The 8th International Conference on Information Communication and Technology (ICoICT)

24-26 June 2020

Yogyakarta, Indonesia

Homepage

It is our great pleasure to announce that the 8th International Conference on Information and Communication Technology (ICoICT 2020) will be held in Yogyakarta, Indonesia on June 24 - 26, 2020.
Yogyakarta is the capital city of Indonesia as well as the largest city in the country. It is the cultural and economic hub of Indonesia with widely recognized landmarks including the iconic Borobudur Temple, Prambanan Temple.

ICoICT 2020 will be jointly organized by Telkom University Indonesia, Multimedia University Malaysia, and Universitas Gadjah Mada Indonesia.
The conference offers a good opportunity to enhance international academic exchange on ICT related topics and to provide a platform for researchers to discuss new problems and solutions.
ICoICT 2020 will feature traditional paper presentations, poster presentations, tutorials, demos, as well as keynote speech by renowned educational experts and industrials.

Papers from the previous ICoICT 2013 until 2019 have been published in **IEEE Xplore** and indexed in **Scopus**:

ICoICT 2013: https://ieeexplore.ieee.org/xpl/conhome/6569393/proceeding
ICoICT 2014: https://ieeexplore.ieee.org/xpl/conhome/6908150/proceeding
ICoICT 2015: https://ieeexplore.ieee.org/xpl/conhome/7203317/proceeding
ICoICT 2016: https://ieeexplore.ieee.org/xpl/conhome/7565234/proceeding
ICoICT 2017: https://ieeexplore.ieee.org/xpl/conhome/8054654/proceeding
ICoICT 2018: https://ieeexplore.ieee.org/xpl/conhome/8509820/proceeding
ICoICT 2019: https://ieeexplore.ieee.org/xpl/conhome/8825728/proceeding

The previous ICoICT conferences have successfully served a forum to bring together a diverse group of people from academics and industrial to share and present the latest issues and recent developments in the area of ICT.

2020 8th International Conference on Information and Communication Technology (ICoICT)
Important Dates:

Call for Paper	OCTOBER 1, 2019

Program

Greetings and Video Presentation

Welcoming Speech by General Chair of ICoICT 2020

Speech by Universitas Gadjah Mada

Speech by Multimedia University

Opening Speech by Rector of Telkom University

Performance by Tel-U Virtual Choir

Keynote Speech #1

Image Recognition for Driving Assistance of Intelligent Vehicles

The driver assistance system through image recognition has become important, as the traffic accident rate of the elderly driver increases in Japan. This may also occur in other countries. On the other hand, the accuracy of image recognition has been improved by the development of deep neural networks or other machine learning technologies. However, a large size of learning data is necessary for these methods. First, I will present an efficient enhancement of learning data for in-vehicle camera images. Then, I will introduce several image recognition methods for driving assistance that we have developed so far, such as the recognition of traffic signs, the recognition of walking pedestrian while operating a smartphone, the recognition of the driver's visibility, the weather recognition, and other research topics. In addition, I will talk about the human pose estimation using a super-low-resolution infrared sensor array for the purpose of watching the elderly people at home.

Keynote Speech #2

Principle of Precise GNSS Positioning and Its Applications to Disaster Monitoring Systems in Japan

Recently, a lot of services are provided based on location information based on GNSS (Global

Navigation Satellite Systems). In this presentation, the principle of satellite positioning is reviewed. There exist many methods and algorithms for satellite positioning, very simple and fundamental method as well as precise positioning methods are focused on. In the latter half of the presentation, as case studies of the application of precise satellite positioning, two systems for disaster monitoring and mitigation operated in Japan are introduced. One is the GNSS reference station network (GEONET), and another is the GPS ocean and tsunami monitoring system.

Keynote Speech #3

Dr. Ardhasena is the deputy director for climate and air quality research in the Indonesian Agency for Meteorology Climatology and Geophysics (BMKG). He is an applied nathematician with wide interest of modelling phenomena in nature, such as water waves, optical waves in photonic devices, and climate. He is internationally active in the field of climate related activities under the World Meteorological Organization (WMO) from climate policy related, climate services and standards developments: presently serving as the chair of the working group on Climate Service of Regional Association - V South West Pasific, co-chair of the Expert Team on Global Climate Statement (responsible for the annual global climate statement for the World's policy reference), and a member of the Interprogramme Task Team on Cataloging Extreme Weather Water and Climate Events (IPTT-CWWCE) of the WMO, where within this team, he contributes to the development of a global framework and standardized approach for systematic cataloging of extreme weather and climate events and their unique identification system, which will be useful for the purpose of attribution of disaster losses and damages due to extreme hydrometeorological events.

1B: Climate Change Monitoring

Prediction of Sea Level by Using Autoregressive Integrated Moving Average (ARIMA): Case Study in Tanjung Intan Harbour Cilacap, Indonesia

Yehezkiel Kevin Purba and Deni Saepudin (Telkom University, Indonesia); Didit Adytia (School of Computing, Telkom University, Indonesia)

Wave Prediction by using Support Vector Regression, Study Case in Jakarta Bay

Elizabeth Manurung and Didit Adytia (School of Computing, Telkom University, Indonesia); Nugrahinggil Subasita (NeXT Waves-ID, Indonesia)

Sea Level Prediction by Using Seasonal Autoregressive Integrated Moving Average Model, Case Study in Semarang, Indonesia

Ronald Tulus and Didit Adytia (School of Computing, Telkom University, Indonesia); Nugrahinggil Subasita (NeXT Waves-ID, Indonesia); Dede Tarwidi (Telkom University, Indonesia)

Utilization of Internet of Thing and Social Media in Designing a Smart System for Identification Pollution Quality of Air, Water, and Temperature

Muhardi Saputra, Wahjoe Witjaksono and Warih Puspitasari (Telkom University, Indonesia)

Wave Height Prediction based on Wind Information by using General Regression Neural Network, study case in Jakarta Bay

Vita Juliani and Didit Adytia (School of Computing, Telkom University, Indonesia); A Adiwijaya (Telkom University, Indonesia)

1C: Emerging Trends - Social Media

Recognizing Personality from Social Media Linguistic Cues: A Case Study of Brand Ambassador Personality

Andry Alamsyah, Rafa Bastikarana, Alya Rysda Ramadhanti and Sri Widiyanesti (Telkom University, Indonesia)

Brand Awareness using Network Modeling Method

Muhamad Fulki Firdaus, Z. k. a. Baizal and Kevin Bratawisnu, Made (Telkom University, Indonesia); Hanafi Abdullah Gusman (Telkom Universtiy, Indonesia)

Measuring the effectiveness of social media owned by local government leaders in communicating smart city programs Case study on the Mayor's Bandung Instagram

Grisna Anggadwita (Telkom University, Indonesia); Brady Rikumahu (School of Economics and Business, Telkom University, Indonesia); Ratih Hendayani and Rayhan Raka Putra (Telkom University, Indonesia)

Classifying the Polarity of Online Media on the Indonesian Presidential Election 2019 Using Artificial Neural Network

Muhammad Afif Farisi and Kemas Lhaksmana (Telkom University, Indonesia)

Ensemble Learning in Predicting Financial Distress of Indonesian Public Company

Dyah Sulistyowati Rahayu (University of Pancasila & University of Indonesia, Indonesia); Heru Suhartanto (Universitas Indonesia, Indonesia)

