p_u = A_u^{-1} V_u,\ \forall u \tag{2}$$

where

$$A_u = Q_{I_u} Q_{I_u}^T + \lambda n_{p_u} E \tag{3}$$
$$V_u = Q_{I_u} R^T(u, I_u) \tag{4}$$

Q_{I_u} denotes the sub-matrix of Q (item feature matrix) consisting of columns $i \in I_u$ (set of items rated by user u). $R(u, I_u)$ denotes the row vector retrieved from the u-th row of R (user-item matrix) for $i \in I_u$ (set of items rated by user u).

Similarly, when Q is updated, individual q_i can be computed via regularized linear least squares solution including the feature vectors of users who rated item i. q_i becomes an expression of (6) and (7), which is given in (5).

$$q_i = A_i^{-1} V_i,\ \forall i \tag{5}$$
$$A_i = P_{I_i} P_{I_i}^T + \lambda n_{q_i} E \tag{6}$$
$$V_i = P_{I_i} R^T(I_i, i) \tag{7}$$

P_{I_U} denotes the sub-matrix of P (user feature matrix) consisting of columns $u \in I_i$ (set of users rated item i). $R(I_i, i)$ denotes the column vector retrieved from the i-th column of R (user-item matrix) for $u \in I_i$ (set of users rated item i) [14].

C. Confidence of Implicit Feedback

As suggested by Hu el al. [1], at this stage, we tried to identify the unique properties of implicit feedback data. The objected function stated in (1), which is based on ALS with weighted λ regularization, is extended in this step. t_{ui} is a binary set which indicates the preference of user u for item i. In other words, if user u has interacted to item i, t_{ui} is equal to 1. On the other hand, if user u never encountered item i, then, that preference is set equal to 0. Preference values are poor in confidence, as having no preference may have a variety of reasons other than not liking an item. Thus, a confidence level model representing the user's preference is required.

Consequently, as r_{ui} grows, the strength of preference should be increased. c_{ui} is measurement for the confidence in t_{ui} equals $(1 + \propto r_{ui})$. The squared error part of the goal function $(r_{ui} - p_u^T q_i)^2$ is extended as $c_{ui}(t_{ui} - p_u^T q_i)^2$. Replacing the ratings with confidence values, A_u, V_u, A_i and V_i are updated, as shown in (8), (9), (10) and (11).

$$A_u = Q_{I_u} C^u Q_{I_u}^T + \lambda n_{p_u} E \tag{8}$$
$$V_u = Q_{I_u} C^u R^T(u, I_u) \tag{9}$$
$$A_i = P_{I_i} C^i P_{I_i}^T + \lambda n_{q_i} E \tag{10}$$
$$V_i = P_{I_i} C^i R^T(I_i, i) \tag{11}$$

C^u is a diagonal $n_i \times n_i$ matrix where $C_{ii}^u = c_{ui}$. C^i is a diagonal $n_u \times n_u$ matrix where $C_{uu}^i = c_{ui}$ [1].

IV. Case Study

A. Business Model & Proposed Framework

In this work, a collaborative filtering based recommendation engine is proposed for an e-retailer. Based on the number of visitors and products, the recommendation engine is utilized for Small Domestic Appliances (SDA) and Fast Moving Consumer Goods (FMCG) categories.

The behavioral events used in this study are listing page visit, product page visit, product added to cart, product saved for later, saved product added to cart and purchase. These events are fitted into a three phase sales funnel, displayed in Figure 1, for customer-product interaction association.

There are three major stages: discovery, intention and purchase. The more a customer carries a product within the funnel, the more involved the customer is. Every event is associated with a stage. Implicit feedback is determined by the last event for each user-item pair. The general framework proposed for this business case is visualized in Figure 2.

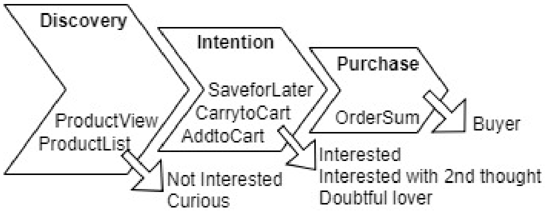


Figure 1. Sales Funnel Design

B. *Training and Prediction Data Preparation*

Implicit feedback is built on event correlations related to the pre-defined products. Collected events are filtered within a time interval of 4 weeks. The start date is selected as January 1st, 2019 and the end date is January 31st, 2019. Interactions are converted to numeric values from 1 to 6. Not interested is 1, curious is 2, interested is 3, interested with second thoughts is 4, doubtful lover is 5 and buyer is 6. The numeric rating values are directly proportional to the incline degree of the interaction portrait. The form of implicit feedback data is a (Customer, Product, Rating) tuple.

Prediction pairs are prepared for users who showed intention or purchased a kitchen appliance product on January 31st, 2019. In short, predictions are generated within one day of interaction by a CF model trained with 4 weeks of feedback data. The users who intended to buy an SDA product but did not purchase one are characterized as inclined users and the ones who purchased one as purchased users. For purchased users, the purchased category is excluded from predictions. For the inclined users, products belonging to the inclined category are predicted.

C. *Recommendation Model Generation*

The proposed model is prepared for evaluating user-item pairs by generating a prediction score. Weighted λ regularization is implemented in the model. Users and items are represented as latent factor vectors for SVD. The rank parameter defines the number of latent factors. The alpha parameter is used as multiplier for rebalancing rating data. The number of iterations represents the number of rotations for ALS. Apache Spark [16] is preferred as it is a large scale data processing platform, which enables parallelized operations.

Mean Squared Error (MSE) is used as optimization metric. When MSE converges, the problem is assumed to be optimized with given parameters. λ is set as 0.01 and model parameters are fine-tuned by MSE convergence. Rank is set from 2 to 128, alpha is sequentially set as 0.01, 0.1, 0.5 and 1. Consequently, how these parameters leverage the MSE is analyzed. Regarding the memory limitations of our sources, the number of iterations is selected as 10 epochs. The model is trained with different rank values when alpha is 1.0 and the number of epochs is 10. The convergence rate is analyzed and the rank is determined to be 60. The MSE is observed to decrease from 6.0488 to 4.3722 with an increased rank and fixed alpha and λ.

D. *Tailored Recommendations*

Despite the fact that predictions are generated for hundreds of products for each user, recommendations are limited to 10 as it is a realistic value for user experience. Regarding the business objectives such as showing the assortment of products and converting more of the cross-sale opportunities, these 10 items are divided into two subgroups. One subgroup represents the products with the highest predicted ranking. The other subgroup is tailored in accordance with business objectives. A total number of 6 items are the ones with highest predictions. 4 items are tailored to ensure that there are both FMCG and SDA products in a 10-product recommendation set.

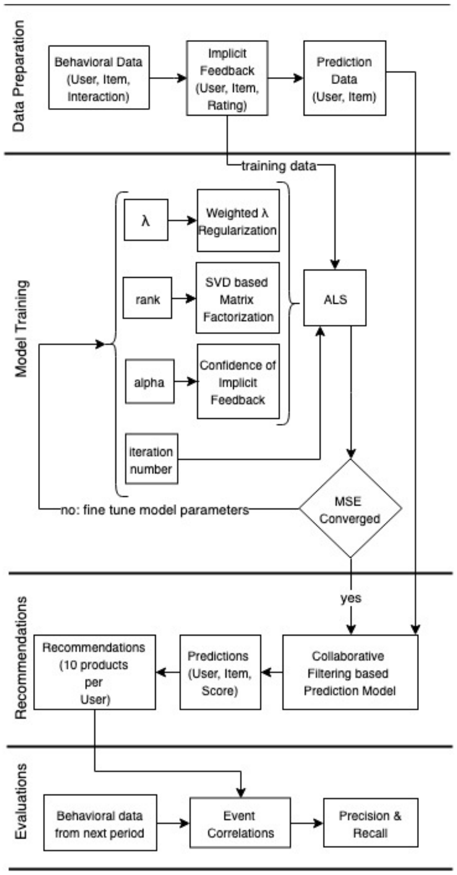

Data Preparation
Behavioral Data (User, Item, Interaction)
Implicit Feedback (User, Item, Rating)
Prediction Data (User, Item)
training data
Model Training
λ
Weighted λ Regularization
rank
SVD based Matrix Factorization
alpha
Confidence of Implicit Feedback
iteration number
ALS
MSE Converged
no: fine tune model parameters
yes
Recommendations
Recommendations (10 products per User)
Predictions (User, Item, Score)
Collaborative Filtering based Prediction Model
Evaluations
Behavioral data from next period
Event Correlations
Precision & Recall

Figure 2. General Framework

E. *Evaluation and Results*

A subtle recommender system should be empowered by customer behavior, up to date, relevant yet unforeseen and personalized. These goals are addressed with the following steps:

- A total of 174626 implicit feedback data points were generated with 39926 customers for 1403 different products.
- With respect to the question of customer taste, ratings are decomposed with latent factors.
- The CF model is optimized for predicting prospective customer-product association strength with respect to proximity and serendipity. As regularization is applied, predictions are considered as adaptive and confident.
- Predictions are generated for 1209 customers and 565 products via the CF model and tailored recommendations are prepared.

In this study, we prepared the recommendations, however, they are not displayed to the customers. In order to evaluate the recommendations, ProductView and AddtoCart events of prospective customers were collected in February, 2019 and interpreted. The Customers who showed an intention to purchase SDA products were targeted as prospective customers. The expected retention of SDA purchasers is 4-5 weeks, given the nature of the purchased product. Therefore, customers were tracked for 4 weeks. 1209 prospective customers were tracked during February, 2019. Comprehensively, 19% of our customers ended up discovering what we predicted for them and 7% of our customers showed a purchase intention to what we predicted for them. Current recommendations perform between 2% and 8% depending on the strategy (e.g. customers who bought this, category based selections, and complementary products) and position (e.g. listing pages, product detail pages, and basket). Precision represents the engagement of the tailored recommendations. This metric is the ratio of true positives over predicted positives. Predicted positive is the customers we made recommendations. True positive is the customers who interacted with the products we recommend. In this study, precision is 0.19. These recommendations are not more than predictions without the proper positioning and marketing communication. Also, prospective customers are not evaluated with a retention perspective. It is inevitable to observe that most of them did not make a secondary purchase. Recall represents the coverage of tailored recommendations over all product interactions. This metric is the ratio of true positives over actual positives. Actual positives represents the total number of customers who interacted with a product from our product spectrum. In this study, recall is 0.76. 76% of prospective customers who visited the website within 4 weeks interacted with a product from our recommendations.

V. Conclusion

A comprehensive recommendation engine for Turkey's leading e-commerce platform hepsiburada is proposed in this study. Considering the accessibility of behavioral data and sophistication of customer taste, latent factors based collaborative filtering is applied. Implicit product feedback from customers is retrieved from data. Customers and products are represented by latent factors. A prediction model is generated by solving a dynamically regularized SVD problem with ALS. The model's training parameters are fine-tuned and predefined predictions are delivered.

This framework can be enhanced with further implementations. One of them is to update the model to display to the user an explanation of the strategy behind the recommendations. The examples are 'you are seeing this because people like you purchased this product' or 'you are seeing this because you purchased that product'. To inform the customer about the reason behind the recommendations is more trustworthy and the customer can know the coherence. The other improvement opportunity is to enrich the implicit feedback model with after sales data, such as review context, return status, replenishment status. Considering the visit numbers and high assortments, millions of events are generated every day. Our case study is limited in data. However, solving the problem with ALS and weighted λ regularization is suitable for big data. This framework can be extended for larger datasets.

Recommendations are generated for the customers who at least added an item to their basket within a given day. Thus, the cold start problem is excluded and the model can be trained on larger datasets and cold starters can be tested.

Acknowledgment

This research has been financially supported by Galatasaray University Research Fund, with the project number 19.402.007.

The data used in this work is provided by hepsiburada, one of Turkey's leading e-commerce platforms. Regarding the EU general data protection regulation and Turkish personal data protection law, customer data and behavioral data was completely anonymized and processed with internal resources.

References

[1] Y. Hu, Y. Koren and C. Volinsky, "Collaborative Filtering for Implicit Feedback Datasets," Eighth IEEE International Conference on Data Mining, pp. 263-272, 2008.

[2] D. Goldberg, D. Nichols, B. M. Oki, and D. Terry, "Using Collaborative Filtering to Weave an Information Tapestry," Commun. ACM, vol. 35, no. 12, pp. 61-70, 1992.

[3] Z. Batmaz and H. Polat, "Randomization-Based Privacy-Preserving Frameworks for Collaborative Filtering," Procedia Computer Science, vol. 96, pp. 33-42, 2016.

[4] J. L. Herlocker, J. A. Konstan, L.G. Terveen, and J. T. Riedl, "Evaluating Collaborative Filtering Recommender Systems," ACM Transactions on Information Systems (TOIS), vol. 22, no.1, pp. 5-53, 2004.

[5] F. Ricci, B. Shapira, L. Rokach, and P. Kantor, "Recommender Systems Handbook,", pp.1-35, 2011.

[6] M. D. Ekstrand, J. T. Riedl, and J. A. Konstan, "Collaborative Filtering Recommender Systems," Foundations and Trends® in Human–Computer Interaction, vol. 4, no. 2, pp. 81-173, 2011.

[7] N. Polatidis and C. K. Georgiadis, "A Dynamic Multi-Level Collaborative Filtering Method for Improved Recommendations," Computer Standards & Interfaces, vol. 51, pp. 14-21, 2017.

[8] B. K. Patra, R. Launonen, V. Ollikainen, and S. Nandi, "A New Similarity Measure Using Bhattacharyya Coefficient for Collaborative Filtering in Sparse Data", Knowledge-Based Systems, vol. 81, pp. 163-177, 2015.

[9] O.Y. Kasap and M. A. Tunga, "A Polynomial Modeling Based Algorithm in Top-N Recommendation," Expert Systems with Applications, vol. 79, pp. 313-321, 2017.

[10] D. Billsus and M. Pazzani, "Learning Collaborative Information Filters," ICML, vol. 98, pp. 46-54, 1998.

[11] Y. Wang, J. Deng, J. Gao, and P. Zhang, "A Hybrid User Similarity Model for Collaborative Filtering," Information Sciences, vol. 418, pp.102-118, 2017.

[12] G. Takács and D. Tikk, "Alternating Least Squares for Personalized Ranking," Sixth ACM Conference on Recommender Systems, pp. 83-90, 2012.

[13] Y. Koren, R. Bell, and C. Volinsky, "Matrix Factorization Techniques for Recommender Systems", Computer, vol. 8, pp. 30-37, 2009.

[14] Y. Zhou, D. Wilkinson, R. Schreiber, and R. Pan, "Large-Scale Parallel Collaborative Filtering for the Netflix Prize," International Conference on Algorithmic Applications in Management, pp. 337-348, 2008.

[15] J. B. Schafer, J.A. Konstan, and J. Riedl, "E-Commerce Recommendation Applications", Data Mining and Knowledge Discovery, vol.5, no. 1-2, pp. 115-153, 2001.

[16] Meng et al., "MLlib: Machine Learning in Apache Spark", Journal of Machine Learning Research, vol.17, pp. 1-7, 2016.

Implementing Artificial Neural Network to Identify Influencers for Crowdfunding Campaigns on Twitter

Frank Yeong-Sung Lin[1], Evana Szu-Han Fang[1], Chiu-Han Hsiao[2], Hsin-Hong Lin[1]
Department of Information Management, National Taiwan University[1]
Research Center for Information Technology Innovation, Academia Sinica[2]
Taipei, Taiwan
e-mail: yeongsunglin@gmail.com, evanafang@gmail.com, chiuhanhsiao@citi.sinica.edu.tw, reckonlin7@gmail.com

***Abstract*—This study proposes an Artificial Neural Network (ANN) model for ranking potential influencers for crowdfunding campaigns on Twitter. Because influencers have a strong connection with their followers and are considered trustworthy key opinion leaders, identifying them provides opportunities for start-up companies to reach highly relevant audiences and promote their campaigns. In this study, the social authority value, a mechanism developed by Followerwonk, was employed to examine the influence strength of a Twitter user. Followerwonk is one of the most popular Twitter marketing platforms in the United States. A total of 20 influence factors of 1969 Twitter users were collected to train the ANN model. The results revealed that 13 of the 20 influence factors were significant for measuring influence strength, which improved the time efficiency of the process of evaluating potential influencers. This model can be effectively and cost- efficiently applied to support start-up companies, thus increasing the success rate of campaigns by utilizing influencer marketing.**

Keywords- social media analysis; influencer marketing; artificial neural network; sentiment analysis.

I. Introduction

Crowdfunding has flourished in the Internet age as a revolutionary means of raising capital and gaining publicity [1][2]. To increase the success rate of fundraising, determining strategies for attracting more people to contribute to fundraising campaigns is crucial for fundraisers [3]; therefore, influencer marketing on social media plays the role of “force multiplier” for crowdfunding.

