THE IEEE
Intelligent Informatics
BULLETIN

IEEE COMPUTER SOCIETY

IEEE Computer Society
Technical Committee
on Intelligent Informatics

October 2020 Vol. 20 No. 1 (ISSN 1727-5997)

On-line version: http://www.comp.hkbu.edu.hk/~iib (ISSN 1727-6004)

The IEEE Intelligent Informatics Bulletin

Aims and Scope

The IEEE Intelligent Informatics Bulletin is the official publication of the Technical Committee on Intelligent Informatics (TCII) of the IEEE Computer Society, which is published once a year in both hardcopies and electronic copies. The contents of the Bulletin include (but may not be limited to):

1) Letters and Communications of the TCII Executive Committee
2) Feature Articles
3) R&D Profiles (R&D organizations, interview profile on individuals, and projects etc.)
4) Book Reviews
5) News, Reports, and Announcements (TCII sponsored or important/related activities)

Materials suitable for publication at the IEEE Intelligent Informatics Bulletin should be sent directly to the Associate Editors of respective sections.

Technical or survey articles are subject to peer reviews, and their scope may include the theories, methods, tools, techniques, systems, and experiences for/in developing and applying biologically and linguistically motivated computational paradigms, such as artificial neural networks, fuzzy logic, evolutionary optimization, rough sets, and self-organization in the research and application domains, such as data mining, Web intelligence, intelligent agent technology, parallel and distributed information processing, and virtual reality.

Publisher: *The IEEE Computer Society Technical Committee on Intelligent Informatics*
Address: *Department of Computer Science, Hong Kong Baptist University, Kowloon Tong, Hong Kong (Attention: Dr. William K. Cheung; Email:william@comp.hkbu.edu.hk)*
ISSN Number: *1727-5997(printed)1727-6004(on-line)*
Abstracting and Indexing: *All the published articles will be submitted to the following on-line search engines and bibliographies databases for indexing—Google(www.google.com), The ResearchIndex(citeseer.nj.nec.com), The Collection of Computer Science Bibliographies (liinwww.ira.uka.de/bibliography/index.html), and **DBLP** Computer Science Bibliography (www.informatik.uni-trier.de/~ley/db/index.html).*

On Data Science and Machine Intelligence Research Innovation and Translation

A SHORT INTRODUCTION TO THE DATA SCIENCE AND MACHINE INTELLIGENCE LAB

Guandong Xu, Shaowu Liu and Qian Li *

Data Science and Machine Intelligence Lab, University of Technology Sydney, Australia.*

E-mail: {Guandong.Xu, Shaowu.Liu, Qian.Li}@uts.edu.au

ABSTRACT

The Data Science and Machine Intelligence (DSMI) lab at the University of Technology Sydney (UTS) is a joint taskforce of excellent academics and researchers dedicated towards excellence and innovation across academia and industry, with research priorities in data science and artificial intelligence. The lab is led by the founding director Professor Guandong Xu and currently has 7 academics and ∼20 research students. In the past few years, the lab has received over $6M external research funding from national funding bodies, governments, corporations, and private sectors. Supported by these funds, researchers at DSMI conducted world-class research and published papers in premier journals, such as TNNLS, TSE, TKDE, and world-top conferences, such as CVPR, AAAI, IJCAI, SIGIR, SIGKDD, WWW, WSDM, ICDE, ICDM, EMNLP. The team also received a number of international and national awards from Data Analytics, Computer, information industry and financial services professional bodies, such as EFMA, ACS, and iAwards.

I. INTRODUCTION

The Data Science and Machine Intelligence (DSMI)[1] lab was founded by Professor Guandong Xu in 2015. The lab has intensively researched on topics of artificial intelligence, behavioural modelling, social computing, recommender systems, natural language processing, predictive and prescriptive analytics, advanced visualization, causality discovery, and causal inference. In the past few years, the lab has received over $6M external research funding from national funding bodies, governments, corporations, and private sectors, such as Australian Research Council Discovery Projects (ARC-DP) and Linkage Projects (ARC-LP), Cooperative Research Centre Project (CRC-P), Colonial First State Investment, NSW Ministry of Health, OnePath, and Providence Asset Group. Supported by these funds, researchers at DSMI continuously publish papers in top conferences and journals in data science and artificial intelligence, such as TNNLS, TKDE, TSE, CVPR, AAAI, IJCAI, SIGIR, SIGKDD, WWW, WSDM, ICDE, ICDM, EMNLP.

In addition to its research excellence, the DSMI lab is also renowned for its translational research in positively impacting the industry and society. For example, the lab's research outcomes have been implemented and operationalised in the following real-world businesses:

- Our AI-based models have significantly simplified the life insurance underwriting process at OnePath, extending the insurance coverage for more vulnerable Australians.
- Our data analytics solutions have optimized the policies of National Emergency Access Target (NEAT) at NSW Ministry of Health, saving lives in emergency departments.
- Our undergoing project with Providence Asset Group incorporates AI into solar farms, delivering renewable energy solutions for world-class energy efficiency.

As a result of impact and recognition, the lab, alongside its industry partners, has received many awards at national and international levels. For example, the work with OnePath was selected by CeBIT (the largest international computer expo) in 2018 to present a feature on the red-carpet showcase, receiving industry awards such as the European Finance Management Association Workforce transformation award (No.3), Australian Computer Society Digital Disruptors Award winner, Australian Information Industry Association Awards, etc.

The above only provides a glance at the research at DSMI, and more details are given in the rest of this article.

II. RESEARCH AREAS

The DSMI lab mainly focuses on data science and artificial intelligence research, though it also involves interdisciplinary research, such as FinTech, digital health, and renewable energy. The key research areas include

- Recommender systems.
- Behavioural modelling, social computing, web mining.
- Text mining and NLP, software code analysis.
- Predictive analytics, time series forecasting.
- Causality and fairness in machine learning discovery.
- Knowledge graph and representation learning.

We briefly introduce the above research areas as follows.

A. Recommender Systems

Recommender System, an active domain of information provision, focuses on modelling user-centric information (e.g., access-logs, purchase history, rating records, and product reviews) to recommend new items. The recommendation aims

[1]lab website: www.dsmi.tech

at improving user experience and loyalty, facilitating decision-making for users and creating more revenues for online businesses and merchants, and so on. Our strengths include

- Cold-start and long-tail recommendation.
- Social network and trust-based recommendation.
- Sequential-based recommendation.
- Geo-based recommendation.
- Explainable recommendation.

See (1; 2; 3) for examples of our research in this area.