1D: Emerging Trends - Health

ECG Based Biometric Identification System using EEMD, VMD and Renyi Entropy

Sugondo Hadiyoso (Telkom University & Institut Teknologi Bandung, Indonesia); Inung Wijayanto (Telkom University & Universitas Gadjah Mada, Indonesia); Tugas Bme (Politeknik Negeri Bandung, Indonesia)

EEG Signal Classification for Alcoholic and Non-Alcoholic Person using Multilevel Wavelet Packet Entropy and Support Vector Machine

Cahyantari Ekaputri (Telkom University, Indonesia); Rahmat Widadi (Institut Teknologi Telkom Purwokerto, Indonesia); Achmad Rizal (Telkom University, Indonesia)

QSAR Study of Fusidic Acid Derivative as Anti- Malaria Agents by using Artificial Neural Network- Genetic Algorithm

Hamzah Azmi, Kemas Lhaksmana and Isman Kurniawan (Telkom University, Indonesia)

Leveraging Textural Features for Mammogram Classification
Sri Frenzilino Mahayyu Akbarisena, Ema Rachmawati and D Utama (Telkom University, Indonesia)

1E: Emerging Trends - NLP/Language Model

Knowing Opposing Arguments in Persuasive Essays Using Random Forest Classifier
Daulat Rachmanto, Ibnu Asror and Anisa Herdiani (Telkom University, Indonesia)

End-to-End Speech Recognition Models for a Low-Resourced Indonesian Language
Suyanto Suyanto (Telkom University, Indonesia); Anditya Arifianto (Telkom University & Artificial Intelligence Laboratory, ICM Research Group, Indonesia); Anis Sirwan (Divisi Digital Service (DDS) PT Telkom, Indonesia); Angga P. Rizaendra (Divisi IT, PT Telkom Akses, Indonesia)

Syllable-Based Indonesian Lip Reading Model
Adriana Kurniawan and Suyanto Suyanto (Telkom University, Indonesia)

Language Modeling for Journalistic Robot based on Generative Pretrained Transformer 2
Raihan Hamid Suraperwata and Suyanto Suyanto (Telkom University, Indonesia)

Academic Expert Finding in Indonesia using Word Embedding and Document Embedding: A Case Study of Fasilkom UI
Theresia V. Rampisela and Evi Yulianti (Universitas Indonesia, Indonesia)

1F: Emerging Trends - IoT

IoT Object Security towards the Sybil Attack Using the Trustworthiness Management
Ridwan Hadiansyah (Telkom University, Indonesia); Vera Suryani (Universitas Telkom, Indonesia); Aulia Arif Wardana (Telkom University, Indonesia)

IoT Object Security towards On-off Attack Using Trustworthiness Management
Anggi Pratama Nasution (Telkom University, Indonesia); Vera Suryani (Universitas Telkom, Indonesia); Aulia Arif Wardana (Telkom University, Indonesia)

Sustainable Internet of Things: Alignment approach using enterprise architecture
Nunung Nurul Qomariyah (Bina Nusantara University Jakarta, Indonesia); Anjar Priandoyo (University of York, United Kingdom (Great Britain))

LoRaWAN Internet of Things Network Planning for Smart Metering Services in Dense Urban Scenario
Alvin Yusri and Muhammad Imam Nashiruddin (Telkom University, Indonesia)

Designing NB-IoT (Internet of Things) Network for Public IoT in Batam Island
Shelasih Winalisa and Muhammad Imam Nashiruddin (Telkom University, Indonesia)

TS1: Tutorial Session #1

The Fourth Industrial Revolution- A Global Revolution in Science, Technology and Society towards a better life by Assoc.Prof.Dr.MD Shohel Sayeed (Multimedia University, Malaysia)

His core research interest is in the area of Biometrics, big data, cloud computing, artificial intelligence, information security, image and signal processing, pattern recognition and classification. He has published over 60 research papers in international peer-reviewed journals and international conference proceedings as a result of his research work. His research works have been published by high ranked peer- reviewed journals such as IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI), International Journal of Pattern Recognition and Artificial Intelligence (IJPRI), Expert Systems with Applications, Discrete Dynamics in Nature and Society (DDNS) as well as several peer- reviewed International journals. Several of his findings have been presented in a number of well recognized IEEE conferences as well. He has been appointed technical paper reviewer for Journal of Pattern Recognition Letters, IEEE Transaction on Neural Networks, IEEE Transactions on Automation Science and Engineering, Journal of Computer Methods and Programs in Biomedicine and International Journal of Computer Theory and Engineering. He has also been invited to review technical papers for several international conferences. In recognition of his professional contribution, he has obtained recognition as a Senior member of IEEE Computer Society, IEEE Communication Society and International Association of Computer Science and Information Technology (IACSIT).

Dr. Shohel has invited as a Chief Guest and keynote speaker at the second International Conference on Advanced Computing (ICAC 2019). Apart from the ICAC 2019, Dr. Shohel was also invited as the keynote speaker for several international conferences such as the International Conference on Recent Trends and Challenges in Healthcare Informatics (RCHI- 2019), International Conference on Recent Trends in Advanced Computing (ICRTAC 2019) and International Conference on Computational Intelligence and Applications (lCCIA 2019). Furthermore, Dr. Shohel was also invited as the keynote speaker for the International Conference on Modern Research (Multidisciplinary) 2019, International Conference on Advanced Computing (ICAC 2015), International Symposium Innovative Management, Information & Production (IMIP 2015) and International Conference on Innovations in Computer Science and Technology (ICICST 2016), respectively.

2B: Disaster Data Science and Management

Automatic First Arrival Picking on P-Wave Seismic Signal Using Support Vector Machine Method

Muhammad Wahyu Putra Indi, Astri Novianty and Anggunmeka Luhur Prasasti (Telkom University, Indonesia)

6D: Emerging Trends - UX

User Interface Design Of Learning Applications For Balinese Traditional Dance using Goal-Directed Design

Ni Nyoman Sri Ayu Asvini Dyatmika, Danang Junaedi and Veronikha Effendy (Telkom University, Indonesia)

Predicting Users' Revisitation Behaviour Based on Web Access Contextual Clusters

Hapnes Toba, Christopher Starry Jomei and Lotanto Setiawan (Maranatha Christian University, Indonesia); Oscar Karnalim (University of Newcastle, Australia & Maranatha Christian University, Indonesia); Hui Li (Guizhou Univrsity, China)

Deep Analysis for Smartphone-based Human Activity Recognition

Chew Yong Shan, Ying Han Pang and Shih Yin Ooi (Multimedia University, Malaysia)

The Analysis of User Intention Detection Related to Conventional Poster Advertisement by Using The Features of Face and Eye(s)

Yolanda Modesty and Dodi Wisaksono Sudiharto (Telkom University, Indonesia); Catur Wirawan Wijiutomo (Telkom University & Institut Teknologi Bandung, Indonesia)

Impact Evaluation of Procedurally Content Generated Against Immersion Games Using ANOVA

Hardianto Wibowo, Dimas Nurpratama, Wildan Suharso, Agus Eko Minarno and Galih Wasis Wicaksono (Universitas Muhammadiyah Malang, Indonesia); Dani Harmanto (De Montfort University, United Kingdom (Great Britain))