Influencer marketing is a trending marketing strategy that entails companies partnering with influential individuals to relay brand messages to the individuals’ audiences [4]. In the age of social media, everyone can be an influencer [5]. An influencer can be a popular fashion photographer on Instagram, a well-known product reviewer who uses Twitter, or a respected marketing executive who frequently shares ideas on LinkedIn. Through recurrent communication, an influencer can influence a prospective consumer by providing campaign information and advice for funding decisions, thereby affecting their beliefs, motivations, attitudes, and opinions [6][7]. However, because of the high intricacy of social media characteristics and the haphazard action of influencers, identifying influencers in a limited time is difficult [8].

Different approaches have been developed for identifying influencers; however, none of such approaches have focused on crowdfunding. Therefore, the objective of this study was to identify suitable influencers who can promote crowdfunding campaigns on Twitter. This study selected Twitter, a representative microblog, because it is a suitable platform for comprehending people’s behaviors in the physical world [9]. To achieve the study objective, an Artificial Neural Network (ANN) model was used to rank the influence strength of Twitter users, and the social authority value on Followerwonk [10] was employed in the training process. A total of 20 influence factors of 1969 Twitter users were collected to train the ANN model. Furthermore, the Marketing Influential Value (MIV) model [11] was applied to classify the 20 influence factors into three primary categories.

The remainder of this paper is organized as follows: Section II reviews related work on surveying influencer identification and measurement. Definitions of 20 key influence factors for measuring the influence strength of a Twitter user are provided in Section III. Section IV details the executed experiment, and Section V presents the results. Finally, conclusions are drawn in Section VI.

II. Related Work

Over the past decade, an increasing number of studies on influencer identification has been a trending research topic. Studies have extensively applied three approaches for identifying influencers: centrality measures [12] applied with graph theory for examining the influence of a given node in a graph; prestige ranking [13] adapted for ranking influencers and inspired by the PageRank algorithm [14], which is the underlying algorithm for the Google search engine; and information diffusion [15] applied to identify the optimal path for spreading information.

A. Measurement of Influence on Blogosphere

Several studies have focused on different social media platforms, such as Facebook [16][17], Twitter [18][19], and other renowned platforms [20]. Moreover, the blogosphere is a widely used target for identifying influencers. Li et al. [11] proposed the MIV model to calculate the strength of influence and identify influential bloggers in the blogosphere. They divided the marketing influence value into three primary categories: network-based factors, referring to the explicit relationship between links or visits and social interaction (e.g., number of comments and citations); content-based factors, including the subjective degree, length, and lifetime a certain blog; and activeness-based factors, including the number of posts and replies. Under the consideration of time-stamped observations of posts and the assumption that transmission was governed by an independent cascade model, Gruhl et al. [21] attempted to construct a transmission network between bloggers. Adar and Adamic [22] used a similar approach to reconstruct diffusion trees among bloggers. Other similar approaches can be found in the next section regarding influencer identification on Twitter.

B. *Measurement of Influence on Twitter*

In earlier studies measuring influence on Twitter, the challenge was to define influence and determine the key factors of influential Twitter users. Anger and Kittl [23] compared three different measures of influence: indegree, representing the popularity of a specific user; retweets, representing the content value of a user's tweets; and mentions, representing the name value of a user. They concluded that indegree is not always related to the ability to engage an audience. This finding suggests that indegree alone reveals little information on user influence.

Kwak et al. [24] compared three different measures of influence: number of followers, page rank, and number of retweets. They observed that the rankings of most influential users differed depending on the applied measure. Similarly, Cha et al. [25] compared the number of followers, number of retweets, and number of mentions. They found that the most followed users did not score the highest on the other measures. Finally, Weng et al. [26] compared the number of followers and page rank with a modified page rank measure that accounted for topics; they also revealed that ranking depended on the influence measure. These studies have provided the foundation for future researchers; nevertheless, their results cannot be easily applied by marketing experts because of the lack of a mechanism to identify influential Twitter users. Moreover, the studies have considered a limited number of factors, which may engender a significant bias in the definition of Twitter user influence.

C. *Machine Learning for Influencer Identification*

Researchers at the Thomas J. Watson Research Center of IBM developed a supervised rank aggregation model for predicting influencers on Twitter; the model combines different influence measures to produce a composite ranking mechanism that is most effective for a desired task [27]. They compared 13 different ranking measures for identifying influencers and concluded that previous retweets were the most effective measure with the highest accuracy. Some studies have focused on analyzing factors that are crucial for increasing the influence of a Twitter user. Such studies have extracted factors from several Twitter marketing platforms: one of them is Twinfluence, which includes the velocity metric that determines the average number of first- and second-order followers [28]; TwitterGrader, which measures the number of followers and friends [28]; and Klout, which provides an influence ranking value [18]. Several regression models have been trained based on different services for evaluating factors that are crucial for increasing the influence of a Twitter user. Bakshy et al. [5] investigated the attributes and relative influence of 1.6 million Twitter users by tracking 74 million diffusion events occurring on Twitter follower graphs. They found that the largest cascades tend to be generated by users with many followers. Moreover, they observed that the most influential users are also the most cost effective, therefore, to achieve cost-effective marketing strategies, managers can increase the degree of influence of ordinary influencers, that is individuals exerting average or below average influence.

In summary, influencer identification has been prominently discussed in academia. On the basis of related work, this study can be addressed by using machine learning and deep learning techniques. These techniques can facilitate the consideration of a relatively high number of measures, which may provide new insights into marketing

III. KEY INFLUENCE FACTORS

Although studies have generally defined influencers as individuals who can have a disproportionate effect on the spread of information, this definition is ambiguous without general measurable standards. A feasible solution to the problem of defining influencers is to apply the ranking mechanism of current influencer marketing services. IZEA [29] has a quality score that ranks potential influencers on different levels from 1 to 5; however, this ranking service cannot be accessed without subscription. Followerwonk is a leading online application that provides several Twitter marketing features, one of which is the "Search Bios" tool. This tool enables users to obtain a list of Twitter users who are relevant to a search keyword. Furthermore, users can search specific Twitter profiles and obtain a summary of its influence. Followerwonk also provides the social authority value, a ranking mechanism that ranks the influence strength of a Twitter user from 0 to 100. A higher social authority value indicates a stronger influence. The score is based on three components:

- The retweet rate of a few hundred of a measured user's last non-@mention tweets [10].
- A time decay to favor recent activity versus ancient history [10].
- Other data for each that are optimized via a regression model trained to retweet rate.

Because retweets constitute a common measure of the effectiveness of a marketing campaign on Twitter, the social authority value is a reasonable reference of the ground truth data for ranking influence.

The Twitter ecosystem is suitable for studying the effect of influencers. This is because interactions between users can be observed using structured data among their tweets and profiles. To examine the degree of influence of Twitter users, this study collected 20 influence factors from profiles of Twitter users and tweets from their user timelines. These factors might exert significant or nonsignificant effects on the social authority value of users. The factors were evaluated using a Backpropagation Neural Network (BPNN). Before their evaluation, the 20 factors were classified into three categories by adjusting the present MIV model [11]: network-based, activeness-based, and content-based factors. These categories are described in the following sections.

A. Network-Based Factors

People tend to follow someone with a fine reputation, which represents their popularity and trustworthiness within a social network. Network-based factors represent the popularity and trustworthiness of a user. In a Twitter network, which comprises a user's followers and followings, tweeting is analogous to spreading seeds on a field. The more influential a user is, the higher is the likelihood that the user's seeds will sprout.

1) Popularity: To follow conversations of other communities and users, Twitter users must subscribe to such communities and users; the tweets of such communities and users would then appear on the users' own newsfeeds. Different from other social media platforms, Twitter users do not require consent to follow other users' activities. Two basic indicators represent the popularity of Twitter users: number of followers, which indicates their reputation but is not necessarily related to their influence; and number of followings (users one follows), which can indirectly increase the visibility of accounts. When users follow other users, they have a relatively high opportunity for interacting with the followed users. The higher the popularity of a Twitter user is, the higher the number of people who can access the user's tweets within a certain period. Therefore, the two aforementioned indicators must be considered:

- Number of followers: The number of followers of a Twitter user.
- Number of followings (users one follows): The number of users followed by the Twitter user.

2) Trustworthiness: Trustworthy Twitter users are responsible when sharing information on Twitter. They are reliable and honest with respect to delivering consistent values and behaviors and understand the importance of nourishing their relationship with subscribers [30]. To evaluate the trustworthiness of a Twitter user, the following factors are usually examined:

- Account age: This refers to the duration for which the account has existed. The credibility of an account can be evaluated using the account age.
- Number of statues: This refers to the number of tweets posted in the lifetime of the account. The number of statues indicates the effort of the Twitter user in managing the account. Twitter users with a high number of statues may be more trustworthy than others.
- Listed number: This refers to the number of times the Twitter account has been added to other users' favorite list in its lifetime. Twitter users can add accounts into their favorite lists. The higher the number of times the account has been listed, the higher the trustworthiness of the account is.

B. Activeness-Based Factor

Twitter is different from other social media platforms or microblogging service providers in that it can highlight some social interactions. First, most interactions occur on tweets. Second, Twitter users can repost other users' tweets to their followers, an action that is popularly known as retweeting. Finally, users can respond to other users' tweets. Users can respond to tweets on Twitter through two approaches: replying and mentioning. Replies can be indicated by tweets starting with @username, excluding retweets. A tweet that starts with @username is not broadcast to all followers but to only the corresponding user. Mentions can be indicated by tweets containing @username in the middle of its text. Such tweets are broadcast to all followers. Twitter users can "like" other users' tweets by clicking or tapping on the "favorite" button. All these interactions can be adequately tracked through the application programming interface (API) of Twitter. These interactions can be further categorized as passive and active.

1) Passive Interactions: When Twitter users tweet, they passively receive likes, retweets, and replies. Influential Twitter users can induce others to interact with them by initiating discussions and creating trending topics. The measurement of passive interactions indicates the ability of Twitter users to induce interactions.

- Most favorited: This represents the number of favorites observed on the most favorited tweet in a Twitter user's account lifetime.
- Average favorites per tweet: This represents the average number of favorites of each tweet in a Twitter user's account lifetime.
- Average favorites per user: This represents the average number of contributed favorites of each follower of a Twitter user in the user's account lifetime.
- Most retweeted: This represents the number of retweets observed on the most retweeted post in a Twitter user's account lifetime.
- Average retweets per tweet: This represents the average number of retweets of each tweet in a Twitter user's account lifetime.
- Average retweets per user: This represents the average number of contributed retweets of each follower of a Twitter user in the user's account lifetime.

An analysis that considers reply measures is comprehensive; nevertheless, an enterprise-level application of Twitter's official API is to acquire relevant objects on Twitter. Because retweeting is considered to be similar to replying, employing retweet measures is sufficient.

2) Active Interaction: In addition to passively receiving favorites, retweets, and replies from others, Twitter users can actively interact with other users to strengthen their influence by favoriting, retweeting, and replying to other users' tweets. Active interactions increase a user's probability of acquiring more followers. The higher the number of active interactions contributed by a user is, the more active the user becomes. Some of the existing active interaction measures can be outlined as follows:

- Average tweets per day: This represents the average number of tweets a Twitter user posts per day.
- Average favorites per day: This represents the average number of favorites a Twitter user contributes per day.
- Average retweets per day: This refers to the average number of retweets a Twitter user contributes per day.

These interactions describe the methods through which users use Twitter. This study was conducted to evaluate the mechanisms through which these interactions affect user influence. Notably, all factors were averaged to moderate the effects induced by the lifetime of an account and the number of followers.

C. Content-Based Factors

The role of content cannot be excluded from this study. Some content types may exhibit a stronger tendency to spread than others. Although Twitter restricts the length of tweets to less than 140 characters, it permits users to include videos, pictures, URLs, and other media on their tweets. Moreover, tweets with emotionally stimulating contents show different tendencies to spread than others [31]. Therefore, VADER [32], an open-source sentiment analysis tool, was applied to provide an averaged sentiment score for each Twitter user; this score indicates the degree to which each user's tweets are positive or negative.

1) Content Analysis: VADER is an open-source Python library for performing sentiment analysis. In VADER, sentiment classification is executed using the lexicon and rule-based sentiment analysis library; the tool performs adequately on text originating from microblogs [32]. This tool was utilized to calculate each Twitter user's average sentiment score, which was measured on a scale ranging from −4 to +4, with the midpoint 0 representing a neutral sentiment. This study considered the following content factors:

- Sentiment score: This represents the average sentiment score of each tweet on a Twitter user's timeline.
- Average length of tweets: This represents the average length of a tweet on a Twitter user's timeline.
- Average number of hashtags per tweet: This represents the average number of hashtags used in each tweet on Twitter user's timeline.

The length of a tweet and number of hashtags used in the tweet may affect the level of influence of the tweeted content. Hashtags are used to express a tweet's similarity with certain clusters of contents and can increase the exposure of the tweet to other users.

2) Types of Media: Twitter allows users to tweet with several types of media. Videos, pictures, and URLs are the most common options. Different levels of difficulty may be experienced in spreading various types of information on social media. To spread information, selecting an appropriate type of media is crucial for a Twitter user. For example, This study considered the following factors to reveal the usage habits of Twitter users and determine the effectiveness of such factors for influence measurement:

- Tweets with hashtags: This represents the number of tweets with hashtags on a Twitter user's timeline, and the corresponding ratio ranges from 0 to 1.
- Tweets with media: This represents the number of tweets containing media (videos or pictures) on a Twitter user's timeline, and the corresponding ratio ranges from 0 to 1.
- Tweets with URLs: This represents the number of tweets containing URLs on a Twitter user's timeline, and the corresponding ratio ranges from 0 to 1. Tweets with media and URLs are separated into two factors, as videos and pictures bring stronger interaction than URLs.

To identify influencers on Twitter, this study collected and processed the 20 aforementioned factors, divided into three categories, from the profiles of Twitter users; the processing results served as inputs for training the neural network model. As mentioned, the effectiveness of some of these factors in revealing the features of influencers may be significant or nonsignificant, and this is discussed in the following sections.

IV. Experiments

Numerous new projects are being implemented on Kickstarter daily, and most of them are in the top five campaign categories: games, technology, design, publishing, and arts. Entrepreneurs must identify different types of influencers during the marketing process. For example, a fundraiser who owns a campaign selling a new smartwatch product might prefer a tech influencer rather than an art influencer. Existing influencer marketing platforms usually rank Twitter users by their general influence rankings, which cannot measure their influence among different categories of campaigns.

To observe the difference between categorical influencers and general influencers, the first step is to train a neural network model, which fits the existing influencer ranking mechanism. Accordingly, this study collected the 20 aforementioned influence factors from the profiles of Twitter users who had recently tweeted about crowdfunding campaigns. These Twitter users were ranked according to the corresponding social authority value derived on Followerwonk. After the datasets were prepared, they were fed into the BPNN. Finally, the predicted social authority value and the actual value on Followerwonk were compared.

This study was considered the expandability of the model for its practicality. The ANN model was applied in this study for the following reasons: First, the proposed marketing research framework could handle more than 20 analyzed factors. Second, accessibility was considered, thus rendering the ANN model the first choice for addressing the proposed research problem for numerous existing free open-source libraries. Finally, the model can help to capture the complex nonlinear relation between this study's input factors and output results.

A. *Data Collection*

With the basic usage limitation of Twitter's API, this study collected only tweets posted within a 7-day period. To obtain a sufficient number of Twitter users, the data collection period was from May 8, 2018, to July 3, 2018, a total of 8 weeks. First, all data of live crowdfunding campaigns were crawled on Kickstarter from the top five categories [33]. These campaigns were filtered using pledged percentages. Campaigns with a pledged percentage of more than 50% were selected to moderate the effect of campaign quality (Table I). A statistical report [34] supported that 95.6% of unsuccessful campaigns did not have a pledged percentage of more than 50%; this thus indicates that unsuccessful campaigns were excluded from the analysis in the present study.

TABLE I. NUMBER OF CAMPAIGNS ABOVE 50% PLEDGED CRAWLED ON KICKSTARTER

Category	***Art***	***Design***	***Technology***	***Game***	***Publishing***
Campaigns	360	754	236	646	302

TABLE II. STATISTICS OF THE TWITTER USER DATASET

Statistics from Our Examined Blogger Set	
Number of total Twitter users	1969
Average account live time	2048.789/per user
Average number of statuses per user	21378.363/per user
Average number of followers per user	39391.371/per user
Average number of friends (followings) per user	3282.032/per user
Average listed number per user	323.231/per user
Average number of favorites per tweet	6.741/per tweet
Average number of retweets per tweet	926.577/per tweet
Average social authority value of all users	43
First quartile of social authority value of all users	32
Third quartile of social authority value of all users	56

Second, using the list of campaigns as search keywords, this study collected all tweets containing search keywords by using Twitter's API Tweepy. Tweepy is a free Python library that allows users to access Twitter and obtain the required data. From the collected tweets, this study acquired the usernames and their profile information for further evaluation. Finally, on the basis of each Twitter username, each Twitter user was ranked in terms of the social authority value on Followerwonk. The initial dataset was prepared by combining the collected data (Table II).

B. *Experimental Design*

The initial dataset was divided into a training set and testing set at a ratio of 8:2. The complete process for training the prediction model involved influence factor calculation, model training, and performance evaluation (Figure 1).