B. Behaviour Modelling and Social Computing

Behaviour modelling and social computing are two fundamental research directions at DSMI. Behaviour modelling is to derive user behavioural patterns, preferences, profiles from user behaviours and user-generated textual data, while social computing is to analyze the social characteristics and trends demonstrated from intra-human or human-computer interactions. We continuously publish papers on these two topics and apply the developed techniques to help our industry partners with their business problems. Our strengths include

- Identifying misbehaviours or abnormal behaviours.
- Discovering underlying behaviour patterns.
- Techniques: topic modelling, representation learning, and embedding, rule mining, sequential pattern mining, anomaly detection, time series analysis.

See (4; 5; 6) for examples of our research in this area.

C. Text Mining and Natural Language Processing

Our research advanced the state-of-the-art in devising new algorithms in analysing meaning, topics, patterns from user-generated content, such as user comments and reviews, product descriptions, social media posts, and customer spoken language, e.g. call-logs for actionable business insights. For example, DSMI has developed analytic models using open-source software to identify risk factors in mental health and customer behaviour patterns in the R&D area, as well as applying to enterprise business operations for identifying high-risks to enable a preventative management approach. Those models have been successfully deployed in Colonial First State Investment (one Australian top-tier superannuation company) for customer churn prediction, and OnePath life Insurance for mental insurance insight analysis.

Our strengths include

- Term indexing, retrieval and search and ranking.
- Sentiment analysis and figurative detection.
- Topic modelling and summarisation.
- Word and context embedding algorithms.
- Latent semantic analysis.

See (7; 8; 9; 10) for examples of our research in this area.

D. Predictive Analytics

The lab has extensive experience in designing customised predictive analytical tools within business operating frameworks to develop automated assistive AI tools to augment the performance of staff (including improving operational resource efficiency, quality assurance and fraud identification), as well as through developing more target and improved personalised customer experiences. Partnered companies include the Australian Taxation Office, OnePath, NSW Health, Colonial First State, Providence Asset Group. Our strengths include

- Domain-specific feature engineering.
- Structured and unstructured data integration.
- Time-series forecasting.
- Multi-label supervised learning.
- What-if analysis, prescriptive analytics.

See (11; 12; 5; 13) for examples of our research in this area.

E. Causality and Fairness in Machine Learning

Our research advances the state-of-the-art causal analysis theory and its interplay with interpretable machine learning (specifically deep learning, transfer learning, and supervised learning). Our research pursues the goal of enabling machine learning methods with causality and human-level intelligence. Although current interpretable models have greatly improved the interpretability landscape, they are unable to provide causal explanations for human-level intelligence. The causality is a clear and mathematically sound mechanism for users to understand the core of machine learning. How to empower the interpretable models with advanced causal explanations is largely unexploited. Fairness becomes a common concern by public applications in various areas, especially some regulated areas, such as online job-seeking and recruitment systems. Such fairness issue is caused by either training data bias, or limitation of machine learning algorithms. Our ongoing research along this line will research devising effective algorithms to alleviate the fundamental fairness challenge in machine learning research.

Our strengths include

- Causal inference and causal discovery.
- Causal reasoning for machine learning.
- Causality in the explainable recommendation.
- Visual causality analysis.
- Fairness in machine learning.

See (14; 15) for examples of our research in this area.

F. Knowledge Graph and Representation Learning

The multi-relationship of Knowledge Graph (KG) brings new challenges for Knowledge Graph analysis, it also makes the research on KG more attractive, because, with this kind of automatically extracted structured human knowledge, we have an opportunity to reveal the human knowledge reasoning patterns with analysis methodologies. As a result of KG analysis, KG can be used as a semantic enhancement for downstream application scenarios, such as Recommender System (RS). In our research, both KG completion and KG based downstream applications are studied. Although a huge amount of human knowledge facts have been collected from multiple open resources, existed KG is still incomplete. Part of our current research focuses on the following proper embedding models for KG completion, 1) entity & relation embedding model, 2) conceptual taxonomy integrated embedding model,

and 3) multi-relational graph sub-structure embedding. See (16; 17; 18; 19) for examples of our research in this area.

III. Research with real-world impact

This section summarises how the lab's research is translated into business solutions, improving business outcomes and benefiting society.

A. Using AI and a hybrid ESS solution to fully integrate solar generation into the distribution system

Funded by Australia Cooperative Research Centres Projects (CRC-P), in this project, we use the combination of AI and hydrogen to unleash the power of solar energy produced at solar farms for better energy productivity, efficiency, operation, and maintenance. DSMI provides the expertise in AI for a hybrid Energy Storage System (ESS) solution to integrate solar generation fully into the distribution system. The project aims to develop a widely applicable integrated package for small-scale solar farming, focusing not just on photovoltaic technologies and solutions, but on the monitoring, control, integration and optimisation of distributed solar farming.

B. Smart Personalized Privacy Preserved Information Sharing in Social Networks

Funded by Australia Research Council Discovery Projects (ARC-DP), this project aims to create a novel and effective method for privacy protection at the individual level, which is now a great concern of persons, businesses, and government agencies in this big data age. The project expects to build an automatic smart and practical personalised privacy-preserving system through removing the fundamental obstacles. The project will significantly advance human knowledge of privacy, and push Australia to the front line of the research field, and protect Australia better.

C. Reshaping Australian superannuation practice via big data analytics

Funded by Australia Research Council Linkage Projects (ARC-LP), this project aims to reform superannuation investment practices in Australia. Using sophisticated data analytics and machine-learning techniques, combined with economic modelling and quantitative finance. The project will try to understand the broad characteristics of Australian superannuation investors and their practice from a 'big data' perspective. The expected outcomes of this project are the identification of key determinants for successful superannuation behaviour to inform decision-making for better superannuation practices and policies. It is expected that this project will contribute to safeguarding the future of Australia's superannuation schemes, and to better financial security at retirement.

D. AI-enhanced Underwriting systems and insights of Mental Health

Funded by OnePath-Zurich, this project aims to optimise the AI-based underwriting risk engine, anti-selection detection, and conduct a pilot study of mental health disorder analysis and visualisation. This project has received the international EFMA workforce transformation award in 2020.

E. AI-enhanced Life Insurance underwriting automation and optimization for ANZ Wealth

Funded by OnePath/ANZ Wealth, this project aims to develop an AI-based underwriting risk engine to improve the current manual underwriting process in life insurance. The data-driven model provides personalised, efficient service with improved quality assurance for customers when they apply for insurance. This project is partnered with ANZ wealth.

F. Longitudinal Study on Taxpayer Behavioural Analysis

Funded by Australia Research Council Linkage Projects (ARC-LP), this project is to investigate the taxpayer behavioural patterns in Australian. Specific analytics lens is developed to reveal the characteristics of interested cohorts, e.g. debtor from longitudinal point of view. This project was jointly funded by ARC and Australian Taxation Office.