6E: Emerging Trends - Telecommunication Network

Outdoor to Indoor Propagation Model of Glass Material Building at 26 GHz for 5G Mobile Technology

Trivia Anggita (University Indonesia, Indonesia); Muhammad Suryanegara (Universitas Indonesia, Indonesia)

Optimizing BTS Placement Using Hybrid Evolutionary Firefly Algorithm

Dzakyta Afuzagani and Suyanto Suyanto (Telkom University, Indonesia)

Deterministic Approach of Indoor Room THz Multipath Channel Model

Dwi Cahyono (AKKA EMC / Mercedes Benz Technology (MBTech) GmbH, Germany); Fawad Sheikh (Universität Duisburg-Essen & The Mobile Terahertz Company UG, ID4US GmbH, Germany); Thomas Kaiser (Universität Duisburg-Essen, Germany)

A Comprehensive Survey of Cellular Network Performance from User's Perspective: A Case Study in 0-km Spot of Yogyakarta

Widyasmoro Widyasmoro and Indar Surahmat (Department of Electrical Engineering, Universitas Muhammadiyah Yogyakarta, Yogyakarta, Indonesia)

An Evidence-Based Technical Process for OpenFlow-Based SDN Forensics
Satria Akbar Mugitama, Niken Cahyani and Parman Sukarno (Telkom University, Indonesia)

TS6: Tutorial Session #6

Deep Learning and Its Applications in Video Analytics by Dr. Lim Kian Ming (Multimedia University, Malaysia)

Deep learning has recently achieved very promising results in a wide range of areas such as computer vision, natural language processing and speech recognition. In contrast to hand-crafted methods, deep learning aims to learn hierarchical representations from large-scale data (e.g. images and videos) via deep architecture models with multiple layers of non-linear transformations. As compared to hand-crafted features, it is easier to achieve higher performance using the learned hierarchical representations. The principle behind the success of deep learning is that it is able to extract different levels of abstractions embedded in the data by carefully design the layer's depth and width. Then, the features that are beneficial for the learning tasks are properly selected.

In recent years, video analytics has attracted increasing interests from both academic and industry. The main goal of video analytics is to automatically recognize the temporal and spatial events in videos. Thanks to the enormous advancements achieved in deep learning, recent improvements in video analytics, ranging from the applications of object tracking, object detection, human-computer interaction, and video surveillance; has automated many tasks in the industries.

This tutorial commences with the concept of deep learning and followed by its applications in video analytics.

Greetings

ICoICT 2021 presentation

Closing by Dean of School of Computing, Telkom University

Performance by Tel-U Virtual Choir


2020 8th International Conference on Information and Communication Technology (ICoICT)

Organized by Telkom University, Multimedia University & Gajah Mada University

Committees

Steering Committee

A Adiwijaya (Telkom University, Indonesia)
Ahmad Rafi (Multimedia University, Malaysia)
Afizan Azman (Melaka International College of Science and Technology Malaysia)
Ari Moesriami Barmawi (Telkom University, Indonesia)
Hairul A. Abdul-Rashid (Multimedia University, Malaysia)
Siong Hoe Lau (Multimedia University, Malaysia)
Maman Abdurohman (Telkom University, Indonesia)
Parman Sukarno (Telkom University, Indonesia)
Rina Pudjiastuti (Telkom University, Indonesia)
Shafinar Ismail (Universiti Teknologi Mara, Malaysia)
Syed Abdul Rahman Al Haddad (Universiti Putra Malaysia, Malaysia)
Kiki Maulana Adhinugraha (La Trobe University, Australia)
Sultan Alamri (SEU, Saudi Arabia)

Conference Committee

General Chair

Ema Rachmawati (Telkom University, Indonesia)

General Co-Chair

Warih Maharani (Telkom University, Indonesia)
Ong Thian Song (Multimedia University, Malaysia)

Technical Program Chair

Bayu Erfianto (Telkom University, Indonesia)

Track Chair

Seno Adi Putra, SSi MT (Telkom University, Indonesia)
Putu Harry Gunawan (Telkom University, Indonesia)
Tee Connie (Multimedia University, Malaysia)
Ying Han Pang (Multimedia University, Malaysia)
Raden Sumiharto (Universitas Gadjah Mada, Indonesia)
Mardhani Riasetiawan (Universitas Gadjah Mada, Indonesia)
Ade Romadhony (Telkom University, Indonesia)

Secretariat Chair

Siti Karimah (Telkom University, Indonesia)
Shih Yin Ooi (Multimedia University, Malaysia)
Rita Rismala (Telkom University, Indonesia)

Publication Chair

Dawam Dwi Jatmiko Suwawi (Telkom University, Indonesia)
Anditya Arifianto (Telkom University, Indonesia)

Finance Chair

Annisa Aditsania (Telkom University, Indonesia)
Siew Chin Chong (Multimedia University, Malaysia)
Siti Sa'adah (Telkom University, Indonesia)

Event and Logistic Chair

Fazmah Arif (Telkom University, Indonesia)
Mira Sabariah (Telkom University, Indonesia)
Prati Gani (Telkom University, Indonesia)

Public Relation Chair

Z. k. a. Baizal (Telkom University, Indonesia)
Mohd Fikri Azli Abdullah (Multimedia University, Malaysia)

CFP Chair

Wikky Fawwaz Al Maki (Telkom University, Indonesia)

Tutorial and Special Session Chair

Didit Adytia (Telkom University, Indonesia)

Sponsorship Chair

Kemas Lhaksmana (Telkom University, Indonesia)

Webmaster

Rahmat Yasirandi (Telkom University, Indonesia)
Yusza Reditya Murti (Telkom University, Indonesia)

Technical Session Committee

Ade Romadhony (Telkom University, Indonesia)
Agung Toto Wibowo (Telkom University - Indonesia, Indonesia)
Agus Harjoko (Universitas Gadjah Mada, Indonesia)
Angelina Prima Kurniati (Telkom University, Indonesia)
Didit Adytia (School of Computing, Telkom University, Indonesia)
Hilal H. Nuha (Telkom University, Indonesia)
Idham Ananta (Universitas Gadjah Mada, Indonesia)
Isman Kurniawan (Telkom University, Indonesia)
Kemas Wiharja (Telkom University, Indonesia)
Mardhani Riasetiawan (Universitas Gadjah Mada, Indonesia)
Niken Cahyani (Telkom University, Indonesia)
Ong Thian Song (Multimedia University, Malaysia)
Parman Sukarno (Telkom University, Indonesia)
Putu Harry Gunawan (Telkom University, Indonesia)
Raden Sumiharto (Universitas Gadjah Mada, Indonesia)
Rendi Yusuf Azhari (Universitas Gadjah Mada, Indonesia)
Risnandar Risnandar (Research Center for Informatics, Indonesian Institute of Sciences, Indonesia)
Shih Yin Ooi (Multimedia University, Malaysia)
Siew Chin Chong (Multimedia University, Malaysia)
Siti Zainab (Universiti Islam Antarabangsa Malaysia, Malaysia)

Tee Connie (Multimedia University, Malaysia)
Wahyono Wahyono (Universitas Gadjah Mada, Indonesia)
Wikky Fawwaz Al Maki (Telkom University, Indonesia)
Ying Han Pang (Multimedia University, Malaysia)
Yunita Sari (Universitas Gadjah Mada, Indonesia)


2020 8th International Conference on Information and Communication Technology (ICoICT)

Organized by Telkom University, Multimedia University & Gajah Mada University
Prepared by EDAS Conference Services.