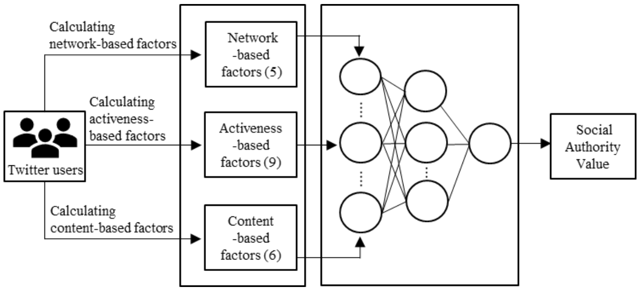


Figure 1. Model training process.

1) *Influence Factor Calculation:* The entire data of Twitter users collected in this study must be evaluated to derive the influence factors. A total of 20 influence factors, comprising the network-based factors, activeness-based factors, and content-based factors, were derived.

2) *Model Training:* In this study, an ANN with a two-hidden layer BPNN was used to address the problem of uncertainty in the weighting process. The 20 influence factors were used for social authority value prediction. The BPNN is the most widely applied ANN model because of its strength [34] of managing complex nonlinear relationships between input data and output results. The original sigmoid function is [0,1], and this study performs rescaling to get the results in the range [0, 100] to fit in the bracket of social authority value.

3) *Performance Evaluation:* To evaluate the performance of the BPNN model, the predicted social authority value and the actual value on Followerwonk were compared, and efficiency of the model was examined. A different combination of parameters of the network structure was tested to fine tune the model. The grid search algorithm was applied for finalizing the parameter settings (Table III).

TABLE III. PARAMETER SETTINGS FOR THE BPNN MODEL

Parameters	**Value**	**Parameters**	**Value**
Number of hidden layers	2	***Loss function***	MSE
Kernel initializer	Normal	***Batch size***	32
Activation function (hidden layer)	Rectifier	***Epochs***	256
Activation function (output layer)	Sigmoid	***Optimizer***	Rmsprop

C. *Structure of BPNN*

Keras on Tensorflow [35], an open-source ANN library written in Python, was applied to construct the three-layer BPNN for training and testing the proposed prediction model. The constructed three-layer BPNN was composed of an input layer, a hidden layer, and an output layer. The input layer had 20 neurons to adopt the 20 influence factors, comprising the network-based factors, content-based factors, and activeness-based factors. In the hidden layer, 10 neurons were used for adaptive weight adjustment. Only one neuron was included in the output layer for the output data, which was the predicted social authority value. Table III details the parameter settings for the BPNN model.

V. Results And Discussion

After the model training process, the model performance was evaluated to optimize the training process. The grid search algorithm was applied to determine the optimal set of parameters to train the BPNN model. A feature selection technique was then applied to determine the factors with relatively high effects and eliminate irrelevant factors. Thus, the resulting factors can help brands to identify influencers by using the minimum required amount of data, thereby improving time efficiency. Finally, the data of categorical influencers were selected. Their Twitter influence was ranked based on the social authority value predicted by the model. The difference between general influencers and categorical influencers was observed with reference to the origin of the social authority value.

A. *Results*

Several parameter sets were tested to optimize the performance of the BPNN model by using the grid search algorithm. Grid search is an approach of parameter tuning that methodically develops and evaluates a model for each combination of algorithm parameters specified in a grid. Because the output value of the BPNN model is a continuous value between 0 and 100, this research problem is naturally defined as a multiple linear regression problem. Therefore, selecting a mean squared error to score each parameter set is the most appropriate approach. The calculated mean squared error value was averaged after the application of 10-fold validation (Table IV).

TABLE IV. Results of Each Testing Parameter Set

Parameters Set	**Batch Size**	**Epochs**	**Optimizer**	**MSE**
Set #1	32	256	Adam	0.0118
Set #2	32	256	Rmsprop	0.0116
Set #3	50	512	Adam	0.0128
Set #4	50	512	Rmsprop	0.0125
Set #5	64	1024	Adam	0.0126
Set #6	64	1024	Rmsprop	0.0133

The variance of all the results was under 0.01, signifying that the model performance was acceptable without overfitting. Moreover, when identifying the influencers on Twitter, a brand is concerned with the accuracy of the model in detecting high-influence Twitter users. Therefore, the best 25 influencers were extracted from the dataset as highly influential users with a social authority value of higher than 82. The performance of the prediction model in identifying these 25 users was examined. A confusion matrix was used for accuracy evaluation, and Table V shows the results.

This study also examined different numbers of layers of the ANN model and determined that the three-layer model exhibited a relatively high performance level (Table VI). The main explanation for the derived results is that the number of highly influential users selected for evaluating the model corresponded to the extreme values of the dataset. The model was trained to fit most of the observed data, but not the extremely large data. Furthermore, a deeper network is required for training a large amount of data, particularly for unstructured data. Accordingly, the three-layer ANN model was deemed suitable for this study's analysis.

TABLE V. Accuracy of Model with High-Influence Users

Parameters Set	**Accuracy**	**True Positive Rate**	**False Positive Rate**
Set #1	94.67%	52.63%	0.53%
Set #2	94.92%	57.89%	0.27%
Set #3	95.18%	42.10%	0.00%
Set #4	95.18%	36.84%	0.00%
Set #5	94.67%	57.89%	0.53%
Set #6	95.18%	42.10%	0.00%

TABLE VI. Examining Different Number of Layers of the ANN Model

Layers	**MSE**	**Variance**	**Accuracy**	**True Positive Rate**	**False Positive Rate**
3	0.0116	0.0029	94.92%	57.89%	0.27%
4	0.0119	0.0025	91.17%	28.07%	5.63%
5	0.0116	0.0025	97.21%	10.52%	2.4%

On the basis of the model evaluation process and the aforementioned results, set#2 was selected as the optimal parameter set for constructing the BPNN model. However, performance could still be improved for highly influential users. The model was still determined to be effective for conducting further analysis.

B. *Feature Selection*

Although the model was appropriately trained and exhibited adequate performance, data collection was the most time-consuming part of this process. The collection of 20 influence factors for the targeted Twitter users required considerable time. To improve time efficiency and avoid the problem of dimensionality, feature selection techniques must be used. This study thus applied the backward elimination process to select the features that were most relevant for measuring the social authority value of Twitter users.

1) Backward Elimination: Backward elimination is a widely used feature selection technique for multiple linear regression problems. Before the initiation of the selection process, the significance level (0.05) required for features to remain in the model must be determined first. The model was fitted with all possible features, and features with the highest p value were considered. If the p value was higher than the significance level, the feature was removed from the model. After the completion of the iterations, 13 features were selected from the 20 influence factors. As shown in Table VII, all network-based factors were selected, and the content-based factors were considered relevant. By contrast, most activeness-based factors were eliminated based on the selected significance level.

TABLE VII. SUMMARY OF FEATURE SELECTION

Features	**Category**	**P-value**
Account age	Network-based factors	0.000
Number of statues		0.000
Number of followers		0.000
Number of followings (friends)		0.000
Listed number		0.003
Average favorites per day	Activeness-based factors	0.000
Average favorites per tweet		0.000
Average favorites per user		0.000
Average length of tweets	Content-based factors	0.000
Sentiment score		0.049
Tweets with hashtags		0.007
Tweets with URLs		0.000
Tweets with media		0.047

2) Eliminate Activeness-based Factors: The preceding result shows that activeness-based factors were less relevant features for measuring the social authority value. The model was retrained to improve its performance in identifying the top 25 influencers. First, the activeness-based factors were eliminated from the features, and the true-positive rate was then improved to 63.19%, without increasing the variance (Model #1). Subsequently, according to the feature selection result, the features that were not eliminated were considered. The true-positive rate was then highly improved to 89.47% (Model #2) (Table VIII).

TABLE VIII. RETRAINED MODEL PERFORMANCE

Model	**MSE**	**Variance**	**Accuracy**	**True Positive Rate**	**False Positive Rate**
#1	0.017	0.010	96.19%	63.16%	2.13%
#2	0.010	0.003	97.46%	89.47%	2.13%

C. Observing Effect of Categories

As mentioned, existing influencer marketing applications, such as Followerwonk, can rank Twitter users based on the general influence. Understating whether a Twitter user is more influential on some topics than others is difficult. To observe the effect of categories, data obtained from 40 Twitter users were selected from the dataset. These users were highlighted for focusing on a certain category of campaigns. Their data were modified and fed into the fine- tuned ANN model to compare the social authority value predicted by our model and the original value on Followerwonk.

Eliminate Tweets without URLs: To examine the impact of categories, tweets without URLs were excluded from the analysis. These tweets were crawled from the timelines of the 40 categorical users. Table IX presents tweets with URLs containing information that users intended to share. These tweets are shown to contain a “call to action.” The objective of a piece of content was to induce followers to perform a specific act. Such tweets are more important for ranking the influence of Twitter users compared with other tweets.

TABLE IX. PREDICTED VALUE OF CATEGORICAL USERS (PARTIAL)

User	**Category**	**Social Authority**	**Predicted Value**
18dMedia	Design	25	64.32
5toclose		33	57.93
bikeradar		66	69.96
designtaxi		76	84.48
gadgetfeedco		39	57.37
werdcom		31	51.24
Ellerium_Games	Games	26	33.33
ETBoard_Games		55	61.09
ssoebmizan		38	47.65
tgn_news		47	57.40
NewsWatchTV	Technology	51	57.14

After predicting the modified data of the 40 categorical users, this study observed that the predicted social authority value of some categorical users was increased (Figure 2).

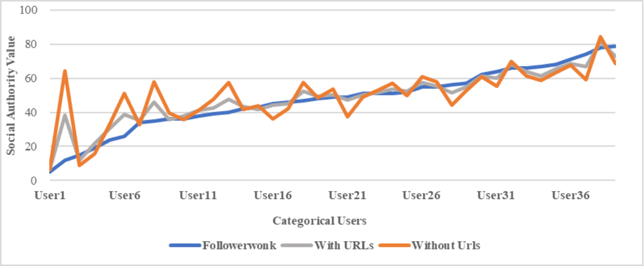


Figure 2. The predicted value of categorical users.

Although the study cannot attribute the differences to the effect of categories, the predicted social authority value of some users increased under the condition where tweets without URLs were excluded.

VI. Conclusion

This study proposes a research framework, involving data collection, influence factor calculation, and ANN model training, for identifying potential influencers for crowdfunding on Twitter. The processes involved in the framework can be easily applied through open-source libraries without incurring any costs. The MIV model and feature selection technique were applied in this study to identify the optimal measures of the influence strength of a Twitter user. Thirteen factors were selected from a total of 20 influence factors. Activeness-based factors were determined to be the least relevant features for measuring influence. These results may be explained by the fact that the quality of tweets by influencers could be inversely proportional to the number of tweets posted. These findings can improve time efficiency for companies in the execution of marketing research. After observing the effects of different categories, this study determined that the social authority value of some of the categorical users increased after the exclusion of tweets without URLs from the analysis.

Future research directions include the development of a fair ranking mechanism because predicted values are limited by the original social authority value, which may be inaccurate in case of promotional activities. Furthermore, this study suggests that future studies monitor changes in crowdfunding campaigns to determine and measure the actual effects of potential influencers. Finally, it would be difficult to conclude that the observed differences between the predicted value and original social authority value were the result of categorical effects. Therefore, this study highly recommends the implementation of a posterior examination framework in future research.

References

[1] P. F. Yu et al., “Prediction of crowdfunding project success with deep learning,” Proc. IEEE 15th International Conference on e-Business Engineering (ICEBE 2018), IEEE Press, Oct. 2018, pp. 1–8.

[2] X. Ren et al., “Tracking and forecasting dynamics in crowdfunding: A basis-synthesis approach,” Proc. IEEE International Conference on Data Mining (ICDM 2018), IEEE, Nov. 2018, pp. 1212–1217.

[3] G. U. T. Sahid, I. Putrì, I. S. Septiana, and R. Mahendra, “Estimating the collected funding amount of the social project campaigns in a crowdfunding platform,” Proc. International Conference on Advanced Computer Science and Information Systems (ICACSIS 2017), IEEE, Oct. 2017, pp. 277–282.

[4] P. Lagrée, O. Cappé, B. Cautis, and S. Maniu, “Algorithms for online influencer marketing,” ACM Transactions on Knowledge Discovery from Data, vol. 13, no. 1, 2019, pp. 3:1–3:30.

[5] E. Bakshy, J. M. Hofman, W. A. Mason, and D. J. Watts, “Everyone’s an influencer: Quantifying influence on twitter,” Proc. The 4th ACM international conference on Web search and data mining (WSDM 2011), ACM, Feb. 2011, pp. 65–74.

[6] J. Villanueva, S. Yoo, and D. M. Hanssens, “The impact of marketing-induced versus word-of-mouth customer acquisition on customer equity growth,” Journal of Marketing Research, vol. 45, no. 1, 2008, pp. 48–59.

[7] T. W. Valente and P. Pumpuang, “Identifying opinion leaders to promote behavior change,” Health Education & Behavior, vol. 34, no. 6, 2007, pp. 881–896.

[8] H. Cao, J. Wang, and Z. Wang, “Opinion leaders discovery in social networking site based on the theory of propagation probability,” Proc. 2nd IEEE Advanced Information Management, Communicates, Electronic and Automation Control Conference (IMCEC 2018), IEEE, May, 2018, pp. 700–704.

[9] L. Tang and Z. Ni, “Emerging opinion leaders in crowd unfollow crisis: a case study of mobile brands in Twitter,” Pattern Analysis and Applications, vol. 19, no. 3, 2016, pp. 731–743.

[10] Followerwonk, https://followerwonk.com/ [retrieved: Jun. 2019]

[11] Y. M. Li, C. Y. Lai, and C. W. Chen, “Discovering influencers for marketing in the blogosphere,” Information Sciences, vol. 181, no. 23, 2011, pp. 5143–5157.

[12] L. C. Freeman, “Centrality in social networks conceptual clarification,” Social Networks, vol. 1, no. 3, 1978–1979, pp. 215–239.

[13] S. Wasserman and K. Faust, Social Network Analysis. Cambridge, NY: Cambridge University Press, 1994.

[14] L. Page, S. Brin, R. Motwani, and T. Winograd, “The PageRank citation ranking: Bringing order to the web,” Technical Report, Stanford InfoLab, 1999.

[15] E. Bakshy, I. Rosenn, C. Marlow, and L. Adamic, “The role of social networks in information diffusion,” Proc. The 21st international conference on World Wide Web (WWW 2012), ACM, Apr. 2012, pp. 519–528.

[16] M. Nouh and J. R. C. Nurse, “Identifying key-players in online activist groups on the Facebook social network,” Proc. IEEE International Conference on Data Mining Workshop (ICDMW 2015), IEEE, Nov. 2015, pp. 969–978.

[17] J. Heidemann, M. Klier, and F. Probst, “Identifying key users in online social networks: A PageRank based approach,” Proc. The 31st International Conference of Information Systems (ICIS 2010), AIS, Dec. 2010, pp. 79.

[18] G. S. Mahalakshmi, K. Koquilamballe, and S. Sendhilkumar, “Influential detection in Twitter using Tweet quality analysis,” Proc. 2nd International Conference on Recent Trends and Challenges in Computational Models (ICRTCCM 2017), IEEE, Feb. 2017, pp. 315–319.

[19] J. Torres, G. Baquerizo, C. Vaca, and E. Peláez,

"Characterizing influential leaders of Ecuador on Twitter using computational intelligence," Proc. 3rd International Conference on eDemocracy & eGovernment (ICEDEG 2016), IEEE, Mar. 2016, pp. 159–163.

[20] X. Lin and W. Han, "Opinion leaders discovering in social networks based on complex network and DBSCAN cluster," Proc. 14th International Symposium on Distributed Computing and Applications for Business Engineering and Science (DCABES 2015), IEEE, Aug. 2015, pp. 292–295.

[21] D. Grugl, R. Guha, D. Liben-Nowell, and A. Tomkins, "Information diffusion through blogspace," Proc. The 13th international conference on World Wide Web (WWW 2004), ACM, May, 2004, pp. 491–501.

[22] E. Adar and L. A. Adamic, "Tracking information epidemics in blogspace," Proc. The 2005 IEEE/WIC/ACM International Conference on Web Intelligence (WI 2005), IEEE, Sep. 2005, pp. 207–214.

[23] I. Anger and C. Kittl, "Measuring influence on Twitter," Proc. The 11th International Conference on Knowledge Management and Knowledge Technologies (i-KNOW 2011), ACM, Sep. 2011, article 31.

[24] H. Kwak, C. Lee, H. Park, and S. Moon, "What is Twitter, a social network or a news media?" Proc. The 19th international conference on World wide web (WWW 2010), ACM, Apr. 2010, pp. 592–600.

[25] M. Cha, H. Haddadi, F. Benevenuto, and K. P. Gummadi, "Measuring user influence in Twitter: The million follower fallacy," Proc. the Fourth International Conference on Weblogs and Social Media (ICWSM 2010), AAAI, Jun. 2010, pp. 10–17.

[26] J. W, E. P. Lim, J. Jiang, and Q. He, "Twitterrank: Finding topic-sensitive influential Twitterers," Proc. the 3rd International Conference on Web Search and Data Mining (WSDM 2010), ACM, Feb, 2010, pp. 261–270.