G. Personality mining via call log analysis

Funded by Colonial First State, this project is to devise a data model of big-five personality scores based on customer call centre logs. Big data analytics on textual, audio, and video data is used to train the personality analysis engine. The predicted personality traits will create value for various business applications.

H. Develop Deep Insights in Customer Retention

Funded by Colonial First State, this project incorporated big data of customers, e.g. demography, transaction, interaction, and behaviour information into predictions of customer churn. Machine learning-based prediction models were developed for various business products.

I. NSW Emergency Department Treatment Performance Analysis

Funded by NSW Ministry of Health, this project works on Emergency Department (ED) Performance Assessment initiative for review and adjustment of '4-hour' ED discharge policy, which is currently applied across Australian hospital ED by patient cohort mining, and patient in-hospital journey and re-admission prediction model based on linked GP data. These analytical results and models are adapted to policy change and decision-making support.

IV. Research Facilities

Researchers at DSMI have access to world-leading facilities at UTS, such as the Data Arena – a 360-degree interactive data visualisation facility set to change the way we view and interact with data, as shown in Figure 2. The lab also has access to world-class computing and storage infrastructure, including a total of 2000+ Intel Xeon cores, 20TB+ of RAM, and 500TB+ of SSD hard drive.

Fig. 1: DSMI is in the UTS Central building in Sydney's CBD.

Fig. 2: A large cylindrical screen, four metres high and ten metres in diameter. A high performance computer graphics system drives six 3D-stereo video projectors, edge-blended to create a seamless three-dimensional panorama.

V. Public Recognition

The lab's research has drawn attention from both academia and industry, evidenced by awards and media release.

National and international awards:

- 2020 Global Efma-Accenture Insurance Innovation Award in Workforce Transformation.
- 2019 Digital Disruptors Winner Award of Skills Transformation of Small Work Teams by Australian Computer Society (Figure 3).

Fig. 3: 2019 Digital Disruptors Winner Award of Skills Transformation of Small Work Teams by Australian Computer Society

- 2019 Digital Disruptors Gold Award of Service Transformation for the Digital Consumer – Corporate by Australian Computer Society.
- 2019 NSW State Merit iAwards in Category of Research & Development Project of the year, Business Service Markets, and Data Insights Innovation of the year by Australian Information Industry Association (AIIA) (Figure 4).

Fig. 4: 2019 NSW State Merit iAwards in Category of Research & Development Project of the year.

- 2018 Digital Disruptors Gold Award for Skill Transformation in Work Team by Australian Computer Society.
- 2018 Best Industry Application of Data Analytics Award by BigInsight Data and Ai Innovation Award (Figure 5).
- 2018 Best Industry Application of AI Award by BigInsight Data and Ai Innovation Award.

Fig. 5: 2018 Best Industry Application of Data Analytics Award by BigInsight Data and Ai Innovation Award.

Media release:

- Global Award for AI Transforming Insurance Underwriting, UTS FEIT News, 2020.
- AAI awards for fintech partnerships, UTS FEIT News, 2018.
- AI-led insurance innovation wins Awards, UTS Newsroom, 2018.
- Super future, UTS Newsroom, 2018.
- AI the future of insurance and underwriting, CeBIT, 2018 (Figure 6).

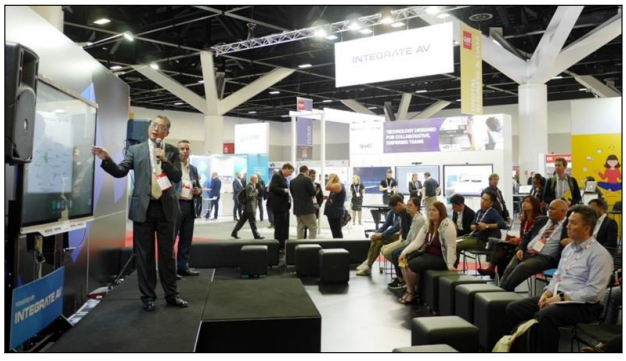

Fig. 6: 2018 CeBIT Sydney red carpet show.

- Colonial First State and UTS use machine learning to predict investors, Financial Review News, 2017.
- Early wins for OnePath's AI insurance underwriting project, ComputerWorld, 2017.
- ANZ Wealth exploring AI for insurance underwriting, ComputerWorld, 2017.
- ANZ and UTS seek AI underwriting, FinanceCareer, 2017.

VI. Looking into the future

The vision of the lab is to become a world-class research lab, delivering high-quality publications and industry innovation to unleash the potential of our partners and benefit society. To maintain and lift our profile in innovative research and industry engagement, we keep recruiting creative, passionate, and self-motivated talented students to grow our team. We also welcome research collaborations internationally.

References

[1] D. Wang, X. Zhang, D. Yu, G. Xu, and S. Deng, "Came: Content-and context-aware music embedding for recommendation," *IEEE Transactions on Neural Networks and Learning Systems*, 2020.

[2] X. Wang, Q. Li, W. Zhang, G. Xu, S. Liu, and W. Zhu, "Joint relational dependency learning for sequential recommendation," in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 2020, pp. 168–180.

[3] H. Ying, F. Zhuang, F. Zhang, Y. Liu, G. Xu, X. Xie, H. Xiong, and J. Wu, "Sequential recommender system based on hierarchical attention network," in *IJCAI International Joint Conference on Artificial Intelligence*, 2018.

[4] N. N. Vo, X. He, S. Liu, and G. Xu, "Deep learning for decision making and the optimization of socially responsible investments and portfolio," *Decision Support Systems*, vol. 124, p. 113097, 2019.

[5] Z. Saeed, R. A. Abbasi, I. Razzak, O. Maqbool, A. Sadaf, and G. Xu, "Enhanced heartbeat graph for emerging event detection on twitter using time series networks," *Expert Systems with Applications*, vol. 136, pp. 115–132, 2019.

[6] J. Yin, Z. Zhou, S. Liu, Z. Wu, and G. Xu, "Social spammer detection: A multi-relational embedding approach," in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 2018, pp. 615–627.

[7] R. Biddle, A. Joshi, S. Liu, C. Paris, and G. Xu, "Leveraging sentiment distributions to distinguish figurative from literal health reports on twitter," in *Proceedings of The Web Conference 2020*, 2020, pp. 1217–1227.

[8] W. Wang, Y. Zhang, Y. Sui, Y. Wan, Z. Zhao, J. Wu, P. Yu, and G. Xu, "Reinforcement-learning-guided source code summarization via hierarchical attention," *IEEE Transactions on Software Engineering*, 2020.

[9] N. N. Vo, S. Liu, J. Brownlow, C. Chu, B. Culbert, and G. Xu, "Client churn prediction with call log analysis," in *International Conference on Database Systems for Advanced Applications*. Springer, 2018, pp. 752–763.