Paper Submission Deadline	FEBRUARY 1, 2020
Extended Paper Submission Deadline	FEBRUARY 29, 2020
Notification of Papes Acceptance	MARCH 15, 2020
Submission of Camera Ready Papers and Author Registration Deadline	APRIL 30, 2020
Conference Date	JUNE 24-26, 2020

The previous ICoICT conferences have successfully served a forum to bring together a diverse group of people from academics and industrial to share and present the latest issues and recent developments in the area of ICT. IColCT 2020 is co-sponsored by IEEE Indonesia Section and IEEE Signal Processing Society Indonesia Chapter. All accepted papers in ICoICT 2020 will be published in the conference proceedings and will be submitted for publication.

We look forward to seeing you in Yogyakarta!

2020 8th International Conference on Information and Communication Technology (ICoICT)

Media Type	Part Number	ISBN
XPLORE COMPLIANT	CFP201CZ-ART	978-1-7281-6142-6
USB	CFP17ICZ-USB	978-1-7281-6141-9

Co-sponsored by

Organized by Telkom University, Multimedia University, & Gajah Mada University
Prepared by EDAS Conference Services.

Predicting Users' Revisitation Behaviour Based on Web Access Contextual Clusters

Hapnes Toba
Faculty of Information Technology
Maranatha Christian University
Bandung 0, Indonesia
hapnestoba it.maranatha.edu

Christopher Starry Jomei
Faculty of Information Technology
Maranatha Christian University
Bandung 0, Indonesia
christopherstarry gmail.com

Lotanto Setiawan
Faculty of Information Technology
Maranatha Christian University
Bandung 0, Indonesia
lo.tanto yahoo.com

Oscar Karnalim
Faculty of Engineering and Built Environment
University of Newcastle
Callaghan, Australia
oscar.karnalim uon.edu.au

Hui Li
College of Computer Science and Technology
Guizhou University
Guiyang, China
cse.huili gzu.edu.cn

***Abstract*— ost modern bro sers record all previousl visited eb pages for future revisitation o ever not all users utilise such feature ne of the reasons is that the records are displa ed at once as a single list hich ma over helm the users This paper proposes a predictive model to decide hether a eb page ill be revisited in the future based on a particular visit The model can be used to filter eb records so that onl eb pages that ma be re-visited are presented ccording to our evaluation the model is considerabl effective t can generate accurac hen measured ith -fold cross validation and meaningful topic identification Further attributes rooted from the same website' access frequency are the most salient ones for prediction n addition conte tual similarities based on k-means clustering and cosine similarit hich are used for defining some attributes are considerabl effective**

Keywords—browsing history, cosine similarity, k-means clustering, multinomial naïve bayes, user log

I. INTRODUCTION

The World Wide Web (WWW) is an ecosystem where individuals can satisfy their information and communication needs . The importance of this ecosystem is growing rapidly as the number of users is e tremely high. In 20 alone, about , 8 million people accessed the ecosystem 2. Conse uently, there are a lot of research related to the ecosystem, which mainly focuses on Information Retrieval . As the users become more familiar with WWW, they tend to access web pages more fre uently . Most modern browsers (such as Google Chrome, Mozilla Firefo , and Microsoft Edge) record all previously visited web pages so that the users can re-visit them easier. However, the records are rarely used based on two reasons . First, the records are not directly visible from the navigation panel. Second, the records are displayed in a way that does not engage the users numerous web pages are presented at once as a single list, which may overwhelm the users.

For the latter issue, some works have been proposed to solve it. They commonly re-visualise or rank the records so that the users can be engaged easily , -0 However, to our knowledge, most e isting works do not try to directly filter the web pages. Only web pages which may be accessed further are kept on the list. We would argue such filtering mechanism may help the users to find some web pages for revisitation as the number of displayed web pages is reduced. It is important to note that the filtering mechanism can also be used together with the e isting works (i.e., revisualising or ranking the records). The web pages are filtered first before passed to the e isting works as the inputs. In order to develop the filtering mechanism, this paper proposes a predictive model to decide whether a web page may be re-visited in the future based on a particular visit.

The model is formed using Multinomial Na ve Bayes , with twenty-four attributes on board. The first eight attributes are based on the access fre uency of the same website while the remaining attributes are based on the access fre uency of websites with the same conte t, *i.e.* the conte tual similarity is defined with either K-Means Clustering or cosine similarity . Na ve Bayes is known for its good performance in dealing with many problems, including the new ones like ours. For some cases, it is even more effective than some advanced algorithms such as Artificial Neural Network. We chose the multinomial one as some of the attributes are numeric. Other algorithms are selected as they are often used as the baselines and we plan to set this work as a baseline for other further research.

Our proposed predictive model may also benefit the website owners. If an owner knows that a user may not re-visit their website anymore, they can attract the user to engage again with a kind of discount (for e-commerce website) or Easter egg (*i.e.* unexpected thing that attracts the user's interest). It is true that this kind of event can also be discovered with server log, checking which users have not accessed the website for a long time. However, we would argue that our proposed mechanism is less time-consuming, as the event can be predicted right after the user has visited the website for the last time. Search engines may also benefit from our predictive model. They may e clude some web pages from their suggestions, which make their computation for web page recommendation faster and time-efficient.

II. RELATED WORK

As the need of internet grows rapidly, web user behaviour has been researched in many works while some of them rely on search engine ueries. For e ample, a work in 2 analyses user ueries from E cite, a search engine, to gain insight about web user behaviour. Another e ample is a work in which discovers how user searches images through their ueries. Several works are more focused on ueries created by a

-- - -- .00

specific group of people. The group can be either young users [illegible] or K-12 students [illegible]. The former group is analysed to quantify any difficulties experienced by young users when interacting with the internet while the latter is analysed to gain an insight about student learning behaviour.

Instead of search engine queries, several research works try to contribute on other technical aspects. First, a work in [illegible] proposes a combination of web application models with runtime navigation logs for understanding user behaviour further. Second, a work in [illegible] introduces SMAP-Mine, a data mining algorithm to efficiently predict user behaviour in a mobile web system. Third, a work in [18] develops visual snippet, a compact representation that combines text snippet's and thumbnail's benefits for finding relevant web pages. Fourth, a work in [illegible] establishes an algorithm for a sparse non-negative matrix factorisation, that is specifically designed to discover web user behaviour. The research on web user behaviour contributes to the growth of other fields. It fasten the research about search engine ranking [20], spam identification [21], user navigation problem [11], the dwell time of web pages [22], and web revisitation.

Web revisitation occurs when an user accesses a previously-visited web page for various reasons [11]. This task can be demanding as the user may not remember the page's URL and the number of previously-visited web pages (accessed through web browsing history) can be extremely high. Hence, several works try to alleviate the burden by either re-visualising or ranking the content of the browsing history. Revisualising is applied by replacing the conventional list-based browsing history with a more intuitive representation. A work in [11], for instance, provides an interactive visualisation. Another work in [10], proposes a graph-based visualisation of the browsing history based on its sessions.