[27] K. Subbian and P. Melville, "Supervised rank aggregation for predicting influencers in Twitter," Proc. Third International Conference on Privacy, Security, Risk and Trust and Third International Conference on Social Computing (PASSAT and SocialCom 2011), IEEE, Oct, 2011, pp. 661–665.

[28] Y. Blanco-Fernández, M. Lopez-Nores, J. J. Pazos-Arias, and M. I. Martín-Vicente, "Spreading influence values over weighted relationships among users of several social networks," Proc. International Conference on Pervasive Computing and Communications Workshops (PerCom 2012), IEEE, Mar. 2012, pp. 149–154.

[29] IZEA, https://izea.com/ [retrieved: Jun. 2019]

[30] L. Li, C. Chen, W. Huang, K. Xie, and F. Cai, "Explore the effects of opinion leader's characteristics and information on consumer's purchase intention: Weibo case," Proc. 15th International Conference on Service Systems and Service Management (ICSSSM 2018), IEEE, Jul. 2018, pp. 1–6.

[31] S. Stieglitz and L. Dang-Xuan, "Emotions and information diffusion in social media—Sentiment of microblogs and sharing behavior," Journal of Management Information Systems, vol. 29, no. 4, 2013, pp. 217–248.

[32] C. J. Hutto and E. Gilbert, "VADER: A parsimonious rule- based model for sentiment analysis of social media text," Proc. International Conference on Weblogs and Social Media (ICWSM 2014), AAAI, May. 2014, pp. 216–225.

[33] Statista. "Distribution of unsuccessfully funded projects on crowdfunding platform Kickstarter as of April 2018, by share of funding reached," https://www.statista.com/statistics/251732/overview-of-unsuccessfully-funded-projects-on-crowdfunding-platform- kickstarter/ [retrieved: Jun. 2019]

[34] J. Devillers, "Strengths and weaknesses of the Backpropagation Neural Network in QSAR and QSPR studies," Neural Networks in QSAR and Drug Design. San Diego, CA: Academic Press, 1996.

[35] Keras, https://www.tensorflow.org/guide/keras [retrieved: Jun. 2019]

An Eye Tracking Study of the Visual Behavior of Children in Social Interaction

Emad Bataineh
College of Technological Innovation
Zayed University
Dubai, UAE
e-mail: Emad.Bataineh@zu.ac.ae

Basel Almourad
College of Technological Innovation
Zayed University
Dubai, UAE
e-mail: Basel.Almourad@zu.ac.ae

***Abstract*— The paper presents an eye tracking analysis study to help us understand the visual behavior and pattern of normal developing children and autistic children while viewing a socially rich stimulus consisting of human and social interactions, as well as the factors that influence their behavior. Eye tracking is a technology that allows the assessment of one's spontaneous visual attention and eye gaze preference and pattern. An eye tracking experiment consists of displaying different images with social stimuli (containing human faces) to the child. The eye tracker captures and tracks the child's eye gaze movements, then analyzes the data to identify where specifically in the stimulus is the child looking at. Sixty-four participants (normal and autistic) were divided into two groups. The participants were asked to view a socially rich information stimulus for a limited and set time. Based on the data analysis conducted in the study, the findings show a significant difference between the two groups viewing patterns and behavior when the subjects were presented with a scene included material with human and social interaction content. The study also reveals that a large percentage of autistic participants expressed minimum interest and time looking at the face area, evident by a significant time spent fixating on non-face regions. This is linked to a lack of interest in socially relevant information, especially the two small areas of interest which are the eyes and the mouth regions, when compared to the normal developing children. The results can be used to help improve the life style of other children who have a potential to develop autism as well as discover earlier signs of autism spectrum disorder.**

Keywords- Autism Spectrum Disorder; Eye Tracking; Socially relevant information; Visual behavior.

I. INTRODUCTION

Autism Spectrum Disorder (ASD) is a neurodevelopmental condition defined by impairments in reciprocal social-emotional interaction and non-verbal communication, alongside with restrictive/repetitive patterns of behavior [1]. Research shows that early diagnosis of ASD enhances the possibility of improving psychosocial functioning in the following developmental years. Furthermore, recent experiments have proved that eye movements and reactions to verbal/visual cues can be used to identify signs of ASD, through an eye tracking software, allowing an early diagnosis.

An eye-tracker is a software system that allows the assessment of one's spontaneous attention and eye gaze preference [2]. In other words, it tracks and captures what the users are looking at. One of the most common signs of ASD is the specific patterns of eye movements and reactions to verbal and non-verbal cues. That is why, in the recent years, many medical researchers turned to eye tracking techniques, as they have the potential to characterize ASD non-invasively and does not require advanced motor responses or language when studying young children and infants [3]-[8].

This research aims to utilize the eye tracking technology combining different stimuli (pictures, films, interpersonal and social interactions) and enlarging computational and analytical capabilities in terms of results – i.e. gaze readings and plots, heat maps, bee swarm. In collaboration with Dubai Autism Centre and other UAE national autism related centers, we have used eye tracking technology in an experimental group and in a control group, aiming to significantly contribute to early ASD diagnosis in UAE and worldwide. The findings may be used to help improve the life of other children who have a potential to develop autism. Furthermore, the research will aid in an earlier diagnosis through an experiment that involves analysis of a child's gaze with an eye-tracker system. The paper is organized as follows: Section 1 presents a broad overview of previous research studies related to ASD and eye tracking technology; Section 2 describes the research motivation and objectives, Section 3 outlines the research methodology; Section 4 presents the data analysis and results; Section 5 discusses the results and findings; and Section 6 presents the conclusion and future work.

II. RELATED WORK

ASD is a neurodevelopmental condition characterized by impaired social interaction, problems with verbal and non-verbal communication, unusual, repetitive, or severely limited activities and interests [9]. The number of children being diagnosed with ASD in the US has risen by 23% since 2009, with one in 88 children affected, according to a report from the US Centre for Disease Control and Prevention [9].

The number of children with ASD in the UK has also risen by twelve-fold in the past 30 years and may be 50% higher than previously suspected, according to a report by the Autism Research Centre at Cambridge University [2]. In the UAE, the head of the community service unit, Dubai Autism Centre [10][11] reported that the UAE is heading in the same direction and autism is on the rise. She also

ISBN: 978-1-61208-728-3

mentioned that Dubai and Abu-Dhabi centers are not coping with the large numbers and several hundreds of patients are on waiting lists. Symptoms are present since early childhood, but the complex nature of this disorder, coupled with a lack of biologic markers for diagnosis and changes in clinical definitions over time, create challenges in monitoring the prevalence of ASD. Furthermore, the evaluation of symptoms involves a multi-disciplinary team of doctors including a pediatrician, a psychologist, a speech and language pathologist, and an occupational therapist.

Fortunately, it is widely reported that early detection is essential for early treatment and symptoms' control and that is why, in recent years, a large number of medical researchers are investigating early symptoms of autism [12][13]. However, most of the research undertaken by medical researchers were not comprehensive, considering only a few aspects of the problem and did not go into deep computational analysis of the data and videotapes recorded on the behavior of early aged children who were later diagnosed with autism. Eye tracking research plays a key role in understanding how individuals view and perceive the world around them. The scientific study of human eye movements provides an insight into the cognitive thought processes and has been established in research domains such as developmental psychology, psycholinguistics, reading research and HCI [14].

Eye tracking is an advanced technology that uses high precision to measure exactly where a user is looking and for how long. It is used to study the relationships between eye movement data and cognitive activity of the user. Frequency of fixations and duration of fixations are two important factors used to decide different aspects of the quality of screen contents. Eye movements are tracked and classified using various significant indicators of ocular behaviors, namely fixations, saccades, pupil dilation, and scan paths [9]. Eye fixations are considered the most relevant indicator for evaluating information acquisition and processing in online search and visualization environment [15]. Fixations are defined as spatially stable gaze which last for approximately 200-300 milliseconds during which visual attention is directed to a specific area of content display [16].

III. RESEARCH OBJECTIVES

Eye tracking technology has been used in many autism studies [4]-[8]. The research in [8] found no differences between gaze behaviors of children with autism and their age and IQ-matched typically developing peers when viewing cartoon like scenes that include a human figure. The research in [4]-[6] contradicted with [7] result as both research studies reported that individuals with autism fixated less on the eye region and more on the mouth, body, and object regions than individuals in the comparison group.

The aim of our research is to exploit the eye tracking technology by analyzing and comparing autistic and normal developing children gaze patterns while viewing a human and social interaction rich material. We have used and combined different stimuli with four still images, each one involving a human and social interaction situation including children and adults, both male and female. The eye tracking technology has significantly improved over the last decade. Our research subjects are based in the Gulf Area and they are a mixture of Emirati and expatriate children. The purpose of this paper is to present the initial study findings and compare it with the previous studies. We have used four different social scenes (each has socially relevant information and human interaction). The long term objective of the project is to design and develop an eye-tracker-based computing model that aims to contribute to early ASD diagnosis. The immediate aim of this research is to address the following research questions:

- How do autistic children (AC) and normal developing children (NDC) eye movement visual patterns differ when processing human faces in a socially rich context?
- What is the first area of the face for the eye gaze fixation of AC and NDC that grab their attention when presented with human and social interaction image situation?
- How fast (fixation time) NDC and AC groups spot the face region?
- Are areas of interest in fixating gaze different between NDC and AC (eye, mouth, ears, body, and off zones)?
- Within the AC group, what are the preferred areas of fixation on a face/person? (duration time).

IV. STUDY DESIGN AND METHODOLOGY

A. Participants

Sixty-five children (34 autistic children and 31 normal developing children) participated in this study. The average age of participants was eight years old (max=16 and min = 4). The gender representation is (73%) females and (27%) males. The participants come from different groups, Expatriates (60%) and Emirati nationals (40%) and come from families with different ethnic background and various social and economic classes.

ISBN: 978-1-61208-728-3

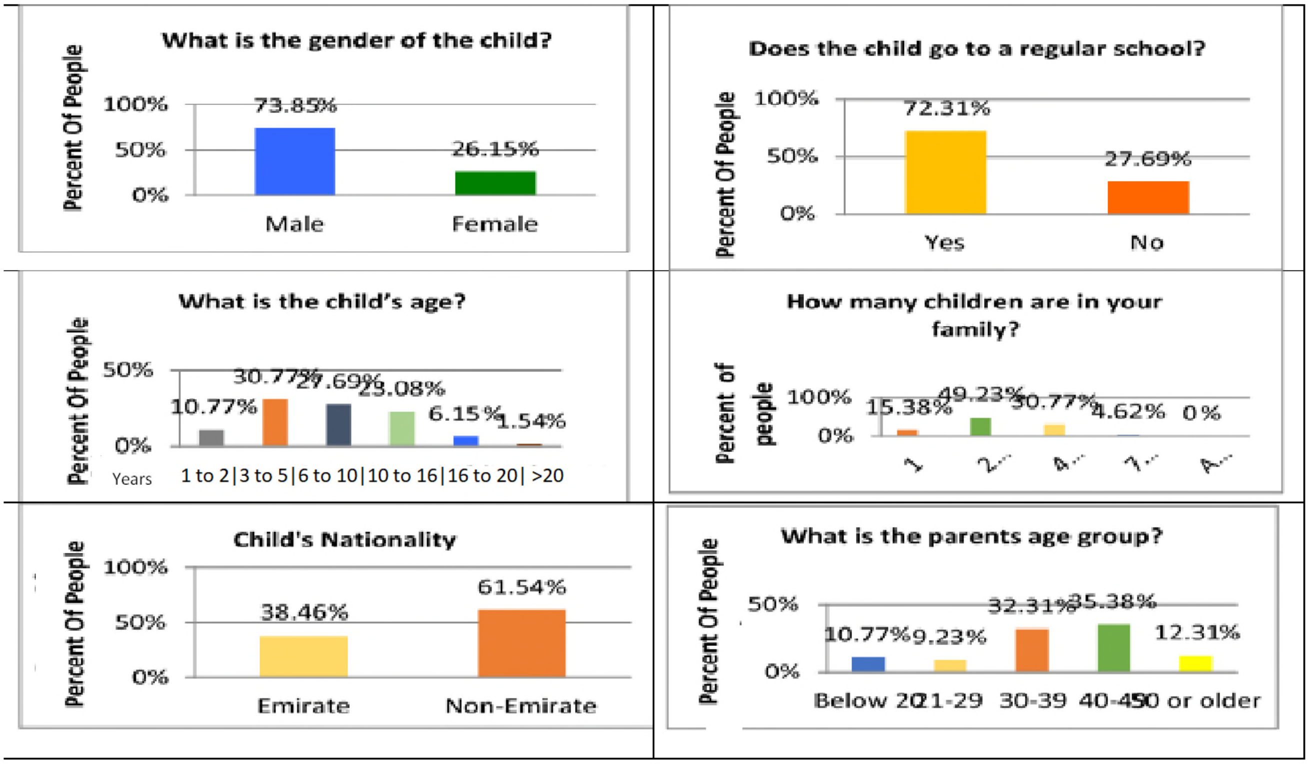


Figure 1. Participants' demographic information and background.

In terms of schooling system, 72% go to a regular school (inclusive education system) and (28%) go to special needs schools. Most participants (80%) come from small to mid-size families, almost (70%) of children parents' age between 30 to 50 years old. Participants were selected from three different Emirates: Dubai, Sharjah and Abu Dhabi. All participants' parents signed a consent form stating the purpose of the study, risks involved, confidentiality and their rights as study participants. See Figure 1 for the details of the participants' demographic and background information.

B. Materials and Tasks

An eye tracking experiment was designed using a Tobii studio environment which consists of a stimulus with four still images each one involving a human and social interaction situations including children and adults and both male and female, as represented in Figure 2. The focus of this research was to study and analyze the eye gaze pattern and visual behavior of autistic and normal developing children while viewing these images. On each image, we created three areas of interest (AOIs) namely: full human face, mouth and eyes, which are the focus of human, and social interactions.

C. Procedure

The experiment started with the moderator explaining the main purpose of the study. An unobtrusive eye-tracking mobile technology and system was used in the experiment, mobile Tobi X2 model, to collect the eye movement data of the children participants. The mobile eye tracker provides the researchers with the flexibility to set up the experiment anywhere off campus, such as homes, schools, and special needs centers to meet the situational needs of the children participants. The eye tracker was calibrated for each participant by the test moderator before the test and collected data covering all the tasks during the experiment, as demonstrated by Figure 3. The experiment consisted of three sections. The three sections of the experiment and type of data collected from each section are outlined and described in Table I. Due to space limitations, this paper will discuss the results from Test 1 only.

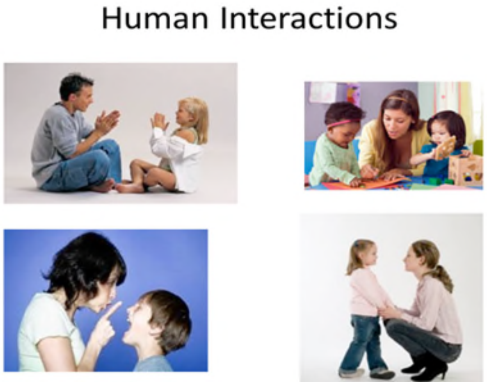


Figure 2. A stimulus involves human and social interaction.

TABLE I. COLLECTION OF TESTS USED IN THE STUDY

Sections	Instrument used for data collection
Test 1	Eye tracking test to collect quantitative data while viewing tasks in a sequence of static colored stimuli created by authors
Test 2	Eye tracking test to collect quantitative data while viewing a sequence of 28 static black and white stimuli used in a previous research study in UK
Test 3	Eye tracking test to collect quantitative data while viewing a sequence of static black and white dynamic stimuli used in a previous research study in the US. Each stimulus expressed a different emotion, followed by a question to determine the type of emotion presented in the motion stimulus
Post -test Questionnaire	To collect qualitative data on the participants' demographics, social status and their physical conditions

Studio software version 3.6 was used to record all eye tracking data for later data visualization and data analysis. This experiment requires the participants to look at a picture (face, banana, football and tomato). During this time, the eye tracker will trace the child's eye gaze, analyzing where specifically the child looks at in the picture. As so, the child only needs to be seated at the computer looking at the different pictures (and can be accompanied by a parent if necessary). The length of the experiment will take approximately between 20 to 30 minutes. For each test, the participant must also complete an eye tracking calibration, so a total of 3 calibrations per experiment/participant are needed.

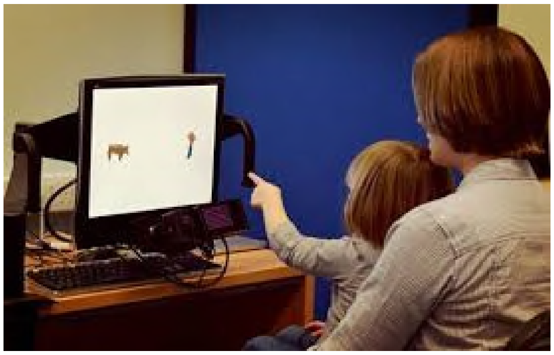


Figure 3. Experiment environment.