[10] Y. Wan, Z. Zhao, M. Yang, G. Xu, H. Ying, J. Wu, and P. S. Yu, "Improving automatic source code summarization via deep reinforcement learning," in *Proceedings of the 33rd ACM/IEEE International Conference on Automated Software Engineering*, 2018, pp. 397–407.

[11] Y. Shu, Q. Li, S. Liu, and G. Xu, "Learning with privileged information for photo aesthetic assessment," *Neurocomputing*, 2020.

[12] I. Razzak, R. A. Saris, M. Blumenstein, and G. Xu, "Integrating joint feature selection into subspace learning: A formulation of 2dpca for outliers robust feature selection," *Neural Networks*, vol. 121, pp. 441–451, 2020.

[13] R. Biddle, S. Liu, P. Tilocca, and G. Xu, "Automated underwriting in life insurance: Predictions and optimisation," in *Australasian Database Conference*. Springer, 2018, pp. 135–146.

[14] G. Xu, T. D. Duong, Q. Li, S. Liu, and X. Wang, "Causality learning: A new perspective for interpretable machine learning," *arXiv preprint arXiv:2006.16789*, 2020.

[15] G. X. Y. Wu, J. Cao, "Fast: A fairness assured service recommendation strategyconsidering service capacity constraints," in *Proceedings of the 2020 ICSOC*, 2020.

[16] Z. Zhou, S. Liu, G. Xu, and W. Zhang, "On completing sparse knowledge base with transitive relation embedding," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 3125–3132.

[17] Z. Zhou, S. Liu, G. Xu, X. Xie, J. Yin, Y. Li, and W. Zhang, "Knowledge-based recommendation with hierarchical collaborative embedding," in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 2018, pp. 222–234.

[18] Z. Wang, Q. Li, G. Li, and G. Xu, "Polynomial representation for persistence diagram," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[19] C. Zheng, Y. Cai, J. Xu, H.-f. Leung, and G. Xu, "A boundary-aware neural model for nested named entity recognition," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019, pp. 357–366.

Web Intelligence Centre Universidad de Chile

A VERY SHORT INTRODUCTION ABOUT WIC-CHILE ACTIVITIES

Francisca González Cohens, Rocío B. Ruiz, Felipe Vera, Victor Hernández, Víctor Cortés, María F. Guiñazu, Juan D. Velásquez*

Web Intelligence Centre, Department of Industrial Engineering, University of Chile.*

e-mail: {francisca.gonzalez, rocio.ruiz, felipe.vera, victor.hernandez, victor.cortes,flavia.guinazu,jvelasqu}@wic.uchile.cl

ABSTRACT

Web Intelligence Centre from Universidad de Chile started in 2010 as professor Juan D. Velásquez received from the Web Intelligence Consotium's academic and orientation support to create a research centre focused on web user behaviour analysis. The centre grew and established alliances with Medicine School at the same university, developing health-engineering projects with the aim to improving people's health through high-end technologies. It has developed 13 projects, given the opportunity to 100 students to develop their thesis, published near 100 new papers and won 16 funds. Now, it has 5 emblematic projects that are currently impacting health systems and health professional's and patient's lives.

I. THE BEGINNING

The Web Intelligence Centre (WIC) was created back in 2010 by the professor Juan D. Velásquez after the Web Intelligence Consortium gave its academic and orientation support, establishing a participation agreement on the Consorium's events. Its initial aim was to research about web users behaviour analysis (1). It began with him, two more researchers, five undergraduate students and two postgraduate students willing to research and generate new knowledge. The first subject that was treated and researched by the engineering group was, as the centre name says, Web Intelligence (2), approaching to sub topics like Web Mining, Text Mining, Natural Language Processing and Web Opinion Mining. In order to finance the research, the team started to applying into public research competitive funds, having a very good winning rate. Those funds allowed the centre to grow and have more students working on their thesis at the centre, students that later on became engineers and stayed working there, generating more knowledge.

The topics on which the centre has expertise are Analytic and Data Science, Business Intelligence, Artificial Intelligence Software Development, Data Architecture and Engineering, Computer Vision, Big Data Analysis, Predictive Analysis and Natural Language Processing; leading to the centre's vision:

"We create information technologies using data science to support decision making at innovative organizations. We believe that this discipline can generate a big impact on society and that passionate us".

As the years went by, important alliances have been carried out, highlighting the one with Medicine School, establishing a new pathway for the centre: health engineering projects, with the aim to improve people's health using new high-end technologies and creating new multidisciplinary knowledge. Thus, the centre's mission is:

"We want to be a relevant stakeholder on data science area and its applications in healthcare. For that, we will have 3 transferred projects to healthcare centres by 2022".

The centre has given the opportunity to 100 undergraduate, postgraduate and PhD students to develop their thesis on investigation projects, 40 students to continue their studies at foreign universities, as University of Tokyo, San Diego University, Merbourne University, KAIST South Korea, KU Leuven University, among others. It has published at least 30 ISI papers, 50 conference papers, 4 books and 12 book chapters, won 16 funds for the different projects it has developed, 13 in total. In 2018 it organized the International Conference on Web Intelligence from IEEE-ACM-WIC, and nowadays it has 5 operating funds and waiting for the results of some more.

Finally, one of the main differentiators of the centre is its focus on technology transfer, in other words, every technology it develops is accompanied by a transfer strategy to the public o private entity that will use it, and this is because one of the WIC's philosophy is to truly impact society and public policies using the developed technologies. Moreover, currently it is participating in the construction of a strategic plan to approach the usage and analyze artificial intelligence impact in Chile at the Senate of the Republic, and also as part of the expert panel summoned by the presidency to elaborate a national intelligence strategy for the country.

II. EMBLEMATIC PROJECTS

The centre has developed a high number of projects, with different objectives, technologies and alliances. Below, we list the most important projects for the centre nowadays, those projects that are reaching the expected impact, that have allowed other technologies to emerge, that are paying back for all the effort invested and that have a promising future.

A. Docode

For the acronym DOcument COpy DEtector, DOCODE is the very first project of the WIC. The first idea of this technology dates back to 2008, when professor Velasquez started researching about plagiarism detection algorithms using text mining (3). It won 4 funds to finance the research needed to build the engine behind and a usable tool

for customers. It took about 8 years for it to debug in what it is now, an automatic plagiarism detector in documents against indexed web pages, repository or other specific documents, the outcome is a side by side comparison between the original document and the ones reviewed, giving a plagiarism index of easily interpretation (4; 5). In 2011 won the first place on the global plagiarism detection contest in Amsterdam, Netherlands. Actually it is been used by Universidad de Chile and by 1 other university that buy Docode's services as well as 5 more big clients from the government and Chile's state, besides 10 smaller clients, being small enterprises and professors. And the most important thing is that this platform is helping not only to reduce plagiarism, from 40% to 1%, but also to assure original content and teach the importance of citing, creating original content and avoiding copy and plagiarism.