Ranking is applied with the help of Information Retrieval. A work in [11] combines ranking and clustering mechanism to predict what web pages may be revisited. A work in [8] utilises context and content keywords to suggest web revisitation with the help of relevance feedback. Not all works about web revisitation are about providing better web browsing history. Several works act as a support to the task. A work [23] discovers four web revisitation patterns through a combination of observing web access log and survey. A work in [24] introduces a new metric for web revisitation as a result of considering tabbed browsing.

To the best of our knowledge, most works have not applied a filtering mechanism to the browsing history (keeping only the most potential pages that will be revisited). We believe that the mechanism may be helpful as the number of web pages is reduced, resulting better revisualising and ranking the content of the browsing history.

III. Methodology

Our work proposes a predictive model to determine the potentiality of a web page to be re-visited. The model can be used to filter user browsing history so that only web pages that would be re-visited are kept. Multinomial Naïve Bayes [illegible] is used as the predictive model's learning algorithm; Naive Bayes (the core model of Multinomial Naive Bayes) is considerably effective in most dataset due to its independence assumption [25]. The algorithm has been applied on many fields such as pharmaceutical industry [26], nuclear power plant [27] and text analysis (either general as in [25] or academic as in [28]). The Multinomial Naïve Bayes is implemented with by utilizing Weka [29]).

The predictive model relies on twenty-four attributes, which are classified further to three groups. The first group (A1-A8) is derived from the access frequency of the same website. The second group (B1-B8) is derived from the access frequency of websites with the same cluster-based context wherein the cluster-based context is determined through K-Means Clustering [illegible]. The third group (C1-C8) is derived from the access frequency of websites with the same cosine-based context wherein the cosine-based context is defined using cosine similarity [illegible].

TABLE I. The Attributes of Proposed Predictive Model

[illegible]	Definition
A1	Delta value (in minutes) between the current and previous access time where both currently- and previously-accessed website are the same
A2	Delta value (in minutes) between the current and previous access time where both currently- and previously-accessed website are different
A3	Average delta time (in minutes) between five latest access times where currently-accessed website and previously-accessed websites are the same
A4	Average delta time (in minutes) between five latest access times where currently-accessed website and previously-accessed websites are different
A5	The proportion of accessed websites with the same name as the currently-accessed one from 2 hours before
A6	The proportion of accessed websites with different name as the currently-accessed one from 2 hours before
A7	The proportion of accessed websites with the same name as the currently-accessed one from 1 days before
A8	The proportion of accessed websites with different name as the currently-accessed one from 1 days before
B1	Delta value (in minutes) between the current and previous access time where both currently- and previously-accessed website share the same cluster-based context
B2	Delta value (in minutes) between the current and previous access time where both currently- and previously-accessed website share different cluster-based context
B3	Average delta time (in minutes) between five latest access times where currently-accessed website and previously-accessed websites share the same cluster-based context
B4	Average delta time (in minutes) between five latest access times where currently-accessed website and previously-accessed websites share different cluster-based context
B5	The proportion of accessed websites with the same cluster-based context as the currently-accessed one from 2 hours before
B6	The proportion of accessed websites with different cluster-based context as the currently-accessed one from 2 hours before
B7	The proportion of accessed websites with the same cluster-based context as the currently-accessed one from 1 days before
B8	The proportion of accessed websites with different cluster-based context as the currently-accessed one from 1 days before
C1	Delta value (in minutes) between the current and previous access time where both currently- and previously-accessed website share the same cosine-based context
C2	Delta value (in minutes) between the current and previous access time where both currently- and previously-accessed website share different cosine-based context
C3	Average delta time (in minutes) between five latest access times where currently-accessed website and previously-accessed websites share the same cosine-based context
C4	Average delta time (in minutes) between five latest access times where currently-accessed website and previously-accessed websites share different cosine-based context
C5	The proportion of accessed websites with the same cosine-based context as the currently-accessed one from 2 hours before
C6	The proportion of accessed websites with different cosine-based context as the currently-accessed one from 2 hours before
C7	The proportion of accessed websites with the same cosine-based context as the currently-accessed one from 1 days before

□□	□efinition
C8	The proportion of accessed websites with different cosine-based conte□t as the currently-accessed one from □days before

Cluster-based conte□tual similarity is defined based on web page clusters provided beforehand. Fig. □shows how the clusters are formed with web page links as the input. Each link's web page is crawled with JSoup □□0□, and tokenised (with all stop words are removed). Since the methodology will be evaluated on Chinese users, only Chinese characters are considered as terms and the stop words are taken from □□□□. These terms are therefore inde□ed as term-fre□uency tuples (where the frequency refers to the term's occurrence fre□uency on a particular web page) and used to form similarity vectors on clustering. The clustering itself is performed with K-Means Clustering □□□□(where K is assigned with □0 and the implementation is based on Weka □2□□).

Two websites are considered to have the same cluster-based conte□tual similarity if, given a number of web page clusters, their links are more fre□uently occurred on web pages which fall on the same clusters than different clusters. Fig. 2 shows how the similarity is calculated.

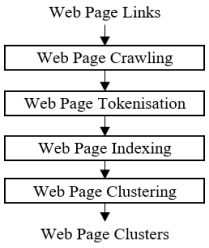

Fig. □ How web page clusters are formed

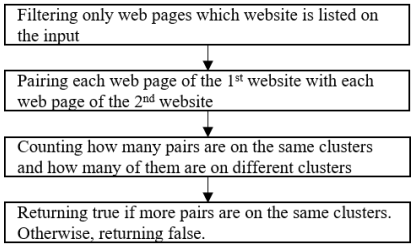

Fig. 2. Defining cluster-based conte□tual similarity between two websites

To illustrate this, let us assume we have two websites (*dom1.com* and *dom2.com*) and their cluster-based conte□tual similarity will be calculated based on two web page clusters (*cluster-1* and *cluster-2*)□ *cluster-1* has three web pages (*dom1.com/a1, dom2.com/b1, & dom3.com/c1*) and *cluster-2* has four web pages (*dom1.com/d2*, *dom1.com/e2, dom3.com/f2,* and *dom4.com/g2*). At first, all unrelated web pages are removed. Conse□uently, cluster-□ and cluster-2 have only two web pages each (dom□com□a□& dom2.com□b□ for cluster-□and dom□com□d2 & dom□com□e2 for cluster-2). Secondly, each web page of the □st website (dom□) will be paired with each web page of the 2nd website (dom2). It results in three pairs□ dom□com□a□ with dom2.com□b□, dom□com□d2 with dom2.com□b□, and dom□com□e2 with dom2.com□b□.

Thirdly, the number of pairs which member fall on the same cluster and different clusters are calculated. In our case, two pairs have their members on different clusters (dom□com□d2 with dom2.com□b□ and dom□com□e2 with dom2.com□b□) and a pair has their members on the same cluster (dom□com□a□with dom2.com□b□). Fourthly, since the number of pairs which members fall on different clusters is higher, dom□com is considered as different to dom2.com based on cluster-based conte□tual similarity. This similarity function would return false.

Considering a large dataset may not be able to be clustered at once due to memory limitation, our cluster-based conte□tual similarity measurement can be tuned to accept multiple cluster groups at once (where each group refers to a result from an e□ecution of clustering algorithm). In such manner, the similarity result will be defined through voting mechanism□a similarity result with higher occurrence fre□uency among groups will be considered as the final result. For e□ample, if there are five cluster groups and three of them consider given two websites are the same, it can be concluded that those two websites share the same cluster-based conte□tual similarity according to those groups.