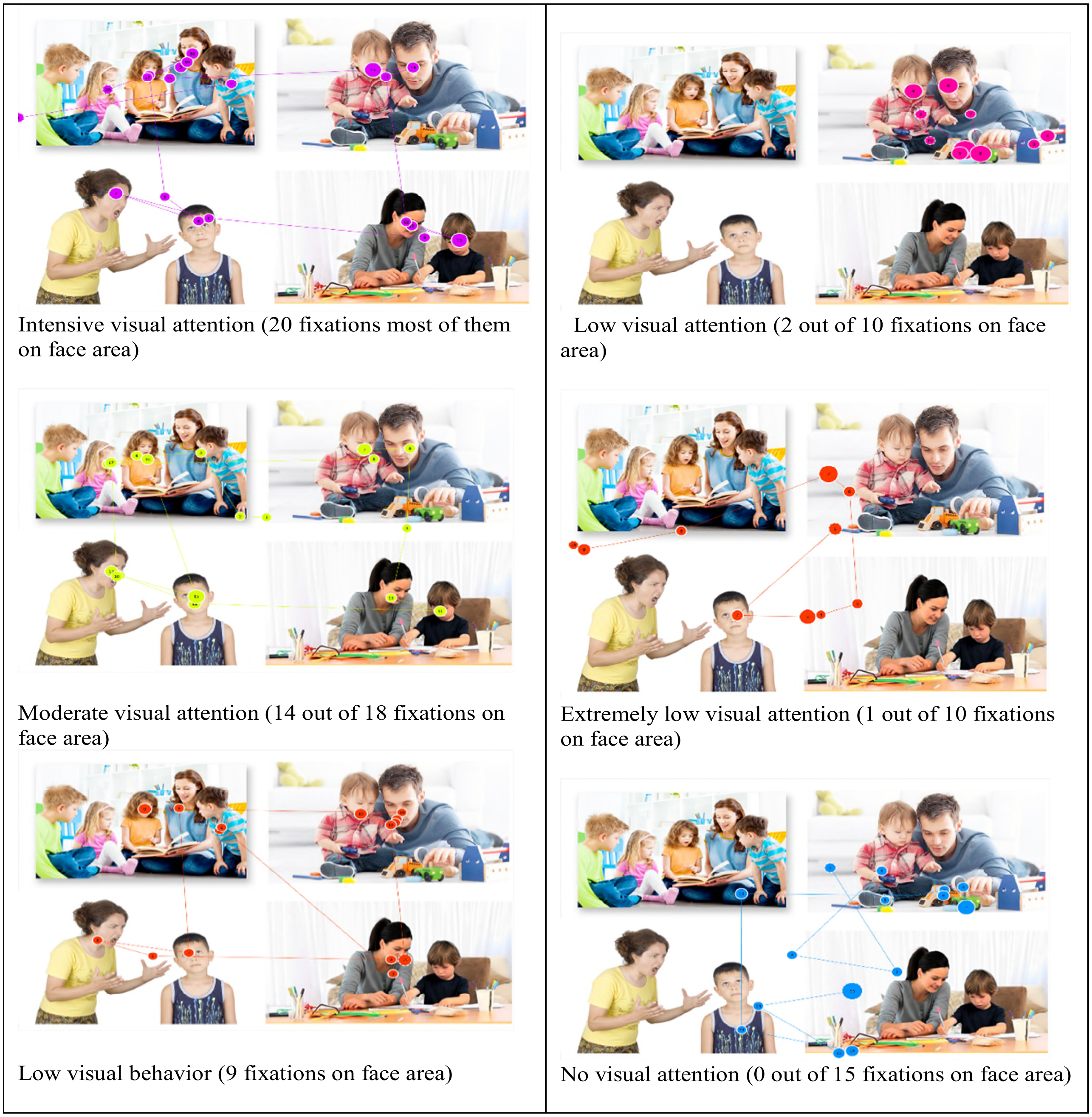
The experiment was conducted in a controlled environment as throughout the study. The participants had control of their environment, allowing them to click and to continue to the next stimulus. Once the test session was completed, the participants were thanked for their participation. At the end of the 3rd test, the parents of the participants were asked to answer 12 demographic, social, financial, educational level, and physical health status questions related to them and their child (questions are not discussed in this paper). For convenience, the questions were presented on the screen in both languages Arabic and English, whereby parents choose their preferred language. The experiment was conducted and eye tracking data were collect between November 2016 and September 2016.

V. DATA ANALYSIS AND DISCUSSION

In this section, we analyze and discuss the visual behavior of autistic and normal developing children when the stimulus contains four still images each one involving a human and social interaction situation including children and adults and both males and females (see Fig 1.). We will discuss and analyze the gaze behavior when the stimulus contained all images. We will then discuss the gaze behavior when the stimulus contained the human's face only. Within the human's face, we have defined three areas of interest (AOI) which includes the full face, face-eyes and face-mouth. Various metrics are used to analyze the difference in the behaviors between the two groups. As mentioned before, we refer to Autistic Participants as AP and Normal Developing Participants as NDC. The collected eye tracking data were analyzed using three data analysis and visualization tools to understand and describe the children's visual and eye gaze behavior.

A. Eye Gaze Analysis

The analysis of eye gaze behavior has been used in this research. Gaze plots showing location, order, time spent looking (fixation time) at different areas (zones) on the stimulus are mostly used for a single eye tracking participant looking at a fixed and dynamic content. It contains a sequence of numbered circles, each one representing a point that the children's eyes fixated on; the larger the circle, the longer the fixation. The numbers represent the order in which children look at various items (areas) on the stimulus that involves human and social interaction used to study. Social interaction stimulus consists of mainly human faces, marked by the two most important areas, namely mouth and eyes. Eye gaze indicates the level of visual activity taking place on certain areas, especially the face area which is the focus of the human and social interaction. It is assumed that more visual activity means higher interest and less visual activity indicates less interest in the content.



Intensive visual attention (20 fixations most of them on face area)

Moderate visual attention (14 out of 18 fixations on face area)

Low visual behavior (9 fixations on face area)

Low visual attention (2 out of 10 fixations on face area)

Extremely low visual attention (1 out of 10 fixations on face area)

No visual attention (0 out of 15 fixations on face area)

Figures 4a and 4b. Eye gaze behavior for normal developing children group vs. autistic children group.

An interesting result has been revealed from the scenes that involve human and social interactions between two or more individuals, including children. It was clear that more visual activity and eye gaze fixations are on areas that are relevant to human interaction, such as the face represented by the eyes and mouths. Normal developing children expressed very high interest in looking and spending more fixation time at the human face, which is the core element for human social interaction. Even within the normal group, there were different levels of fixation intensity on the face area. It ranged from highly intensive fixation, which demonstrated strong interest, to minimum fixation, which indicated low interest in social interaction, as depicted in Figures 4a and 4b. On the other hand, Autistic children have expressed minimal interest in looking at the human face as a result have much lower fixations with extremely no fixations, as indicated by the Figures 4a and 4b. The interest was measured in how many fixation points resulted from the eye gaze behavior spent looking at various images involving human and social interaction. It is worth noting that this pattern of visual attention is evident in all four images regardless of ages and genders of the persons in the picture. The results indicate that autistic children avoid looking at the human faces, including mouth and eyes.

B. *Heat Maps Analysis*

Heat maps show how the viewing and looking is distributed over the given stimulus. Heat maps are a visualization tool that can effectively reveal the focus (hot spots) of visual attention and viewing behavior for a group of participants. Children's visual attention to socially relevant information stimulus can be represented with heat maps. Four heat maps are presented for each group, see Figures 5a and 5b. Figures 5a and 5b show the heatmap analysis with the hot spot and fixation areas and patterns on the social interaction stimulus for the Normal Developing group and Autistic group. There are 11 different human faces presented on the social stimulus, which include 3 girl faces, 4 boy faces, one man face and 3 women faces.

The heat maps show that the two boy faces looking sideways did not receive any visual attention from both groups. A close look at the heat maps in Figure 5a shows what areas of the stimulus attracted children's attention during the entire time that they viewed the content. The results in Figure 5a reveal that the Normal Developing Children group was more visually engaged (they tended to spend more time looking at the eyes) with the social content, especially at the human faces compared to the Autistic group. The finding indicates that the majority of ND children focused their visual attention on the eyes of the human face. Figure 5b exhibits more fixation activities on areas away from the faces. Autistic children tended to spend more time looking at areas other than the eye regions. This visual behavior demonstrate that autistic children lack interest in looking at socially relevant information represented by the human eyes, evident by the time spent fixating mostly on the mouth region.

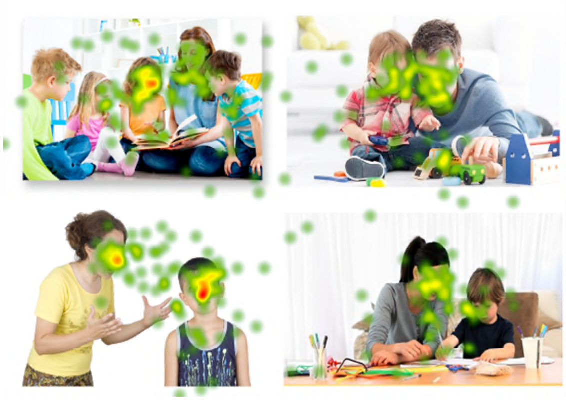


Figure 5a. Heat maps of visual attention for the NDC group.

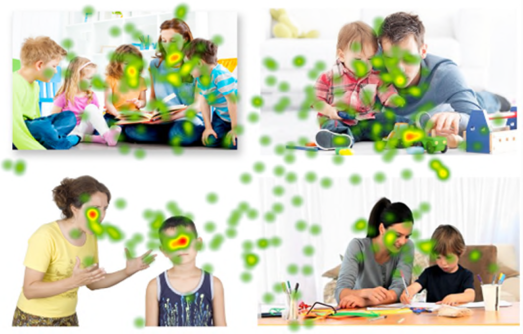


Figure 5b. Heat maps of visual attention for AC group.

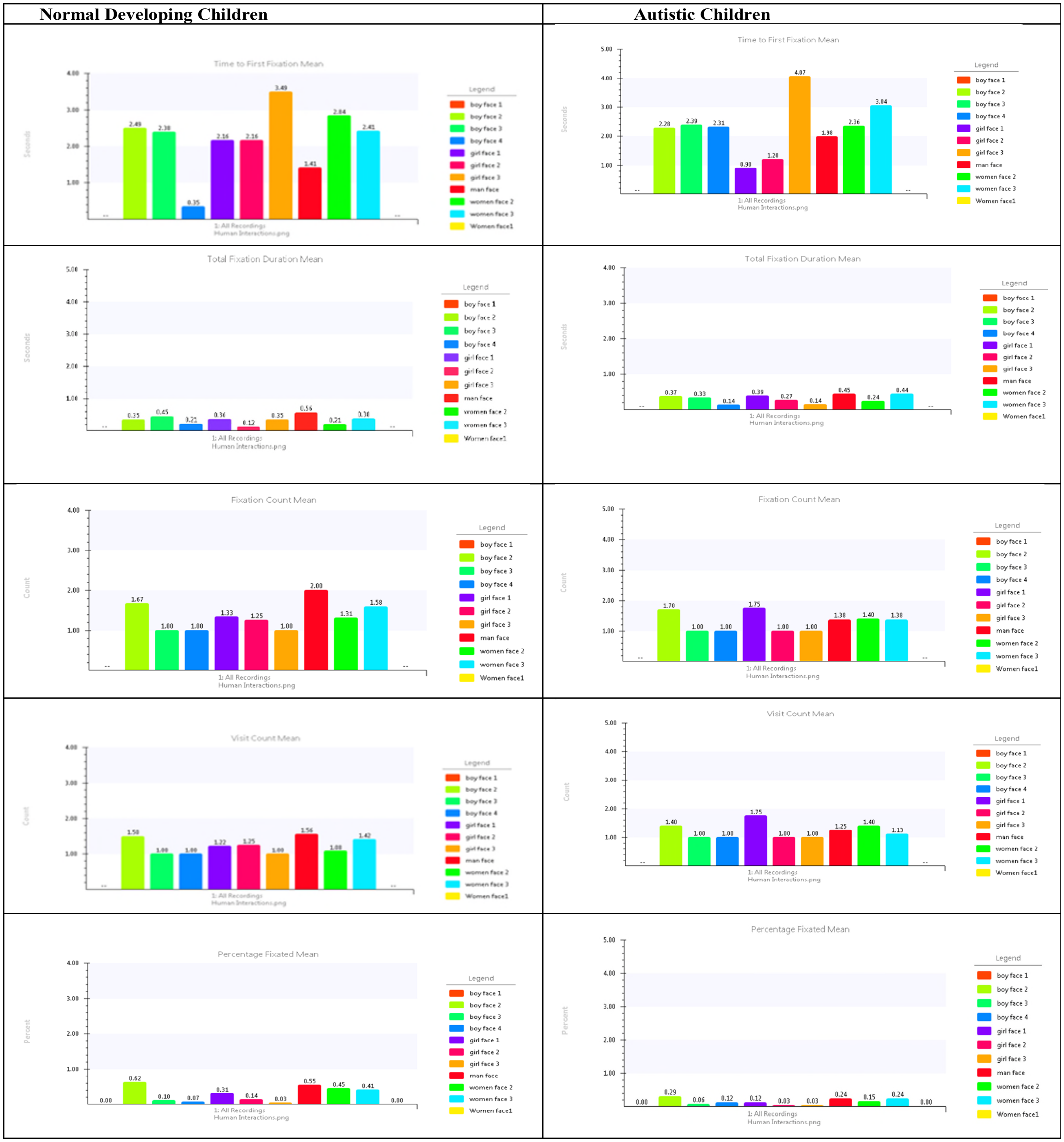
C. *Descriptive Statistical Data Analysis*

Eye movement analysis was conducted in the study to analyze the children's visual attention. The average total viewing time (fixation duration) and the average number of fixations in each social image were calculated and compared across the two groups. Statistical analysis tools using AOIs and various metrics data, like time for first fixation, no. of fixations, fixation duration on specific content area, no. of visits, etc. were used to measure the child dwelling time and decision-making behavior activities. To quantify visual fixations, four regions of interest (eye, mouth, face and non-face) were created. The areas of interest (AOIs) were defined on the stimulus and include all human faces for all people shown on the human and social interaction image. For this research, we use five different metrics. The types of metrics and their meaning and description are outlined in Table II.

A minimal difference was found in the time taken by the AC and NDC groups to have their first fixation on all human face regions. The average mean time for the first fixation looking at any of the faces, all combined for the AC group

TABLE II. LIST OF METRICS USED IN THE STATISTICAL ANALYSIS.

Metric name	**Meaning/Description**
Time to First Fixation	This metric measure how long it takes before a participant fixates on an AOI or object for the first time.
Total Fixation Duration Mean	This metric measures the sum of the duration for all fixations within an AOI or object
Visit Counts Mean	A visit is defined as the time interval between the first fixation on the active AOI and the end of the last fixation within the same active AOI (Area of Interest)
Total Visits Duration Mean	Total visit duration is defined as the sum of visit durations of an active AOI
Percentage Fixated Mean	This metric measures the number of recordings in which participants have fixated at least once within the AOI and expresses it as a fraction of the total number of recordings



Figures 6a and 6b. Metrics analysis for normal developing children group vs. autistic children group.

was 2.28 sec whilst it was 2.15 sec for the NDC group. It indicates that, on average, there was an 0.13 second delay in the first fixation by the AP group. Another interesting observation, the first face among the 11 faces that grabbed the attention of the NDC group was a boy face with 0.35 sec. mean time to first fixation, while the AP group got fixated first at a girl face with average of first fixation time 0.90 sec (see Figures 6a and 6b). It means that there are more boys in the NDC group than the AP and vice versa. Another area of minimal difference is that the NDC group expressed slightly more interest in looking at all faces combined than the AC group, as demonstrated by the mean total fixation duration time 0.34 sec vs. 0.30 sec (Figures 6a and 6b).

In addition, the NDC group made on average more fixations on human faces compared to the AC group. The average fixation counts were 1.34 fixations for the NDC group vs. 1.28 for the AC group (see Figures 5a and 5b). These results support the fact that NDC children expressed more interest in looking at human faces than the AC children, within a social context. The number of visits to human face AOIs were also analyzed. The results show no significant difference in the average of visits counts on all faces for both groups, with an average of 1.22 visits for NDC and 1.21 visits for AC children. The percentage of participants of both groups who had at least one fixation within the faces regions were also analyzed. A significant difference was noticed between the two groups. The results show that 30% of the NDC children got fixated at least once on one of the human faces comparing to only 14% of the AC children (see Figures 5a and 5b). This indicates that even though there is a small difference in average fixations duration on human faces combined between the two groups, the interest among the NDC group is almost twice that of the AC group. It also means that a large percentage of the AC children were not attracted and did not pay any attention to the human faces in the social stimulus, and that is linked to a lack of interest in socially relevant information, especially the face and, more specifically, the eyes.