B. Akori

From the acronym Advanced Kernel for Ocular Research and web Intelligence, Akori also means "Falcon" in "Mapudungún"[1], and was the first joint project between WIC and the Faculty of Medicine, back in 2012. It started with the idea of getting to know if it was possible to identify if a web user liked or disliked something in a webpage by looking at his pupils, thing that was studied using eye-tracking, electroencephalogram and key object identification by web mining (6) and determining if click intention could be predicted (7; 8). Nowadays, Akori is a platform that uses web intelligence and physiological variables to determine the web page salient objects for users, it analyses a web page image and returns a heat map whether the looking preferences of the users should be. This project won 2 funds and, by now, is at the stage of improving the algorithms to give faster results.

C. Sonama

Sonama, Social Network Analysis for Marijuana and Alcohol, is a platform that analyses social media information to study marijuana and alcohol prevalence in Chile. Its aim is to give a high value and complex prediction about the potential behaviour of social media users with respect of marijuana and alcohol consumption, opinion and dependency. The technologies used for this purpose are data mining and information fusion, and the main results obtained from the development of the research are (9):

- Data extraction algorithms from Twitter.
- Behavioural model for predicting marijuana consumption.
- Web application prototype to visualise the above results. It allows to see the evolution of some indicators trough time.

The platform extracts Twitter user information, like posts, user data and relationship between users to create indicators of prevalence and risk perception of marijuana (9) and a predictive model to improve substance use surveillance (10). The purpose of this project is to help health authorities take fast and well informed decisions with respect of narcotic policies, which, nowadays are taken using a national survey, taken every 2 years. The National Service for Prevention and Rehab of Drugs and Alcohol Consumption, the health entity in charge of this issue, participated in the design stage of the project and validated the results obtained.

D. Prevedel

Prevedel, for "Preventing Delirium", also developed jointly with the Medicine Faculty, aims to preventing the "delirium" condition affecting senior hospitalized patients. It first started in 2017 from Medicine's idea that some other measures combined in an app could prevent delirium condition in a better way and less invasive than medication, establishing the alliance with WIC and winning 2 funds for its development. The first one finished with a complete developed and proven software for Android devices with little games that stimulate reorientation, cognitive aspect, early moving, sensory help promotion, sleep hygiene and pain management optimization. The software incorporated horizontal disposition, color contrast, big interaction interest areas and customization. The results were that 91% of the population studied could reach the software performances without instructions, and 100% after them, and there was 60% less delirium in the group that used the app (11). The second part of the study, the one financed by the second fund, is taking place right now, and looks forward to proving the efficacy of the software in truly preventing delirium.

E. Kefuri

Kefuri, kidney in Mapudungún, is a web platform to warn the Procurement team about the arrival of a possible organ donor to the Emergency Room (ER). This project, developed jointly wit Medicine school, began as a student project that was sponsored by the centre, whose aim was to increasing organ donation rate in a country, Chile, that has had a historically low donation rate despite communicational efforts encouraging donation (12). The first steps of the project were research for all the processes that take place to achieve an actual organ donor and discover the bottle neck stages that could have solution using high-end technologies. In those researches, the investigation team discovered various inefficiencies and inefficiencies, but that the very first stage, the one where the possible organ donor came in to the hospital was the weakest one because ER personnel did not warn Procurement unit, the one in charge of organ donation in hospitals, about the arrival of the patient (13). The project won one fund to develop and test the software, now available for Android and iOS, at the "Hospital del Salvador", where physicians and nurses can easily and quickly warn about this type of patients just by entering 4 variables and pressing "advise", where the Artificial Intelligence algorithm makes the calls to the procurement unit and Intensive Care Unit to speed up the entrance of the patient for his metabolic and hemodynamic stabilization.

Improving the first stage of the process will show other process bottle necks, which have been studied and will be tackled during the second part of the project. The important thing is that

[1]It means *"land language"* in the language or Mapuche, the Chilean original habitants

Kefuri aims to increasing the number of organ donors available for transplantation, fact that can save the lives of those patients with end stage organ diseases waiting for a transplant, and, as a second derivative, making the system more efficient by procuring several patients at low cost due to innovative technologies, making public health insurance save money.

III. Looking to the future

The mail goal of the centre is to become the most important Health-Engineering centre in Chile, and all the steps that have been taken during this years point that way. WIC works with vision, empowerment, resilience, excellence and compromise; all its workers and students do a first class research to generate new knowledge about novel high-end technologies, how to develop them for the user, understanding the context in which they will be inserted, how to apply them into real world, how to improve them every day and how to transfer them in a proper way, a way that ensures its usage and their possibility to change the world and improve health professionals' and patient's lives.

References

[1] P. Loyola, G. Martinez, K. Muñoz, J. D. Velásquez, P. Maldonado, and A. Couve, "Combining eye tracking and pupillary dilation analysis to identify website key objects," *Neurocomputing*, vol. 168, pp. 179–189, 2015.

[2] J. D. Velásquez and L. C. Jain, *Advanced techniques in web intelligence*. Springer, 2010, vol. 311.

[3] G. Oberreuter and J. D. Veláquez, "Text mining applied to plagiarism detection: The use of words for detecting deviations in the writing style," *Expert Systems with Applications*, vol. 40, no. 9, pp. 3756–3763, 2013.

[4] J. D. Velásquez, Y. Covacevich, F. Molina, E. Marrese-Taylor, C. Rodríguez, and F. Bravo-Marquez, "Docode 3.0 (document copy detector): A system for plagiarism detection by applying an information fusion process from multiple documental data sources," *Information Fusion*, vol. 27, pp. 64–75, 2016.

[5] J. D. Velásquez *et al.*, "Docode 5: Building a real-world plagiarism detection system," *Engineering Applications of Artificial Intelligence*, vol. 64, pp. 261–271, 2017.

[6] G. Slanzi, C. Aracena, and J. D. Velásquez, "Eye tracking and eeg features for salient web object identification," in *International Conference on Brain Informatics and Health*. Springer, 2015, pp. 3–12.

[7] G. Slanzi, J. A. Balazs, and J. D. Velásquez, "Combining eye tracking, pupil dilation and eeg analysis for predicting web users click intention," *Information Fusion*, vol. 35, pp. 51–57, 2017.

[8] C. Aracena, S. Basterrech, V. Snáel, and J. Velásquez, "Neural networks for emotion recognition based on eye tracking data," in *2015 IEEE International Conference on Systems, Man, and Cybernetics*. IEEE, 2015, pp. 2632–2637.

[9] V. D. Cortés, J. D. Velásquez, and C. F. Ibáñez, "Twitter for marijuana infodemiology," in *Proceedings of the International Conference on Web Intelligence*, 2017, pp. 730–736.