Cosine-based conte□tual similarity is defined based on predefined conte□tual groups. Two websites are considered similar if their links fall on the same conte□tual group when measured using cosine similarity □□□. It is important to note that the links are compared instead of the web pages since comparing web pages is time-inefficient.The conte□tual groups should depict various topics. Hence, in our case, these groups are defined from various software descriptions on Wikipedia □□2□. One hundred most fre□uently-accessed software are selected for this task (where, in our case, the fre□uencies are calculated based on user software access log from our evaluation dataset). Since given website links are commonly written with English characters, the software descriptions are limited to English-written descriptions.

Collected software descriptions are then tokenised, inde□ed, and clustered to conte□tual groups. The tokenisation works by splitting the descriptions with non-alphanumeric characters and removing all stop words □□□□. The inde□ing converts resulted terms to term-fre□uency tuples where the frequency is the term's occurrence frequency toward a particular description. The clustering categorises these descriptions to five conte□tual groups with the help of K-Means Clustering □□□□. The conte□tual group for each given website link is determined in twofold. At first, the link is tokenised and inde□ed in the same way as software descriptions are tokenised and indexed. Later, the link's index will be compared to each conte□tual group's index and a group which leads the highest cosine similarity will be considered as the conte□tual group of the link □□□.

I□. E□ALUATION AND DISCUSSION

This section is divided to seven sub-section. The first sub-section will discuss about evaluation dataset. Remaining sub-sections will discuss about specific evaluation mechanisms and their results. One of them is about cluster-based conte□tual similarity□ one of them is about cosine-based conte□tual similarity□and the last four are about our prediction model for web revisitation.

A. Evaluation Dataset

Our evaluation dataset was e□tracted from software and web access logs of □80 Chinese users for four weeks. The access logs were recorded using a dedicated monitoring application installed on their laptop □computer. In total, there are □,□□2,□8□ software access entries and □,2□0,□□0 web access entries. Each entry has three fields□user ID, access timestamp, and accessed software □web page link. For referencing purpose, this dataset will be labelled as main dataset. In addition to main dataset, another dataset called cosine dataset is also used in this evaluation. As its name states, it is dedicated for evaluating cosine-based conte□tual similarity. Thirty web page pairs were selected by the second author of this paper□two thirds of them share the same topic on their respective web page contents while the others do not.

B. The Cluster Descriptiveness of Cluster-based Contextual Similarity

Clusters resulted from our proposed techni□ue for determining cluster-based conte□tual similarity are considered descriptive if their respective terms either have a meaning or lead to a particular topic. To determine how many cluster terms are meaningful, all web page links from the main dataset are grouped per one thousand links wherein each group are categorised further to ten clusters with our proposed clustering techni□ue. In total, there are □□8 groups with □□80 clusters. A cluster will be selected randomly for each group, and the meaning of its most fre□uently-occurred term will be rated by two e□aminers. The e□aminers have three options to choose on□ meaningful, not meaningful, and undecidable. If the e□aminers share different perspectives toward some cluster terms, a discussion will be performed till a consensus is met.

Our e□periment reveals that the proportion of meaningful cluster terms (□0.8□□) is higher than that proportion for meaningless cluster terms (22.88□). Hence, it can be stated that our proposed techni□ue may lead to descriptive clusters. Several cluster terms (26.27%) are still on "undecidable" status since these terms are either overly general or have multiple interpretations.To determine whether terms on a cluster may lead to a more-general topic, the first three thousand web pages from the main dataset are divided to three groups (with one thousand pages per group) wherein each group's pages are categorised to ten clusters with our proposed techni□ue. For each cluster, we define a general term based on its top-□0 fre□uently-occurred words, and we could also able to define the topic of □□□ clusters in Chinese. In other words, our proposed techni□ue may lead to descriptive clusters□each cluster has its own specific topic.

C. The Accuracy of Cosine-based Contextual Similarity

Cosine-based conte□tual similarity is accurate if it can distinguish web page topic as human does. In this conte□t, cosine dataset is used where the accuracy is defined based on how many web page pairs are detected to have the same result as human judgement. Considering cosine similarity returns a proportion degree while human judgement returns a boolean value, cosine similarity's result will be discretised to the boolean one. The proportion degree that is higher or e□ual to □□□ will be converted to true (which means given two web pages share the same topic). Otherwise, it will be converted to false. Our e□periment shows that cosine similarity can correctly predict the topics of □□□ web page pairs (2□□ for true positive and 2□□ for true negative). In other words, cosine-based conte□tual similarity may be considerably accurate□its prediction is higher than □0□.

D. The Effectiveness of The Same Website's Access Frequency for Predicting Web Revisitation

Eight attributes used for predicting web revisitation (A□-A8 from Table 1) are rooted from the same website's access fre□uency. In this section, the effectiveness of those attributes will be evaluated. A prediction model based only on those attributes will be built and evaluated toward web access log entries from our main dataset (which are □,2□0,□□0 entries in total). An entry indicates that a web page may be re-visited if at least one entry with the same user accesses the same website at later timestamp. Otherwise, it indicates no revisitation. In the dataset, the number of web access entries for revisitation is significantly higher than that number for no revisitation since, per user, a particular website was usually accessed more than once and only one of those indicates no revisitation (the last access). Therefore, to balance the dataset, only N□2th web access entries are considered for revisitation (where N refers to the number of web access entries for a user toward a particular website). This phase results in □8,□00 entries in total. Half of them indicates web revisitation while another half indicates no web revisitation. For convenient referencing, these entries will be labelled as prediction dataset in the rest of this paper.

An attribute is considered effective for predicting web revisitation if its ine□istence leads to a huge accuracy reduction □□□□. The accuracy will be defined through □0-fold cross validation (□0 is selected due to its popularity as a threshold for k-fold cross validation □□□□). Prior determining the accuracy reduction, an accuracy when A□-A8 are used together should be measured as the baseline. On our dataset, the accuracy is □□2□□□, which is ade□uately acceptable. The accuracy reduction for each attribute can be seen on Table 2. Each value is calculated by subtracting the baseline accuracy with an accuracy where given attribute is e□cluded. A□ (average delta time in minutes between five latest access times where currently-accessed website and previously-accessed websites are the same) is the most significant one for predicting web revisitation. It gains the largest reduction (□□□8□) compared to other attributes. A□ (delta value between the current and previous access time where both currently- and previously-accessed website are the same), on the contrary, should not be used since its ine□istence leads to higher accuracy and the largest negative reduction. A□and A□ should not also be used as they still lead to negative reduction (even though it is not as significant as A□).

TABLE II. ACCURACY REDUCTION FOR ATTRIBUTES FROM THE SAME WEBSITE'S ACCESS FRE□UENCY

□ttribute □□	□ccurac□Reduction □□ □
A□	-0.00□
A2	0.0□□
A□	□.□□8
A□	-0.0□□
A□	-0.00□
A□	-0.00□
A□	0.000
A8	0.000

Among those attributes, A□ and A8 are two attributes which ine□istence does not affect resulted accuracy. One of the possible reasons is that these attributes need longer data collection period to show their impact (since their value is calculated for □ days and our whole data collection period is only 28 days).When each of those attributes are correlated with target classes (using correlation-based feature selection from Weka □2□), that is rooted from Pearson correlation □□□□, only A□and A□indicate a non-zero correlation degree (which is 0.00□□ and 0.00□2 respectively). Hence, from feature selection perspective, it can be stated that both attributes may be effective for predicting web revisitation (even though A1's impact is still need to be observed further as its ine□istence leads to higher accuracy).