V. CONCLUSION

The paper presents a research analysis study using eye tracking technology in ASD diagnosis which, is the first of its kind in the region. The aim of the study was to help us understand the visual behavior and pattern of Normal Developing children and Autistic children while viewing a static social stimulus consisting of human and social interaction. Eye tracking was employed to record children’s visual attention. An eye tracking experiment was designed and conducted using social interaction content (human faces) to collect an eye tracking data on children’s visual behavior. Sixty-four participants (normal and autistic) were divided into two groups. The participants were asked to view a static stimulus for a limited and set time.

Based on the data analysis conducted in the study, the finding shows a significant difference between the two groups (NDC vs. AC) viewing patterns and behavior when presented with a situation including material with human and social interaction content. The study reveals that a large percentage of autistic participants expressed minimum interest and time viewing the face area, especially the two Areas of Interests (AOIs) eyes and mouth regions comparing to the normal developing children. The findings revealed a significant difference between the two groups which is in line with the research results found in [4][6]. Autistic children fixate less on eyes (strong tendency to avoid fixation on the eyes) and expressed more interest in looking at the mouth than the normal developing children. The finding of this study is in line with the majority of international eye tracking studies indicating that individuals with ASD exhibited decreased visual attention to social stimuli relative to NDC. The results also provide quantitative assessment of how children with potential ASD process facial regions information when presented with socially rich context.

The study also provided practical evidences with respect to the speed of human face visual scanning and recognition. It appears that autistic children need more time to recognize and process facial and social information compared to normal developing children. The findings can be used to help improve the life style of other children who have a potential to develop autism as well as earlier ASD diagnosis. As future research, the results from the study can be used to develop an eye tracking-based framework to assist specialists to look for some early signs of potential children with ASD. As a future work, it would be worth pursuing to include other facial expression data such as human emotions in the data collection and analysis to investigate their impact on the children’s’ visual behavioral patterns in a social context that were not addressed by this study.

ACKNOWLEDGMENT

This research was supported by Zayed University Incentive Grant No. UL1RR024982. We would like to thank all the children, parents and RAs (Shaikha, Deena, Hessa, and Tania) who have ensured that accurate eye tracking data was collected. Finally, we give our thanks all the local special needs centers who provided the opportunity to meet with children and parents.

REFERENCES

[1] A. Susac, A. Bubic, J. Kaponja, M. Planinic, and M. Palmovic, "Eye Movements Reveal Students' strategies in Simple Equation Solving," Int. Jrnl. of Science & Mathematics Education, vol. 12, pp. 555-577, 2014.

[2] L. Speer, A. Cook, W McMahon, and E. Clark, "Face processing in children with autism: effects of stimulus contents and type," Autism : the international journal of research and practice, vol. 11, May 2007.

[3] A. Rizvi, “Autism being wrongly diagnosed,” Sep. 2017.

[4] K. Rayner, "Eye movements in reading and information processing: 20 years of research," Psychological bulletin, vol. 124, pp. 372-422, Nov. 1998.

[5] K. Rayner, "Eye movements in reading and information processing," Psychological bulletin, vol. 85, pp. 618-660, May 1978.

[6] K. Pelphrey et al., "Visual scanning of faces in autism," Journal of Autism and Developmental Disorders, vol. 32, pp. 249-261, Aug. 2002.

[7] B. Pan et al., "The Determinants of Web Page Viewing Behavior: An Eye-tracking Study," , New York, NY, USA, 2004, pp. 147–154.

[8] A. Klin, D. Lin, P. Gorrindo, G. Ramsay, and W. Jones, "Two-year-olds with autism orient to nonsocial contingencies rather than biological motion," Nature, vol. 459, pp. 257-261, 2009.

[9] A. Klin, W. Jones, R. Schultz, F. Volkmar, and D. Cohen, "Visual fixation patterns during viewing of naturalistic social situations as predictors of social competence in individuals with autism," Archives of General Psychiatry, vol. 59, Sep. 2002.

[10] J. Geest, C. Kemner, G. Camfferman, M. Verbaten, and H. Engeland, "Looking at images with human figures: comparison between autistic and normal children," Journal of Autism and Developmental Disorders, vol. 32, Apr. 2002.

[11] J. Geest, C. Kemner, M. Verbaten, and H. Engeland, "Gaze behavior of children with pervasive developmental disorder toward human faces: a fixation time study," Jrnl. of child psychology & psychiatry, & allied disciplines, vol. 43 ,July 2002.

[12] T. Frazier et al., "A Meta-Analysis of Gaze Differences to Social and Nonsocial Information Between Individuals With and Without Autism," Journal of the American Academy of Child and Adolescent Psychiatry, vol. 56, pp. 546-555, July 2017.

[13] J. Fedor et al., "Patterns of fixation during face recognition: Differences in autism across age," Autism: the Int. journl of research and practice, Aug. 2017.

[14] T. Falck-Ytter, E. Rehnberg, and S. Bölte, "Lack of visual orienting to biological motion and audiovisual synchrony in 3-year-olds with autism," vol. 8, 2013, URL: http://www.ncbi.nlm.nih.gov/pubmed/23861945/ [accessed: 2015-02-07].

[15] A. Duchowski, "Eye Tracking Methodology - Theory and Practice" Andrew Duchowski | Springer.: Springer-Verlag London, 2007.

[16] S. Chaudhary, May 2012 "The rise of autism in the UAE," URL: http://gulfnews.com/leisure/health/the-rise-of-autism-in-the-uae11/ [accessed: 2014-04-12].

MO2MD: Message-Oriented Middleware for Dynamic Management of IoT Devices

Imen Ben Ida
Electronic systems and
communications networks laboratory
(SERCOM),
Polytechnic School of Tunisia,
Carthage University
Tunis, Tunisia
Email: Imen.benida@gmail.com

Takoua Abdellatif
Electronic systems and
communications networks laboratory
(SERCOM),
Polytechnic School of Tunisia,
Carthage University
Tunis, Tunisia
Email: takoua.abdellatif@ept.rnu.tn

Abderrazek Jemai
Electronic systems and
communications networks laboratory
(SERCOM),
Polytechnic School of Tunisia,
INSAT, Carthage University
Tunis, Tunisia
Email: Abderrazek.Jemai@insat.rnu.tn

***Abstract*— Middleware are fundamental components for Internet of Things (IoT) solutions. They provide general and specific abstractions through which smart devices and their related applications can be easily interconnected. However, due to the wide variety of software and hardware technologies of IoT solutions, the management of the connected devices is still a challenging task, especially for Cloud-Edge based solutions. In this paper, we take advantage of message-oriented computing and Web technologies to propose a solution for devices management without having to worry about the underlying infrastructures or implementation details. In particular, we describe a Web-based middleware that enables configurability and manageability of connected devices in a dynamic manner at the Cloud level.**

Keywords-Message-Oriented Middleware; IoT; Cloud comuting; Edge computing; Devices Management.

I. Introduction

IoT devices, also known as smart objects, are projected to grow exponentially both in terms of quantity as well as variety [1]. Some examples include wireless body sensors, smart vehicles, and surveillance cameras. By connecting devices in an Internet-like structure, a variety of data can be collected, which is of great benefit in industry and daily life. To support such data streams, the Cloud infrastructure provides several services namely, Software as a Service (SaaS), Platform as a Service (PaaS) and Infrastructure as a Service (IaaS) for massive-scale and complex data computing. It takes advantage of virtualized resources, parallel processing and data service integration [1].

However, due to the explosive growth of connected lightweight devices, Cloud computing is facing increasing challenges, especially in managing and reconfiguring IoT devices. The challenge of flexible devices configuration from the Cloud layer is due to several domain requirements, such as context-awareness and delay-sensitive control [2]. In addition, designing and implementing an appropriate mechanism that can dynamically manage resources across the Cloud-IoT spectrum is a challenging task to resolve due to the highly dynamic behaviors of the devices [1].

As response to this challenge, the Edge computing paradigm enables the Cloud resources to move closer to the devices network by offering localized devices configuration. The Edge layer is exploited by reinforcing edges, such as gateways, with sufficient processing power, intelligence, and communication capabilities to integrate efficiently the Cloud layer with the sensors layer and to provide efficient configuration. It provides not only local data processing and data storage, but also it ensures local management of the IoT devices [3].

In this paper, we explore the Edge computing paradigm to implement a distributed solution for a dynamic management of IoT devices. In particular, we propose a Message-Oriented Middleware for Dynamic Management of IoT Devices (MO2MD). Middleware are widely used in distributed systems and they are considered as fundamental tools that provide general and specific abstractions for the design and the implementation of smart environment applications [4]. A middleware can ease the development process by integrating heterogeneous computing and communications devices and by supporting the interoperability within the diverse applications and services [5].

More specifically, our proposed middleware MO2MD offers a flexible configuration of IoT devices to support a dynamic management of the Edge layer. The remainder of this paper is organized as follows. In Section 2, we describe the message-oriented approach for IoT middleware and the challenges of IoT devices. Some related works are highlighted in Section 3. Our proposed solution is detailed in Section 4. Section 5 presents concluding remarks and future work.

II. BACKGROUND

A. *Message-oriented Middleware for Internet of Things*

A middleware is a software component that allows programming IoT solutions with a higher level of abstraction. It provides a simpler interface to ensure appropriate abstractions and mechanisms for dealing with the heterogeneity of IoT devices. It simplifies the development and execution of distributed applications and hides their complexity. In particular, middleware removes the programmer from the complex aspects of IoT, such as the handling of wireless communications, power management and hardware programming.

In event-based middleware, components, applications, and all the other participants interact through events. Each event has a type, as well as a set of typed parameters whose specific values describe the change to the producer's state [5]. Events are propagated from the sending application components (producers) to the receiving application components (consumers).

Message-Oriented Middleware (MOM) is a type of event-based middleware. In this model, the communication is based on messages. Generally, messages carry publisher and subscriber addresses and they are delivered by a particular subset of participants, whereas events are broadcast to all participants.

B. Challenges of IoT Devices Management

We present in the following paragraphs the main challenges of implementing middleware components to manage IoT devices [2][3][5]-[7]:

1) Heterogeneity:

Device heterogeneity emerges from different characteristics, such as differences in capacity, features, and application requirements. Added to that, the emergence of new protocols as an important lightweight support for the IoT devices communication requires appropriate communication mechanisms for the delivery, support and management of the different types of resources.

2) Dynamic behaviors:

In an IoT environment, thousands of devices may interact with each other even in one local place (e.g., in a building, supermarket, and hospital), which is much larger scale than most conventional networking systems. The interactions among a large number of devices will produce an enormous number of events. This may cause problems, such as event congestion and reduced event processing capability. Consequently, any predefined and fixed set resource management policies will be rendered useless for a dynamic management of devices.

3) Scalability:

Scalability is a significant challenge for current IoT platforms, where lightweight IoT devices can hardly extend their functionalities by adding new hardware modules. For example, it is very difficult to integrate a temperature detection service on an IoT device by simply attaching a temperature sensor. Added to that, other components, such as resource discovery and data analytics of IoT solutions need to be scalable to achieve system-wide scalability.

4) Resource Constrains:

With smaller, more compact sensors, the available battery power is always limited. The IoT systems must be designed to manage limited power by designing efficient processes and capabilities of the sensors. Mechanisms to ensure efficient power consumption are necessary for IoT-based services.

III. RELATED WORK

Numerous middleware proposals have been put forward for IoT-based applications. They provide abstractions through which connected devices and their related applications can be easily built up and managed.

In database-oriented middleware, the devices are considered as virtual objects in a relational database, so that an application can perform queries using a syntax in SQL language, allowing complex queries to be performed. For example, Global Sensor Net-works (GSN) [8] is an IoT middleware that aims to provide flexible management of heterogeneous IoT devices. It enables developers to specify XML-based deployment descriptors to deploy a sensor. An implementation of the wrapper in Java is required in order to add a new type of sensor to the middleware.

In event-driven middleware, components, applications, and other participants interact through events. In [9], the authors present an event-driven user-centric middleware for monitoring and managing energy consumption in public buildings and spaces. The proposed middleware allows the integration of heterogeneous technologies in order to enable a hardware independent interoperability between them.

UbiSOAP (Service-Oriented Middleware for Ubiquitous Networking) is a service-oriented middleware [10] that provides complete integration of the network with Web Services. The architectural resources layer has the necessary functions, including a unified abstraction for simple services (sensors, actuators, processors or software components) to help integrate applications and services with resources. A service support component facilitates the discovery and dynamic composition of resources (eg services). Dynamic composition and instantiation of new services are facilitated by semantic models and descriptions of sensors, actuators and processing elements.

In [11], the authors present a QoS aware publish/subscribe middleware for Edge computing called EMMA (edge-enabled publish–subscribe middleware). They show that EMMA can provide low-latency communication for devices in close proximity, while allowing message dissemination to different locations at minimal overhead costs. Gateways allow existing publish/subscribe client infrastructure to transparently connect to the system.

Another message-oriented middleware used for communication between the system services of a proposed framework is presented in [12]. The authors analyze the driver real-time data using a distributed system architecture and send alert messages in a timely manner. According to their experimental results, the middleware design increases the speed and stability of information transmission.

Other types of middleware are used in a distributed environment, such as Transaction-Oriented Middleware (TOM), which is used to ensure the correctness of transaction operations and Object-Oriented/Component Middleware (OOCM), which is based on object-oriented programming models requests [13].

Considering the characteristics of middleware introduced above and the goal of a dynamic management of IoT devices, it is possible to argue the suitability of message-oriented middleware for a flexible configuration. The interactions of database-driven and object-oriented models are synchronous, which limits the scalability to large volumes of data. Not being designed for concurrent event management, these usually do not attain the same level of performance as systems designed for the event-based interaction paradigm.

IV. MO2MD PRESENTATION

Our proposed message-oriented middleware for Dynamic Management of IoT Devices, named MO2MD, extends the capabilities of messages-oriented middleware and provides high flexibility for adding new configurations of devices at the Edge layer. The proposed middleware considers all connected devices, such as sensors and actuators, as data providers. We focus on the challenge of dynamic behavior and scalability.

A. Architecture

The proposed architecture is depicted in Figure 1. The implementation of MO2MD is achieved through a distributed architecture which integrates global Cloud services with local services in different Edge nodes. All participating Edge nodes communicate with the Cloud layer to exchange devices data and to receive configuration data. The following paragraphs describe the principal components of the proposed architecture:

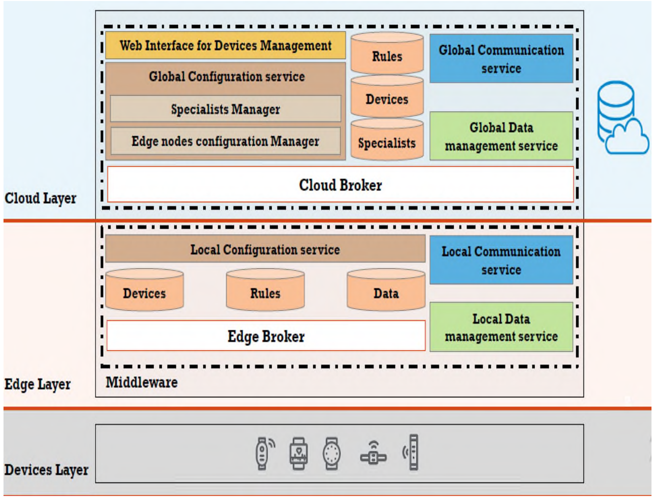

Web Interface for Devices Management
Rules
Global Communication service
Global Configuration service
Devices
Specialists Manager
Edge nodes configuration Manager
Specialists
Global Data management service
Cloud Broker
Cloud Layer
Local Configuration service
Local Communication service
Devices
Rules
Data
Edge Broker
Local Data management service
Edge Layer
Middleware
Devices Layer

Figure 1. MO2MD architecture.

1) Cloud Layer

The Cloud layer is composed of a cluster of servers with massive computing and storage resources. This layer communicates with three types of participants:

- *Admins*

They are the middleware users who control the different IoT devices. They are responsible of adding or modifying the middleware configuration.

- *Specialists*

They are responsible for interventions in case of devices malfunction. They are considered as data subscribers and receive notifications about the devices status depending on their subscriptions.

- *Gateways at the Edge layer*

They are local gateways, in which local services process the collected data from different devices and apply the configuration requests received from the Cloud layer.

A Configuration Web application is the entry point of the Cloud layer services. It allows the middleware admins to view everything as a single large system that provides basic and advanced services. The Web application provides management tools to control the whole IoT platform and to request new configurations that will be processed by a global configuration service and published to the corresponding Edge Node. The exchange of messages between the Cloud layer and the Edge layer is ensured using the publish/subscribe pattern [14]. A global broker handles the requests of configuration from the admins and publishes the submitted configurations as messages to the corresponding Edge node.

2) Edge Layer

The Edge layer reinforces the devices layer with storage, processing and communication capabilities. It is the intermediate layer between the IoT devices and the Cloud services. A local broker service in each Edge node regularly updates the global broker about the availability and status of their devices and it receives the data provided by IoT devices. The Edge layer aims to provide low latency responses.

3) Distributed Services for dynamic management of IoT devices

a) Data management service

Data are the key of efficient devices management. In the proposed middleware, we consider 3 types of exchanged data, which are presented in Table 1.