[10] M. F. Guiñazú, V. Cortés, C. F. Ibáñez, and J. D. Velásquez, "Employing online social networks in precision-medicine approach using information fusion predictive model to improve substance use surveillance: A lesson from twitter and marijuana consumption," *Information Fusion*, vol. 55, pp. 150–163, 2020.

[11] E. A. Alvarez, M. Garrido, D. P. Ponce, G. Pizarro, A. A. Córdova, F. Vera, R. Ruiz, R. Fernández, J. D. Velásquez, E. Tobar *et al.*, "A software to prevent delirium in hospitalised older adults: development and feasibility assessment," *Age and Ageing*, vol. 49, no. 2, pp. 239–245, 2020.

[12] F. F. González and F. C. González, "Analysis of organ donation for transplantation in chile during 2017," *Revista medica de Chile*, vol. 146, no. 5, pp. 547–554, 2018.

[13] F. González Cohens, F. Vera Cid, R. Alcayaga Droguett, and F. González Fuenzalida, "Critical analysis of the low organ donation rates in chile," *Revista médica de Chile*, vol. 148, no. 2, pp. 242–251, 2020.

[14] P. Loyola, P. E. Román, and J. D. Velásquez, "Clustering-based learning approach for ant colony optimization model to simulate web user behavior," in *2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology*, vol. 1. IEEE, 2011, pp. 457–464.

K-Initialization-Similarity Clustering Algorithm

Tong Liu and Gaven Martin

***Abstract*—As one of the most used clustering algorithms, K-means clustering algorithm has been applied in variety areas. Its clustering result depends on the predefined cluster number, the initialization, and the similarity measure. Previous research focused on solving parts of these issues but has not focused on solving them in a unified framework. However, fixing one of these issues does not guarantee the best performance. To improve K-means clustering algorithm, we propose the K-Initialization-Similarity (KIS) clustering algorithm to solve the issues of the K-means clustering algorithm in a unified way. Specifically, we propose to learn the similarity matrix based on the data distribution, to automatically output the cluster number using a robust loss function, and to fix the initialization by using sum-of-norms which outputs the new representation of the original samples. The proposed algorithms outperformed the state-of-the-art clustering algorithms on real data sets. Moreover, we theoretically prove the convergences of the proposed optimization methods for the proposed objective function.**

***Index Terms*—Clustering, K-means, Spectral clustering, Machine learning, Initialization, Similarity measure.**

I. INTRODUCTION

K-means clustering algorithm is considered one of the most used clustering algorithms. It has been successfully applied to broad areas such as artificial intelligence, machine learning, data mining, etc.

K-means clustering algorithm partitions the dataset into *K* distinct clusters in the following steps: First, it initializes cluster centers via randomly selecting *K* data points as the *K* cluster centers. Second, it assigns each data point to its nearest cluster center according to a similarity measure, e.g., Euclidean distance. Third, it revises the *K* cluster centers using the mean of assigned data points in each cluster. K-means clustering algorithm keeps repeating the last two steps until the algorithm achieves convergence [1, 2].

As one of the most famous and widely used clustering algorithm, K-means clustering algorithm still has its limitations. It is difficult to determine the cluster number *K* without prior knowledge. Different initializations may obtain completely different clustering results. K-means clustering results depend on the similarity measure such as Euclidean distance measure which does not account for the factors such as cluster sizes, dependent features or density [3, 4]. Thus K-means clustering algorithm is not good for indistinct or not well-separated data sets [5, 6]. Existing methods only solved some of these problems. All these issues of K-means clustering algorithm are important to be addressed to improve the performance of K-means clustering algorithm. Many literatures have solved some parts of these issues of K-means clustering algorithm [7-9]. For example, Duan et al. developed an algorithm to calculate the density to select the initial cluster centers [10]. Lakshmi et al. proposed to use nearest neighbors and feature means to decide the initial cluster centers [8]. Other works also addressed the issues of K-means clustering algorithm [11-14].

However, previous clustering algorithms only fixed parts of the issues of the K-means clustering algorithm. When a clustering algorithm addresses those problems separately, it is easily to be trapped into the sub-optimal results, which means it is hard to obtain a global optimal solution, for example, even if a best initial value is found or the best similarity matrix is found, but the final optimal results may not be obtained. Because the results of the individual steps are not obtained according to the requirements of the next steps. It would be significant if we could fix the issues of the initialization, cluster number determination and similarity measure problems of K-means clustering algorithm in a unified framework to achieve global optimal results.

Our proposed new K Initialization Similarity (KIS) algorithm is aimed to develop an improved K-means clustering algorithm while solving the issues of the cluster number determination, the initialization, the similarity measure in a unified way. Specifically, the cluster number is automatically generated by using a robust loss function, the initialization of the clustering using sum-of-norms (SON) regularization, the similarity matrix based on the data distribution. Furthermore, we employ the alternative strategy to solve the proposed objective function. The experimental results on real-world benchmark data sets also demonstrates that our KIS clustering algorithm outperforms the related clustering methods in terms of accuracy (Acc), the assessment evaluation metric for clustering algorithm [1].

We briefly summarize the contributions of our proposed KIS algorithm as follows:

- A unified way addresses the cluster number determination, initialization, and similarity measure issues around clustering.
- The cluster number is automatically generated using a robust loss function.
- The initialization is fixed by using the sum-of-norms regularization
- The similarity measure is generated based on the data distribution
- The proposed clustering algorithm outperforms comparison clustering algorithms. It implies that simultaneously addressing the three issues (cluster number determination, initialization, and similarity measure) is feasible and robust.

This section has laid the background of our research inquiry. The remainder of this paper is organized as follows. Section 2 discusses the existing relevant clustering algorithm. Section 3 introduces our KIS algorithm. Section4 discusses the experiments we conducted and present the results of our

experiments. The conclusions and future research direction are presented in Section 5.

II. Related Work

Clustering can be generally categorized into the prototype-based and the non-prototype-based approaches, based on whether the clustering algorithm is center-based or not.