E. The Effectiveness of The Same Website's Access Frequency for Predicting Web Revisitation

Attributes that are related to the access fre□uency of websites with the same cluster-based conte□t (B□-B8 from Table □) are evaluated in this section. The evaluation will be conducted in a similar manner as evaluating A□-A8 (with the

prediction dataset as instances). An attribute is considered effective if its ine□istence results in significant accuracy reduction □□□□when measured with □0-fold cross validation □□□□. When B□-B8 are used altogether to form a predictive model for web revisitation, the model leads to □0.02□□ accuracy. It is lower than the baseline accuracy of A□-A8 (□□2□□□). This phenomenon may be caused by at least one of two reasons. First, B□-B8 may be less effective than A□-A8 for predicting web revisitation. Second, the currently-used techni□ue for determining cluster-based conte□tual similarity may be ineffective when applied on the dataset. Table □shows the accuracy reduction for each attribute. B2 (delta value between the current and previous access time where both currently- and previously-accessed website share different cluster-based conte□t) is the only attribute which leads to a positive accuracy reduction. Other attributes lead to either zero impact (B□, B□, and B□) or negative reduction (B□, B□, B□, and B8).

TABLE III. ACCURACY REDUCTION FOR ATTRIBUTES FROM THE ACCESS FRE□UENCY OF WEBSITES WITH THE SAME CLUSTER-BASED CONTE□T

□ttribute □□	□ccurac□ Reduction □□ □
B□	0.000
B2	0.□□□
B□	0.000
B□	-0.0□□
B□	-0.008
B□	0.000
B□	-0.008
B8	-0.008

When correlated with target classes (using correlation-based feature selection from Weka □2□□), all of these features generate zero correlation degree. In other words, from feature perspective, none of these attributes may be effective for predicting web revisitation.

F. The Effectiveness of The Access Frequency of Websites with the Same Cosine-based Context for Predicting Web Revisitation

Attributes rooted from the access fre□uency of websites with the same cosine-based conte□t (C□-C8 from Table □) will be evaluated in this section. The evaluation will be conducted in a similar manner as evaluating A□-A8 and B□-B8, involving the prediction dataset as instances. An attribute is considered effective if its ine□istence results in significant accuracy reduction □□□□when measured with □0-fold cross validation □□□□. The baseline accuracy is defined by utilising C□-C8 altogether to form a model and evaluate the model with □0-fold cross validation. The resulted accuracy is □□.□□□□, which is lower than the baseline accuracies for A□-A8 (□□2□□□) and B□-B8 (□0.02□□). In other words, C□-C8 may be less effective than other attribute categories. Further, proposed cosine-based conte□tual similarity may be not accurate for determining similarity. This finding is supported by the fact that it only leads to □□□ accuracy when compared to human judgment.

TABLE I□. ACCURACY REDUCTION FOR ATTRIBUTES FROM THE ACCESS FRE□UENCY OF WEBSITES WITH THE SAME COSINE-BASED CONTE□T

□ttribute □□	□ccurac□ Reduction □□ □
C□	0.0□□
C2	-0.□□□
C□	0.0□□
C□	-0.□□□□
C□	0.0□□
C□	-0.□□□
C8	0.0□2

As seen on Table □, C□(delta value between the current and previous access time where both currently- and previously-accessed website share the same cosine-based conte□t) and C□(average delta time between five latest access times where currently-accessed website and previously-accessed websites share the same cosine-based conte□t) generate the largest positive reduction, followed by C□, C□, and C8 respectively. Only three attributes yield negative reduction□C2, C□, and C□. It can be therefore stated that, according to accuracy reduction, all attributes e□cept C2, C□, and C□may be effective for predicting web revisitation. It is important to note that the previous finding is not aligned with our finding from correlation-based feature selection from Weka □2□□. When correlated with their target classes, all attributes show no correlation.

G. The Effectiveness of Proposed Model for Predicting Web Revisitation

This section aims to evaluate the overall effectiveness when all attributes are used together. With □0-fold cross validation on board, these attributes lead to □□.□□□□ accuracy when evaluated toward the prediction dataset. It is higher than the baseline accuracies of B□-B8 (□0.02□□) and C□-C8 (□□.□□□□). However, it is still slightly lower than the baseline accuracy of A□-A8 (□□2□□□). To sum up, according to our dataset, only attributes rooted from the same website's access fre□uency are enough to generate the highest accuracy. Other kinds of attributes may worsen the result even though the reduction is insignificant.

□. CONCLUSION

This research proposes a predictive model to determine whether a website will be revisited by a particular user. The model relies on 2□ attributes which are grouped to three categories. The first group is derived from the access fre□uency of the same website□the second group is derived from the access fre□uency of websites with the same cluster-based conte□t□and the third group is derived from the access fre□uency of websites with the same cosine-based conte□t.

According to our evaluation, four findings can be concluded. First, the predictive model is considerably accurate. It generates □□.□□□□ accuracy when all attributes are involved. Second, attributes derived from the access fre□uency of the same website may be potential for predicting web revisitation. These attributes alone generate □□2□□□ accuracy, which is slightly higher than the accuracy of all attributes. Third, cluster-based conte□tual similarity (that is used for defining attributes) generates considerably descriptive clusters. The clusters have □0.8□□ meaningful terms while the topics of □□□ clusters can be defined based on their top-□0 fre□uently-occurred words. Fourth, cosine-based conte□tual similarity (that is used for defining attributes) is considerably effective. It generates □□□ accuracy. These findings confirmed the effectiveness of our approach. As a comparison, some problems, such as keyphrase e□traction often have lower accuracy (smaller than □0□) □28, □□-□□□.

For future works, we plan to incorporate more-advanced clustering algorithm on cluster-based conte□tual similarity for higher accuracy via a comparative study. Further, we also plan to implement this predictive model as a browser plug-in so that it can be used to filter web browsing history.

ACKNOWLEDGMENT

The authors would like to thank Guizhou University, for providing the dataset and Maranatha Christian University, for providing the international student e□change grants.

REFERENCES

□□□ M. Vigo and S. Harper, "Real-time detection of navigation problems on the World 'Wild' Web," *International Journal of Human-Computer Studies*, vol. □0□, pp. □–□, 20□□.

□2□ J. Clement, "Number of internet users worldwide 2005-2018," *Statista*, 0□-Jan-2020. □Online□. Available□ https□□www.statista.com□statistics□2□□0□8□number-of-internet-users-worldwide□. □Accessed□□□-Feb-2020□.

□□□ B. Croft, D. Metzler, and T. Strohman, *Search engines: Information retrieval in practice*. Boston□Pearson, 20□0.

□□□ W. Du, Z. □ian, P. Parsons, and Y. Chen, "Personal Web Library: Organizing and Visualizing Web Browsing History," *International Journal of Web Information Systems*, pp. 2□2–2□2, Feb. 20□8.

□□□ S. S. Won, J. Jin, and J. I. Hong, "Contextual web history," Proceedings of the 2□th international conference on Human factors in computing systems - CHI 0□, 200□.