TABLE I. DATA TYPES

Sensed data	Configuration data	Notifications
Data collected by the devices.	The configuration parameters requested by the admins.	Notifications of the devices dysfunction.

The proposed middleware provides data management services, such as data acquisition, data processing and data storage.

The configuration data are the parameters and the rules which must be taken in consideration for data management at the Edge layer. In case of non-respect of configured rules, notification messages are sent to the concerned specialists.

b) Configuration service

The core component of the middleware is the global configuration service which is composed of two major units. The first unit is the specialist's manager, which allows the admin to add, update or delete the corresponding specialists of each type of device. The registered specialists will be notified in case of an abnormal behavior of any device at the Edge layer. The second one is the Edge nodes manager, which is responsible for processing the configuration requests of the admins. We choose as configuration requests:

- The interval of saving data in the local data base of each device.
- The priority of each device.
- The corresponding specialist to notify in case of device disfunction.
- The interval of sending data to the Cloud layer.
- Specific rules for each device.

All the requests may be introduced through the Web application, depending on the decision of the admins. This Web-based configuration enables a dynamic management of the device's behaviors. For example, in the case of smart buildings, the buildings' owner can change the interval of saving temperature data depending on the building location. Another example, in a smart hospital, a doctor can modify the priority of each medical device depending on the patient situation.

The configuration data, the specialist's subscriptions and the list of active devices are saved in both the Cloud and the Edge layer. Each configuration is considered as a rule to respect at the Edge layer. In Figure 2, we illustrate 2 different scenarios of data exchange between the Edge layer and the Cloud layer.

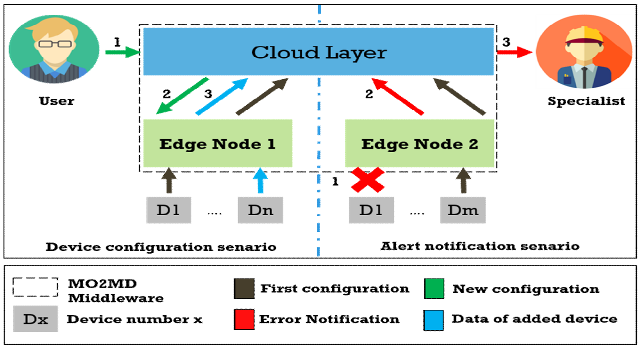


Figure 2. Scenarios of dynamic data management.

c) Communication service

The communication service supports a publish-subscribe messaging service that ensures data-centric communication between the Cloud layer and the Edge nodes. It also supports the link between external databases and the Cloud layer.

With the publish-subscribe pattern, the local communication managers act as publishers on a given topic and send messages without the need to know about the existence of the receiving clients, who are the Specialists. At the Edge layer, the communication service ensures the data storage in a local database and the synchronization between the Cloud broker and the local services to send the collected data and any dysfunction detection.

B. Implementation and evaluation

The implemented middleware is based on Node.js which allows multi-platform development and supports JavaScript-based webservers. The I/O architecture of Node.js is based on non-blocking asynchronous event-driven which makes it suitable for data-intensive and real-time applications in lightweight and efficient way. The Edge is a Raspberry Pi 3 which is a low-cost small-sized single board with 1 GB of Ram and 1.2 GHz processor [15]. The messages exchanged are ensured by MQ Telemetry Transport (MQTT) protocol [16]. The MQTT protocol is a lightweight application layer protocol designed for resource-constrained devices. It uses the publish-subscribe messaging system combined with the concept of topics to provide one-to-many message distribution. It supports a range of 10 to 100 messages per second.

We install the InfluxDB database, which is an open-source Time Series Database. At its core is a custom-built storage engine called the Time-Structured Merge (TSM) Tree, which is optimized for time-series data. InfluxDB provides support for mathematical and statistical functions across time ranges; also it is developed for custom monitoring, metrics collection and real-time analytics [17].

To prove the benefits of the dynamic configuration of the Edge layer, we consider three different scenarios of controlling 10 devices in one day. We suppose that the static configuration of the interval of data storage in the local database is a second. In each scenario, we change this interval for a certain number of devices. Table 2 shows the interval configuration for each scenario, as well as the percentage of gain in terms of memory compared to the static configuration.

TABLE II. CONFIGURATION SCENARIOS

Scenario	Number of devices per time			Gain of memory
	hour	*minute*	*second*	
Fixed case	2	1	7	29.83 %
Custom case 1	5	2	3	69.65 %
Custom case 2	2	7	1	88.82 %

Figure 3 shows that the use of a single data processing strategy in a fixed and non-custom way can result in an unnecessary use of gateway memory which obviously affects performance and the time reaction e in emergency cases.

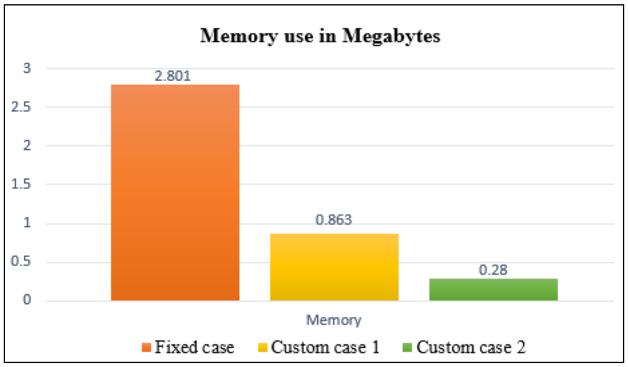


Figure 3. Scenarios of dynamic data management.

V. CONCLUSION AND FUTURE WORK

Flexible configuration requires the development of a dynamic mechanism of devices management. In this paper, we describe a message-oriented middleware that offers distributed services for IoT devices management. The proposed architecture supports a flexible configuration of connected devices in different locations from a Web application. In particular, we show that our middleware gives the possibility to modify the data storage interval at the Edge layer in order to personalize the devices behavior and realize resources optimization.

Future work includes the complete implementation of the proposed middleware, including its security and reliability guarantees, as well as automatic resource discovery to realize autonomous devices deployment at the Edge layer.

REFERENCES

[1] H. El-Sayed et al., "Edge of Things: The Big Picture on the Integration of Edge, IoT and the Cloud in a Distributed Computing Environment," IEEE Access, vol. 6, pp. 1706-1717, 2018. doi: 10.1109/ACCESS.2017.2780087

[2] J. Ren, H. Guo, C. Xu and Y. Zhang, "Serving at the Edge: A Scalable IoT Architecture Based on Transparent Computing," IEEE Network, vol. 31, no. 5, pp. 96-105, 2017. doi: 10.1109/MNET.2017.1700030

[3] S. Shekhar and A. Gokhale, "Dynamic Resource Management Across Cloud-Edge Resources for Performance-Sensitive Applications," 2017 17th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGRID), Madrid, 2017, pp. 707-710.

[4] G. Fortino, A. Guerrieri, W. Russo and C. Savaglio "Middleware for Smart Objects and Smart Environments: Overview and Comparison," G. Fortino,P. Trunfio,eds. , Internet of Things Based on Smart Objects. Internet of Things (Technology, Communications and Computing), Springer, Cham, 2014, pp. 1-27.

[5] M. A. Razzaque, M. Milojevic-Jevric, A. Palade and S. Clarke, "Middleware for Internet of Things: A Survey," in IEEE Internet of Things Journal, vol. 3, no. 1, pp. 70-95, Feb. 2016. doi: 10.1109/JIOT.2015.2498900

[6] L. Alonso et al., "Middleware and communication technologies for structural health monitoring of critical infrastructures: A survey," Computer Standards & Interfaces, vol. 56, pp. 83-100, 2018, doi: 10.1016/j.csi.2017.09.007

[7] N. Naik, "Choice of effective messaging protocols for IoT systems: MQTT, CoAP, AMQP and HTTP," 2017 IEEE International Systems Engineering Symposium (ISSE), pp. 1-7, Vienna, 2017. doi: 10.1109/SysEng.2017.8088251.

[8] A. H. Ngu et al, " IoT Middleware: A Survey on Issues and Enabling Technologies," IEEE Internet of Things Journal, vol. 4, pp. 1-20, Feb. 2017, doi: 10.1109/JIOT.2016.2615180

[9] E. Patti et al., "Event-Driven User-Centric Middleware for Energy-Efficient Buildings and Public Spaces," IEEE Systems Journal, vol. 10, pp. 1137-1146, Sept. 2016, doi: 10.1109/JSYST.2014.2302750

[10] M. Caporuscio, P.G. Raverdy and V. Issarny, "Ubisoap: a service-oriented middleware for ubiquitous networking," IEEE Trans. Serv. Comput., vol. 5,pp. 86–98, 2012. https://doi.org/10.1109/TSC.2010.60

[11] X. Xu et al., "EAaaS: Edge Analytics as a Service," 2017 IEEE International Conference on Web Services (ICWS), Honolulu, HI, 2017, pp.9-356.

[12] P. Lai, C. Dow and Y. Chang. "Rapid-Response Framework for Defensive Driving Based on Internet of Vehicles Using Message-Oriented Middleware," IEEE Access, vol. 6,pp. 18548-18560, 2018, doi: 10.1109/ACCESS.2018.2808913

[13] M. Albano, L.L Ferreira, L. M. Pinho, and A. R. Alkhawaja,

" Message-Oriented Middleware for smart grids," Computer Standards & Interfaces, vol.38, pp. 133-143, 2015. https://doi.org/10.1016/j.csi.2014.08.002

[14] A. Hakiri, P. Berthou, A. Gokhale and S. Abdellatif, "Publish/subscribe-enabled software defined networking for efficient and scalable IoT communications," IEEE Communications Magazine, vol. 53, no. 9, pp. 48-54, September 2015. doi: 10.1109/MCOM.2015.7263372

[15] J. Bermúdez-Ortega et al., "Remote Web-based Control Laboratory for Mobile Devices based on EJsS, Raspberry Pi and Node.js", IFAC-PapersOnLine,vol. 48, no. 29, pp. 158-163, 2015.

[16] A. Banks and R. Gupta, MQTT Version 3.1. 1. OASIS standard, vol. 29, 2014.

[17] C. Rudolf, "SQL, noSQL or newSQL–comparison and applicability for Smart Spaces," Network Architectures and Services, 2017.

Microservices for Multimobility in a Smart City

Cristian Lai, Francesco Boi, Alberto Buschettu and Renato Caboni

CRS4, Center for Advanced Studies, Research and Development in Sardinia,
Pula Italy
Email: {cristian.lai, francesco.boi, alberto.buschettu, renato.caboni}@crs4.it

Abstract—**In this paper, we discuss our thoughts of microservice architecture and the way to build Internet of Things (IoT) services for multimobility in a smart city. We briefly introduce a draft architecture that we have used to develop an experimental Web application, designed to be used in a real case study and capable of interfacing with a wide range of heterogeneous IoT devices and services.**

Keywords*–*IoT; Smart City; Microservices; Multimobility.

I. Introduction

Information and Communication Technology (ICT) is evolving toward more scalable architectures, and those based on microservices are paving the way to modern applications [1]–[3]. The number of connected devices available in real life has significantly grown [4] as crucial parts of the IoT. In this paper, we discuss how we figure out microservice architectures and efficient applications specifically designed for the IoT field in front of a large number of users and a high demand of resources, as opposed to monolithic architectures. We introduce the draft of an architecture used to develop an experimental Web application for multimobility services in a smart city. This architecture is capable of integrating both physical and logical devices as well as of managing typical operations, such as device registration, data storage, and data retrieval.

In a smart city, multimobility combines different modalities of transportation [5], e.g., private cars, bus, carsharing, and bikesharing. The shift from homogeneous to multimodal mobility is growing in popularity, especially in urban centers with recurring problems associated with congestion, parking, and an overall lack of space [6]. Drivers moving by car towards an area with high traffic volume, such as the city centre, have the available information necessary to elude high traffic intensity areas avoiding time and fuel wasting besides preventing the increase of traffic. Every single element, such as a parking area, a bus stop, a car or bike sharing station, takes part in an extremely sophisticated network of miscellaneous connected IoT devices. Within a scalable system, a single device must be managed independently from the others. With the microservice paradigm, each device is managed by a dedicated microservice.

This paper is organized as follows. Section II reviews state of the art service-oriented architectures. Section III describes our microservice architecture. Finally, Section IV provides conclusions and future perspectives.

II. State of the Art

In IoT systems, a centralized architecture is responsible for offering one or more services to the user while the necessary data is produced by a set of devices deployed in different locations. These devices generate an amount of data readily available, used to create vertical applications. One way to access these data is usually using cloud-based platforms [7]–[9] (service-oriented centralized architectures). These platforms provide Application Programming Interface (API) for storing and retrieving data and interfacing with existing systems using the most common IoT protocols, such as HyperText Transfer Protocol (HTTP) and Message Queuing Telemetry Transport (MQTT) [10]. In this kind of architecture, device management is centralized, and this can generate an overload and bottlenecks. To date, the centralized architectures are successful state of art technologies but built as monolithic solutions, not offering the flexibility required to deal with heterogeneous devices efficiently. Often, they consist of three main parts: a client-side user interface, a database and a server-side application. Changes to the system necessitate building and deploying a new version of each component.

On the contrary, microservice architectures are divided into some small independent services, each of which implements a specific feature. A scalable architecture achieves more efficient management of available resources by allocating more only to those modules that are overstressed, rather than unconditionally to all the sub-components [2] [11]. To overcome the previously discussed limitations, we have been adopting the microservice paradigm and developing an architecture that provides independently deployable and loosely coupled basic services.

III. Microservice Architecture

Our concept of microservice architecture is a collection of independently deployable and loosely coupled basic services. Each microservice runs in its process, communicates with lightweight mechanisms, such as HTTP resource API [12] and manages significant operations. Our microservice architecture, namely *CRS4 Microservice Core for Iot* (CMC-IoT), is composed of basic microservices: i) **Cmc Auth** is a token generator that protects microservices from unauthorized access using a token-based authentication technique; ii) **Cmc App** manages resources and applications sign-up and the subsequent sign-in phase; iii) **Cmc User** manages user access to protected microservices. Basic microservices have been used for developing a specialized set for the IoT, including: i) **Cmc Devices** manages the device functionalities providing the Representational State Transfer (REST) create, read, update and delete (CRUD) operations; ii) **Cmc History** stores and retrieves historical data produced by devices; **Cmc Persistence** is a scheduler for general-purpose devices that do not directly provide their data. Once a device has been added to CMC Devices, it can send its data through a token authorized write operation. Then, data can be read by other authorized devices, applications, etc. Cmc Devices uses the basic microservices to check if tokens are valid and authorized to query a specific device (see Figure 1). CMC-IoT is the foundation of our first Web application designed for citizen multimobility in a smart city, called SmartMobility. Smartmobility is the name of the application built on top of the presented architecture. It presents on a User Interface (UI) smart city data belonging to diverse mobility services. Currently, the contemplated services are parking areas, traffic sensors, public bus transportation, and car and bike sharing services. The application is supposed to be used in the following scenario. Before entering the city centre, the driver, using SmartMobility, can check the availability of free parking spots in the monitored parking areas close to his destination. In this way, she does not have to drive around the city looking for a free parking spot. Once the parking lot has been chosen, she can plan the fastest path to reach it according to real-time traffic information in the main city roads shown in SmartMobility. From the parking area, the driver can again check on the application the availability of mobility services, such as bus stops and sharing services, and their reliability so that she can choose the one most suitable for his needs or alternatively walk to his final destination. All the devices, managed by CMC-IoT and composing the considered mobility services, are shown on a map in the UI. Each graphical element shows additional information. For example, the parking lot shows in real-time not only the overall parking capacity but also further information about other mobility services in the area close to it. The bus stop provides the lines, the timeschedules and the service reliability. Carsharing and bikesharing show the number of available vehicles.

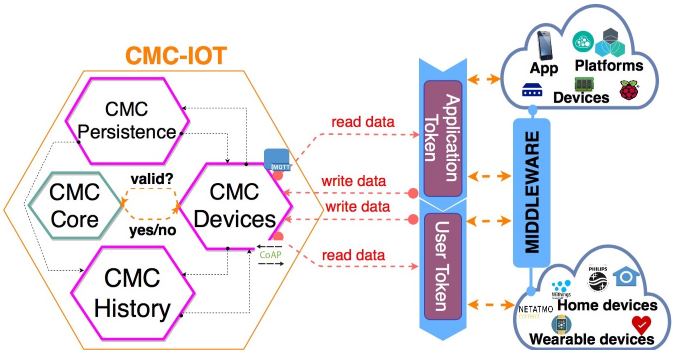


Figure 1. CMC IoT.

IV. CONCLUSION

Our application approaches the effectiveness of the drafted microservice architecture. The potential of the adopted microservice architecture allows designing modular systems. They can grow incrementally without continuous redesign, development, and deployment of the entire application. The system is divided into small and lightweight services, purposely built to perform a very cohesive business function. Every single element, i.e., a parking area, a bus stop, a car or bike sharing station could be added as well as removed independently, while the system can be further enriched with new services. Both these actions can be performed by modifying only the directly interested modules without affecting the others. Services independence allows reaching for an extremely sophisticated network of connected IoT devices. As a follow-up, we will develop a real case scenario demonstrating the implementation of the microservice architecture. Several tests will be performed to determine performances under an increasing number of requests per unit of time.