A. Prototype-based clustering Algorithms

The prototype of the corresponding cluster is the center of the data points in each cluster. The prototype-based clustering algorithms assign data points to their closest prototypes, such that data points in the group are closer to the prototype of the cluster than to the prototype of any other group. K-means clustering algorithm is one of the most famous representatives of this kind of clustering approaches [15, 16]. It keeps recalculating the prototypes followed by assigning each data point to a cluster represented by a prototype until the algorithm achieves convergence [1]. There are numbers of other algorithms based on prototype clustering algorithms, e.g. K-medoids, COTCLUS, and Tabu search. K-medoids chooses the data points located near their prototypes to represent the clusters. The rest of remaining data points are clustered with the representative prototype to which they are the most similar based on the minimal sum of the dissimilarities between data points and their corresponding cluster prototypes [17]. Instead of using only one center for each class, COTCLUS, an improved prototype-based clustering algorithm, uses suitable prototypes from another cluster. It finds two prototypes from one cluster and replace them by two prototypes from the other cluster in such a way that maximum decreases the mean square error of the first clustering. It constructs a clustering from two suboptimal clustering results based on the belief that each suboptimal clustering has benefits regarding to containing some of the correct clusters [18]. After modifying centroids, it applies K-means clustering algorithm for final fine-tuning [18]. A Tabu based clustering algorithm employs the prototype driven approach of the K-means clustering algorithm with the guidance of Tabu search, which is a local or neighborhood search algorithm that accepts the worsening searches of no improving search is available and discourages the search from going back to previously visited search [19]. The K-medoids, COTCLUS, and Tabu search example like other K-means clustering algorithm need to specify the cluster number *K* before the execution of the algorithms.

B. Non-Prototype-based clustering algorithms

Instead of conducting clustering based on the cluster centers by the prototype approaches, some non-prototype-based clustering algorithms use links or graph. Robust clustering using links (ROCK) [20]. ROCK clustering algorithm draws a number of data points randomly from the original data set as inputs along with the desired cluster number *K*. Instead of using distances to conduct clustering, ROCK uses the number of links which is defined as the number of common neighbors as the similarity measure [20]. But ROCK ignores the possible differences in the similarity measure of different clusters inside the same data set. Graph is also used for representing the high-order relationship among data points [21]. A graph is a set of nodes with connected edges which have weights associated with them [22]. Spectral clustering algorithm is an example of clustering algorithms using graph. It creates a similarity matrix first and then defines a feature vector. Then it runs the K-means clustering algorithm to conduct clustering [23]. It creates the spectral representation and conducts the final clustering in separate stages, and it requires the cluster number beforehand because it uses of the K-means clustering algorithm. Low-rank representation (LRR) identifies the subspace structures from data points and then finds the lowest rank representation among data points to represent the original data points [24]. A low-rank kernel learning graph-based clustering (LKLGC) algorithm is based on a multiple kernel learning with assumption that the consensus kernel matrix is a low-rank matrix and lies in the neighbourhood of the combined kernel matrix [25]. The spectral clustering algorithm is applied to get the final clustering results for LKLGC algorithm, hence it is a multi-stage clustering and the cluster number needs to be predefined [25]. A low-rank kernel learning for Graph-based Clustering (LKG) iteratively constructs graph and kernel learning which exploits the similarity of the kernel matrix and an optimal kernel from the neighboring candidate kernels [11]. It requires the cluster number beforehand. A hybrid representative selection based ultra-scalable spectral clustering (U-SPEC) constructs a sparse affinity sub-matrix by using a hybrid representative selection strategy and a K-nearest representatives approximation method, and then interprets the sparse sub-matrix as a bipartite graph, which is partitioned using transfer cut to obtain the clustering result [12]. It is a multi-stage clustering and cluster number is prerequisite.

Current prototype-based and non-prototype-based clustering algorithms do not simultaneously solve the initialization, similarity measure or cluster number issues of non-graph-based clustering algorithms.

III. K-Initialization-Similarity Clustering

Given a data matrix $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\} \in \mathbb{R}^{n \times d}$, where n and d, respectively, are the number of samples and features, we denote boldface uppercase letters, boldface lowercase letters,

TABLE I
Description of symbols used in this paper

Symbol	Description
$\mathbf{X}$	Data matrix
$\mathbf{x}$	A vector of $\mathbf{X}$
$\mathbf{x}_i$	The i-th row of $\mathbf{X}$
$x_{i,j}$	The element in the i-th row and j-th column of $\mathbf{X}$
$\|\mathbf{x}\|_2$	L_2-norm of $\mathbf{x}$
$\|\mathbf{X}\|_F$	The Frobenius norm or the Euclidean norm of $\mathbf{X}$
$\mathbf{X}^T$	The transpose of $\mathbf{X}$
K	Cluster number

and italic letters as matrices, vectors and scalars, respectively, and also summarize the symbols used in this paper in Table I.

A. Motivation

To find out how other algorithms improve K-means clustering algorithm, we investigate Spectral clustering algorithm, Robust Clustering using links, Low-rank kernel learning for graph-based clustering (LKG), and Ultra-scalable spectral clustering (U-SPEC) beside K-means clustering algorithm in detail.

K-means algorithm aims at minimizing the total intra-cluster variance represented by an objective function known as the squared error function shown in Eq. (1):

$$E = \sum_{j=1}^{K} \sum_{i \in C_j} \left\| i_l - w_j \right\|^2 \quad (1)$$

where K is number of clusters, $j \in \{1, \dots, K\}$, n is number of data points, $l \in \{1, \dots, n\}$, $(w_1, \dots, w_K)$ is the K prototypes. C_j is the j^{th} cluster. K-means clustering algorithm operates in the following steps: First, it initializes k prototypes $(w_1, \dots, w_K)$ via randomly selecting K data points from $l \in \{1, \dots, n\}$. Second, it assigns each i_l to the cluster C_j with w_j, each C_j is associated with w_j. Third, it updates the prototype w_j for each cluster C_j using the mean. K-means clustering algorithm keeps repeating the last two steps until the E doesn't change or change insignificantly [12].

Instead of using original data points, Spectral clustering algorithm conducts K-means clustering on spectral representation. To do this, Spectral clustering algorithm first creates a similarity matrix, and then constructs a diagonal degree matrix using the sum of all the weights on each row of the similarity matrix and a feature vector by computing the first K eigenvectors of its Laplacian matrix, which is the degree matrix subtracting the similarity matrix. Finally, it runs K-means clustering on these features to separate objects into K clusters [23]. Spectral clustering algorithm is a multi-step algorithm and it requires the cluster number to be predefined.

Robust clustering using links (ROCK) obtains a number of random data points from the original data, then uses the link agglomerative approach with a goodness measure, which determines which pair of points is merged at each step as shown in Eq. (2). Finally, the remaining data points are assigned to these clusters [20].

$$g(K_i, K_j) = \frac{\text{link}(K_i, K_j)}{(n_i + n_j)^{1+2f(\theta)} - n_i^{1+2f(\theta)} - n_j^{1+2f(\theta)}} \quad (2)$$

where $\text{link}(K_i, K_j)$ is the number of links between the two clusters. n_i and n_j are the number of points in each cluster. The function f satisfies the property that each item in K_i, has approximatelt $n_i^{f(\theta)}$ neighbors in the cluster. The reasoning behind using link is that the data points belonging to the same cluster most likely have a large number of common neighbours, thus more links. Hence the larger the number of links between data points, the greater likelihood they belong to the same cluster. But ROCK ignores the possible differences the similarity of different data points and it require the cluster number beforehand as well.