□□□ M. □. Kleek, B. Moore, C. □u, and D. R. Karger, "Eyebrowse," Proceedings of the 28th of the international conference e□tended abstracts on Human factors in computing systems - CHI EA □0, 20□0.

□□□ M. Mayer, "Web History Tools and Revisitation Support: A Survey of Existing Approaches and Directions," *Foundations and Trends® in Human–Computer Interaction*, vol. 2, no. □, pp. □□□–2□8, 200□.

□8□ L. Jin, L. Feng, G. Liu, and C. Wang, "Personal Web Revisitation by Context and Content Keywords with Relevance Feedback," *IEEE Transactions on Knowledge and Data Engineering*, vol. 2□, no. □, pp. □□08–□□2□, Jan. 20□□.

□□□ G. Papadakis, R. Kawase, E. Herder, and W. Nejdl, "Methods for web revisitation prediction: survey and experimentation," *User Modeling and User-Adapted Interaction*, vol. 2□, no. □, pp. □□□–□□□, 20□□.

□□0□ J. Šimko, M. Tvarozek, and M. Bielikova, "Semantic History Map: Graphs Aiding Web Revisitation Support," *2010 Workshops on Database and Expert Systems Applications*, 20□0.

□□□□ T. M. Mitchell, *Machine Learning*. New York□McGraw-Hill, □□□□.

□□2□ B. J. Jansen, A. Spink, and T. Saracevic, "Real life, real users, and real needs: a study and analysis of user queries on the web," *Information Processing & Management*, vol. □□, no. 2, pp. 20□–22□, 2000.

□□□□ J. Y. Park, N. Ohare, R. Schifanella, A. Jaimes, and C.-W. Chung, "A Large-Scale Study of User Image Search Behavior on the Web," *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems - CHI 15*, 20□□.

□□□□ S. D. Torres, I. Weber, and D. Hiemstra, "Analysis of Search and Browsing Behavior of Young Users on the Web," *ACM Transactions on the Web*, vol. 8, no. 2, pp. □–□□, Jan. 20□□.

□□□□ J. Jadav, C. Tappert, M. Kollmer, A. M. Burke, and P. Dhiman, "Using te□t analysis on web filter data to e□plore K-□2 student learning behavior," *2016 IEEE 7th Annual Ubiquitous Computing, Electronics & Mobile Communication Conference (UEMCON)*, 20□□.

□□□□ C. Bernaschina, M. Brambilla, A. Mauri, and E. Umuhoza, "A Big Data Analysis Framework for Model-Based Web User Behavior Analytics," *Lecture Notes in Computer Science Web Engineering*, pp. □8–□□□, 20□□.

□□□□ V. S. Tseng and K. W. Lin, "Efficient mining and prediction of user behavior patterns in mobile web systems," *Information and Software Technology*, vol. □8, no. □, pp. □□□–□□□, 200□.

□□8□ J. Teevan, E. Cutrell, D. Fisher, S. M. Drucker, G. Ramos, P. Andr□, and C. Hu, "Visual snippets: summarizing web pages for search and revisitation," *Proceedings of the 27th international conference on Human factors in computing systems - CHI 09*, pp. 202□-20□2, 200□.

□□□□ A. Demachi, S. Matsushima, and K. Yamanishi, "Web Behavior Analysis Using Sparse Non-Negative Matrix Factorization," *2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pp. □□□-□8□, 20□□.

□20□ E. Agichtein, E. Brill, and S. Dumais, "Improving Web Search Ranking by Incorporating User Behavior Information," *ACM SIGIR Forum*, vol. □2, no. □, pp. □□–□8, 20□□.

□2□□ Y. Liu, R. Cen, M. Zhang, S. Ma, and L. Ru, "Identifying web spam with user behavior analysis," *Proceedings of the 4th international workshop on Adversarial information retrieval on the web - AIRWeb 08*, pp. □-□□, 2008.

□22□ C. Liu, R. W. White, and S. Dumais, "Understanding web browsing behaviors through Weibull analysis of dwell time," *Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval - SIGIR 10*, pp. □□□-□8□, 20□0.

□2□□ E. Adar, J. Teevan, and S. T. Dumais, "Large scale analysis of web revisitation patterns," *Proceeding of the twenty-sixth annual CHI conference on Human factors in computing systems - CHI 08*, pp. □□□□-□20□, 2008.

□2□□ H. Zhang and S. Zhao, "Measuring web page revisitation in tabbed browsing," *Proceedings of the 2011 annual conference on Human factors in computing systems - CHI 11*, pp. □8□□-□8□□, 20□□.

□2□□ L. Jiang, C. Li, S. Wang, and L. Zhang, "Deep feature weighting for naive Bayes and its application to text classification," *Engineering Applications of Artificial Intelligence*, vol. □2, pp. 2□–□□, 20□□.

□2□□ H. Zhang, Z.-□. Cao, M. Li, Y.-Z. Li, and C. Peng, "Novel naïve Bayes classification models for predicting the carcinogenicity of chemicals," *Food and Chemical Toxicology*, vol. □□, pp. □□□–□□□, 20□□.

□2□□ F.-C. Chen and M. R. Jahanshahi, "NB-CNN□Deep Learning-Based Crack Detection Using Convolutional Neural Network and Na□ve Bayes Data Fusion," *IEEE Transactions on Industrial Electronics*, vol. □□, no. □, pp. □□□2–□□00, 20□8.

□28□ F. C. Jonathan and O. Karnalim, "Semi-Supervised Keyphrase E□traction on Scientific Article using Fact-based Sentiment," *TELKOMNIKA (Telecommunication Computing Electronics and Control)*, vol. □□, no. □, p. □□□□, Jan. 20□8.

□2□□ I. H. Witten, E. Frank, M. A. Hall, and C. J. Pal, *Data mining: practical machine learning tools and techniques*. Amsterdam□ Morgan Kaufmann, 20□□.

□□0□ "jsoup Java HTML Parser, with best of DOM, CSS, and jquery," *jsoup Java HTML Parser, with best of DOM, CSS, and jquery*. □Online□. Available□https□□jsoup.org□. □Accessed□□□-Feb-2020□.

□□□□ Xangis, "Xangis□e□tra-stopwords," *GitHub*. □Online□. Available□ https□□github.com□□angis□e□tra-stopwords□blob□master□chinese. □Accessed□□□-Feb-2020□.

□□2□ "Main Page," *Wikipedia*, 0□-Feb-2020. □Online□. Available□ https□□en.wikipedia.org□wiki□Main□Page. □Accessed□□□-Feb-2020□.

□□□□ Xangis, "Xangis/extra-stopwords," *GitHub*. □Online□. Available□ https□□github.com□□angis□e□tra-stopwords□blob□master□english. □Accessed□□□-Feb-2020□.

□□□□ "Notes on regression and inheritance in the case of two parents," *Proceedings of the Royal Society of London*, vol. □8, no. □□□-□□2, pp. 2□0–2□2, □8□□.

□□□□ O. Karnalim, "Software Keyphrase Extraction with Domain-Specific Features," *2016 International Conference on Advanced Computing and Applications (ACOMP)*, pp. □□-□0, 20□□.

□□□□ G. J. Mclachlan, K.-A. Do, and C. Ambroise, "Analyzing Microarray Gene Expression Data," *Wiley Series in Probability and Statistics*, 200□.