ACKNOWLEDGMENT

This work has been partially supported under the ERDF (European Regional Development Fund), POR Sardegna FESR 2007-2013, PIA 2013 (project: n.295 PIA ENTANDO).

REFERENCES

[1] S. Baskarada, V. Nguyen, and A. Koronios, “Architecting microservices: Practical opportunities and challenges,” Journal of Computer Information Systems, 09 2018, pp. 1–9.

[2] N. Dragoni et al., “Microservices: How to make your application scale,” Perspectives of System Informatics, 2018, pp. 95–104.

[3] M. Kalske, N. Mäkitalo, and T. Mikkonen, “Challenges when moving from monolith to microservice architecture,” in Current Trends in Web Engineering, I. Garrigós and M. Wimmer, Eds. Cham: Springer International Publishing, 2018, pp. 32–47.

[4] M. Hung, “Leading the IoT,” Gartner, Tech. Rep., 2017. [Online]. Available: http://www.gartner.com/imagesrv/books/iot/iotEbook_digital.pdf [accessed: 2019-06-17]

[5] A. Zanella, N. Bui, A. Castellani, L. Vangelista, and M. Zorzi, “Internet of things for smart cities,” IEEE Internet of Things Journal, vol. 1, no. 1, Feb 2014, pp. 22–32.

[6] S. Shaheen, A. Stocker, and A. Bhattacharyya, “Multimobility and Sharing Economy,” Transportation Research Board, 500 Fifth Street, NW, Washington, DC 20001, Tech. Rep. E-C120, 2016.

[7] A. I. Abdul-Rahman and C. A. Graves, “Internet of things application using tethered msp430 to thingspeak cloud,” in SOSE. IEEE Computer Society, 2016, pp. 352–357.

[8] C. Lai, A. Pintus, and A. Serra, “Using the web of data in semantic sensor networks,” in Complex, Intelligent, and Software Intensive Systems CISIS 2017. Advances in Intelligent Systems and Computing, vol 611. Springer, Cham, L. Barolli and O. Terzo, Eds., 2018, pp. 106–116.

[9] C. Perera, A. Zaslavsky, P. Christen, and D. Georgakopoulos, “Sensing as a service model for smart cities supported by internet of things,” Transactions on Emerging Telecommunications Technologies, vol. 25, no. 1, 2013, pp. 81–93.

[10] A. H. Ngu, M. Gutierrez, V. Metsis, S. Nepal, and Q. Z. Sheng, “Iot middleware: A survey on issues and enabling technologies,” IEEE Internet of Things Journal, vol. 4, no. 1, Feb 2017, pp. 1–20.

[11] A. Krylovskiy, M. Jahn, and E. Patti, “Designing a smart city internet of things platform with microservice architecture,” in Proceedings of the 2015 3rd International Conference on Future Internet of Things and Cloud, ser. FICLOUD ’15. Washington, DC, USA: IEEE Computer Society, 2015, pp. 25–30.

[12] M. Fowler and J. Lewis, Microservices, 2014. [Online]. Available: http://martinfowler.com/articles/microservices.html [accessed: 2019-06-17]

Semantic-based Linked Data Management Platform

Yongju Lee, Hyunseung Seok, Seonghyeon Nam
School of Computer Science and Engineering
Kyungpook National University, Daegu, Korea
e-mail: yongju@knu.ac.kr, seokhyunseung@hanmail.net, connernam@gmail.com

***Abstract*—The growing number of available Linked Datasets raises a challenging data management problem, namely, how should the big data be stored and the desired resources be located. Although many platforms, architectures, and mechanisms are proposed for Linked Data, there are still a number of limitations in the area of Linked Data. We propose a novel semantic-based Linked Data management platform, which is composed of a storage-indexing-retrieval system, an ontology learning system, and an automatic composition system. This platform serves as a basis for implementing other more sophisticated applications required in the area of Linked Data.**

***Keywords*—Linked Data; platform; storage and retrieval system; ontology learning; automatic composition.**

I. INTRODUCTION

A very pragmatic approach towards achieving the Semantic Web has gained some traction with Linked Data. Linked Data refers to a set of best practices for publishing and interlinking structured data on the Web [1]. The basic idea of Linked Data is to apply the general architecture of the Web to the task of sharing structured data on a global scale. Technically, Linked Data employ Uniform Resource Identifications (URIs), Resource Description Frameworks (RDFs), and the Hypertext Transfer Protocol (HTTP) to publish structured data and connect related data that are distributed across multiple data resources.

RDF [2] is the data model for Linked Data, and SPARQL [3] is the standard query language for this model. All data items in RDF are represented in triples of the form (*subject, predicate, object*). Spurred by efforts like the Linked Open Data (LOD) project [4], a large amount of semantic data are available in the RDF format in many fields such as science, business, bioinformatics, and social networks. These large volumes of RDF data motivate the need for scalable RDF data management solutions capable of efficiently storing, searching, and integrating RDF data. This paper introduces our current project entitled "semantic-based Linked Data management platform." Work in our project focuses on the development of a storage-indexing-retrieval system, and an ontology learning system, and an automatic composition system. These topics are described in detail in Sections 2, 3, and 4.

II. STORAGE-INDEXING-RETRIEVAL SYSTEM

This section starts by providing a software architecture for the storage-indexing-retrieval system (see Figure 1). Our system consists of four subsystems: data acquisition, RDF storage, ontology construction, and analysis subsystems.

A. Data Acquisition Subsystem

Information represented in unstructured or structured form must be mapped to the RDF data model. We crawl Web sites and extract unstructured data. This procedure is based on a crawler such as Scrapy [5]. The extracted properties are then transformed into RDF triples. Structure data (e.g., relational databases) are transformed into RDF triples using the D2R Server [6]. The D2R Server is an open source tool for publishing relational databases on the Linked Data.

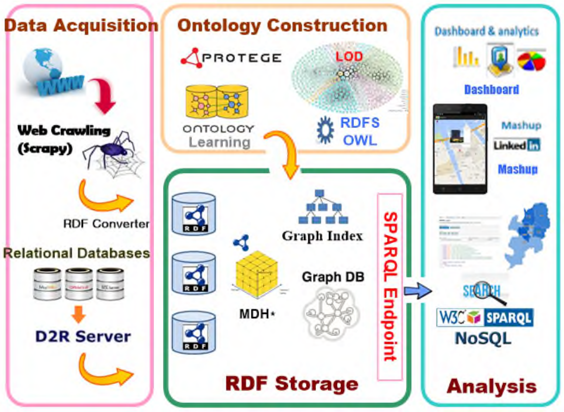


Figure 1. Architecture of storage-indexing-retrieval system.

B. RDF Storage Subsystem

Once there is a critical mass of RDF data, mechanisms have to be in place to store and index this data efficiently. Our system uses the graph-based RDF store. With this store, the platform provides a SPARQL endpoint that allows any user to access the stored RDF triples. Our system indexes the triples stored in the RDF store. These triples are mapped into multi-dimensional histograms stored in an MDH* (Multi-Dimensional Histograms for Linked Data) index structure [7].

C. Ontology Construction Subsystem

RDF Schema (RDFS) and Web Ontology Language (OWL) are key semantic Web technologies that provide a way to write down rich descriptions of your RDF data. Protégé [8] is a leading ontological engineering tool. In spite of using this ontological engineering tool, the construction of ontologies is a very expensive task which hinges on the availability of domain experts. In Section 3, we investigate an ontology learning method to generate ontologies automatically.

D. Analysis Subsystem

A number of applications provided the browsing Linked Data (e.g., Tabulator [9], Marbles [10], Magpie [11]), advanced searching facilities (e.g., Sindice [12], Sig.ma [13], Watson [14]), and meshup Linked Data (e.g., DbpediaMobile [10], LinkedGeoData [15], ActiveHiring [16]). Nevertheless, these applications hardly go beyond presenting together data gathered from different sources. Our system provides search capabilities, such as SPARQL and NoSQL (Not Only SQL) over the RDF triples. In addition, we can advance to build various applications based on Linked Data. For instance, we implement an aggregation of dashboards that presents various business analytics computed on the Linked Data [17].

III. Ontology Learning System

The successful employment of Linked Data is dependent on the availability of high quality ontologies. Building such ontologies is difficult and costly, thus hampering Linked Data deployment. This research automatically generates ontologies from RDF datasets and their underlying semantics. We focus on adapting Linked Data mining techniques to the syntactic descriptions of RDF triples. Since RDF was not designed for the ontology, it does not provide placeholders for high level syntaxes of the resources. We propose an ontology learning method to semantically describe Linked Data (see Figure 2).

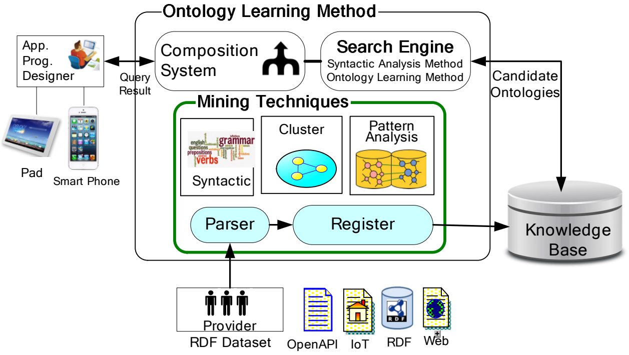


Figure 2. Ontology learning system.

We have developed a clustering technique [18] to derive several semantically meaningful concepts from RDF triples. We consider the syntactic information that resides in triples and apply a mining algorithm to obtain their underlying semantics. The main idea is to measure the co-occurrence of terms and then cluster the terms into a set of concepts. The pattern analysis technique [18] also captures relationships between terms contained in triples and matches items if both terms are similar and the relationships are equivalent. The ontology is generated from the set of triples created in accordance with the pattern analysis rules.

IV. Automatic Composition System

With the growing popularity of Linked Data, the number of RDF datasets has increased significantly. As a result, finding and composing the right resources has become an increasingly complex task. Recent approaches have stressed the importance of Linked Data composition, data mediation, and the semantic annotations of Linked Data. Although they try to overcome the limitations of traditional mashup solutions, there are several challenging issues. First, because LOD cloud may have a large number of datasets, manually searching and composing RDF triples can be a tedious and time-consuming task. Therefore, developers wish to quickly find the desired items and easily integrate them. Second, portal sites typically only support keyword or category search. The keyword search is insufficient because of bad recall and precision. Returned lists from the category search are generally based on criteria that have no relevance to the developer's desired goals. To create mashups more efficiently, a semantic-based approach is needed such that agents can reason about the capabilities of the items that permit their discovery and composition. Third, most mashup developers want to figure out all the intermediate steps needed to generate the desired mashup automatically. An infrastructure that allows users to provide interesting or relevant composition candidates is needed.

Our research investigates algorithms for automatic Linked Data discovery and composition using the ontology learning method [19]. A common issue is how to locate the desired items. Efficient discovery can play a critical role in conducting further RDF composition. Our discovery algorithm adopts strategies that rapidly filter out items that are guaranteed not to match the query. The composition algorithm consists of constructing a Compatible Similarity Graph (CSG) and searching composition candidates. The composition process can be described as generating Directed Acyclic Graphs (DAGs) that can produce the output satisfying the desired goal. The DAGs are gradually generated by forward and backward searching over the graph (see Figure 3).

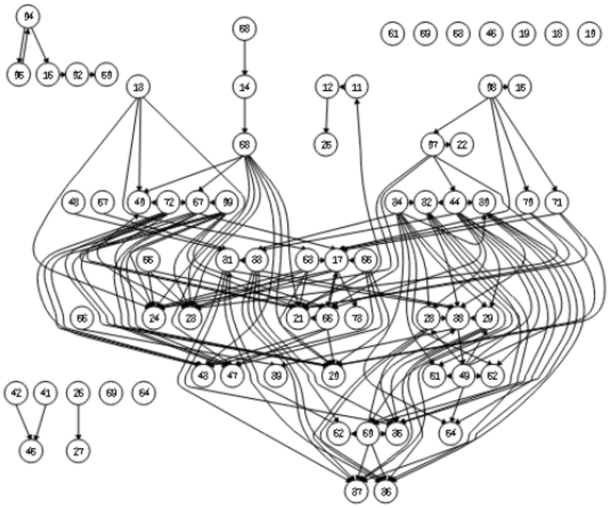


Figure 3. Compatible similarity graph for Linked Data composition.

V. Conclusion

The evolution of Linked Data has created a strong wave of research approaches in Semantic Web community. We introduced a semantic-based Linked Data management platform, which consists of the storage-indexing-retrieval system, onto-logy learning system, and automatic composition system. In this paper, we describe briefly the overview of our ongoing project. Our platform has an integrated hybrid architecture which supports the whole life-cycle of Linked Data from data acquisitions, building ontologies, data storage and retrieval, and implementing applications. The main components of the platform are open source in order to facilitate wide usage and ease the scalability. This proposal is a first phase of our research aiming at increasing tool coverage and developing real Linked Data applications.

Acknowledgment

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (No. 2016R1D1B02008553).

References

[1] S. Auer, J. Lehmann, A. N. Ngomo, and A. Zaveri, “Introduction to Linked Data and Its Lifecycle on the Web,” Proc. 9th Int. Conference on Reasoning Web, Aug. 2013, pp. 1-9.

[2] W3C. *Resource Description Framework (RDF)*. Available from: https://www.w3.org/RDF/ 2014.03.15

[3] W3C. *SAPRQL 1.1 Query Language*. Available from: https:// www.w3.org/TR/sparql11-query/ 2013.03.21

[4] Lod-cloud.net. *The Linked Open Data Cloud*. Available from: https://lod-cloud.net/ 2019.03.01

[5] Scrapy. *Scrapy*. Available from: https://scrapy.org/

[6] C. Bizer and R. Cyganiak, “D2R Server – Publishing Relational Databases on the Semantic Web, Poster at the 5th Int. Semantic Web Conference (ISWC), Nov. 2006.

[7] Y. Lee and S. YuXiang, “Hybrid Index Structured on MBB Approximation for Linked Data,” Proc. 10th Int. Conference on Computer Modeling and Simulation, Jan. 2018, pp. 101-104.

[8] Stanford University. *A Free, Open-source Ontology Editor and Framework for Building Intelligent Systems*. Available from: https://protege.stanford.edu/

[9] T. Berners-Lee, Y. Chen, L.Chilton, D. Connolly, R. Dhanaraj, J. Hollenbach, A. Lerer, and D. Sheets, “Tabulator: Exploring and Analyzing Linked Data on the Semantic Web,” Proc. 3rd Int. Semantic Web User Interaction Workshop, Nov. 2006.

[10] C. Becker and C. Bizer, “DBpedia Mobile: A Location-Enabled Linked Data Browser,” Proc. 1st Workshop about Linked Data on the Web (LDOW), Apr. 2008.

[11] M. Dzbor, J. Domingue, and E. Motta, “Magpie – Towards a Semantic Web Browser,” Proc. 2nd Int. Semantic Web Conference (ISWC 2003), Lecture Notes in Computer Science (LNCS), vol. 2870, pp. 690-705, 2003.

[12] G. Tummarello, R. Delbru, and E. Oren, “Sindice.com: Weaving the Open Linked Data,” Proc. 6th Int. Semantic Web Conference (ISWC 2007), Lecture Notes in Computer Science (LNCS), vol. 4825, pp. 552-565, Nov. 2007.

[13] G. Rummarello, R. Cyganiak, M. Catasta, S. Danielczyk, R. Delbru, and S. Decker, “Sig.ma: Live Views on the Web of Data,” Proc. 19th Int. Conference on World Wide Web (WWW 2010), Apr. 2010, pp. 1301-1304.

[14] M. d’Aquin and E. Motta, “Watson, More Than a Semantic Web Search Engine,” Semantic Web Journal, vol. 2, no. 1, pp. 55-63, Jan. 2011.

[15] C. Stadler, J. Lehmann, K. Höffner, and S. Auer, “LinkedGeoData: A Core for a Web of Spatial Open Data,” Semantic Web Journal, vol. 3, no. 4, pp. 333-354, Oct. 2012.

[16] A. D. Mezaour, J. Law-To, R. Isele, T. Schandl, and G. Zechmeister, “Revealing Trends and Insights in Online Hiring Market Using Linking Open Data Cloud: Active Hiring a Use Case Study,” Proc. 11th Int. Semantic Web Conference: Semantic Web Challenge (SWC 2012), submission 9, Nov. 2012.

[17] E. Jung and Y. Lee, “Linked Data Based Storage/Application Platform and Implementation of Analysis System Visualization,” Journal of KIIT, vol. 16, no. 9, pp. 95-102, Sep. 2018.

[18] Y. Lee, “Semantic-based Data Mashups using Hierarchical Clustering and Pattern Analysis Methods,” Journal of Information Science and Engineering, vol. 30, no. 5, pp. 1601-1618, Sep. 2014.

[19] Y. Lee, “Semantic-based Web API Composition for Data Mashups,” Journal of Information Science and Engineering, vol. 31, no. 4, pp. 1233-1248, Jul. 2015.