Low-rank kernel learning for graph-based clustering (LKG) constructs graph and learns consensus kernel in a unified framework. Its low-rank kernel matrix is learnt by exploiting the similarity of the kernel matrix and seeking an optimal kernel from the neighboring of candidate kernels. The graph and kernel are iteratively enhanced by each other. LKG runs the spectral clustering algorithm to achieve the final clustering results.

$$\min_{Z,K,\mathbf{g}} \frac{1}{2} Tr\|K - 2KZ + Z^T K Z\|_F^2 + \alpha\rho(Z) + \beta\|K\|_* + \gamma\left\|K - \sum_i \mathrm{g}_i H^i\right\|_F^2,\ s.t.\ Z \geq 0, K \geq 0, \mathrm{g}_i \geq 0, \sum_i^r \mathrm{g}_i = 1 \quad (3)$$

where Z is self-expression coefficient, K is nonnegative kernel matrix, H is kernel matrix, the weight of kernel H^i is g_i, the constraints for g are from standard Multiple kernel learning method. The corresponding $\mathrm{g}_i$will be assigned a small value if a kernel is not appropriate. $\|K\|_*$ is the structure of the kernel matrix, where K will respect the correlations among data points with the cluster structure. The last term in Eq. (3) seeks an optimal kernel K in the neighborhood of $\sum_i \mathrm{g}_i H^i$. Z and K repeatedly learnt in a unified model. LKG reinforces the underlying connections between the optimal kernel learning and graph learning.

To improve the randomness and efficiency of K-means cluster algorithm, a hybrid representative selection based ultra-scalable spectral clustering (U-SPEC) was designed. It interprets the sparse sub-matrix as a bipartite graph, which is partitioned using transfer cut to obtain the clustering result. U-SPEC algorithm conducts in three phases. In the first phase, a hybrid representative selection strategy is applied by randomly selecting candidates and obtaining representatives from candidates via K-means. In the second phase, a coarse-to-fine method is used to approximate the K-nearest representatives for each data points, and to construct a sparse affinity sub-matrix between the data points and the representatives. The sparse affinity sub-matrix is represented by the Eq. (4). In the third phase, the sub-matrix is interpreted as a bipartite graph, which is partitioned to the final clusters.

$$B = \{b_{i,j}\}_{Nxp}, b_{i,j} = \begin{cases} \exp\left(-\frac{\|x_i - r_j\|^2}{2\sigma^2}\right), \text{if } r_j \in N_K(x_i) \\ 0, \text{ otherwise} \end{cases} \quad (4)$$

where $N_K(x_i)$ represents the set of K-nearest representatives of x_i and is the average Euclidean distance between the data points and their K-nearest representatives.

Previous clustering algorithms only fixed part of the issues of the K-means clustering algorithm. It would be significant for our KIS clustering algorithm fixing the issues of the initialization, cluster number determination and similarity measure problems of K-means clustering algorithm in a unified framework to achieve global optimal results.

B. K-Initialization-Similarity Clustering Algorithm

This paper proposes a new clustering algorithm (i.e., K-Initialization-Similarity (KIS)) to simultaneously solve the cluster number determination, the initialization issue, and the similarity measure issue of K-means clustering algorithm in a unified framework. Specifically, KIS clustering algorithm

generates the new representation of original data points, applies sum-of-square error estimation to minimize the difference between the original data and its new representative, learns the similarity matrix based on the data distribution, and generated the cluster number *K* by using the robust loss function. To achieve our goal, we form the objective function of the KIS clustering algorithm as follows:

$$\min_{\mathbf{S},\mathbf{U}} \frac{1}{2}\|\mathbf{X}-\mathbf{U}\|_F^2 + \frac{\alpha}{2}\sum_{i,j=1}^{n} s_{i,j}\rho(\|\mathbf{u}_i-\mathbf{u}_j\|_2) + \beta\|\mathbf{S}\|_2^2, \quad s.t., \forall i, s_{i,j} \geq 0, \mathbf{s}_i^T\mathbf{e} = 1 \tag{5}$$

where $\mathbf{X} \in \mathbb{R}^{n\times d}$ is the data matrix, $\mathbf{U} \in \mathbb{R}^{n\times d}$ is the new representation of $\mathbf{X}$, and $\mathbf{S} \in \mathbb{R}^{n\times n}$ is the similarity matrix to measure the similarity among data points, and $\rho\left(\|\mathbf{u}_i-\mathbf{u}_j\|_2\right)$ is a robust loss function, used for automatically generating clusters. The smaller the value of $\|\mathbf{u}_i-\mathbf{u}_j\|_2$ is, the closer the distance is, and the higher the similarity $\mathbf{s}_i$ and $\mathbf{s}_j$is. With the update of other parameters in Eq. (5), the distance $\|\mathbf{u}_i-\mathbf{u}_j\|_2$ for some i and j, will be very close, or even $\mathbf{u}_i = \mathbf{u}_j$. The clusters will be determined. $\boldsymbol{e} = [\mathbf{1}, \dots, \mathbf{1}]^T$.

Eq. (5) fixes the initialization of clustering, automatically learns the new representation **U** and the similarity matrix **S**, and generates the cluster number. The similarity matrix **S** learning is based on the data distribution, i.e., iteratively updated by the updated **U**. This produces an intelligent new representation of the original data matrix.

Moreover, Eq. (5) will keep the distance of indicator vectors similar if the data belongs to the same cluster, possibly making them equal. The distance of indicator vectors is as separated as possible if data belongs to the different clusters.

Several robust loss functions have been proposed to avoid the influence of noise and outliers in robust statistics [26]. Here we employ the Geman-McClure function [27]:

$$\mathrm{P}\left(\|\mathbf{u}_p-\mathbf{u}_q\|_2\right) = \frac{\mu\|\mathbf{u}_p-\mathbf{u}_q\|_2^2}{\mu+\|\mathbf{u}_p-\mathbf{u}_q\|_2^2} \tag{6}$$

The literature of half-quadratic minimization and robust statistics explains the reason for selecting Geman–McClure loss function instead of other loss functions [28]. Eq. (6) measures how well a model predicts the expected outcome. The smaller the value of $\|\mathbf{u}_p-\mathbf{u}_q\|_2^2$ is, the closer the distance is, and the higher the similarity s_p and s_q is. With the update of other parameters in Eq. (6), the distance $\|\mathbf{u}_p-\mathbf{u}_q\|_2^2$ for some p and q, will be very close, or even $\mathbf{u}_p = \mathbf{u}_q$. The clusters will be determined.

The optimization of the robust loss function is challenging. To address this, it is normal practice to introduce an auxiliary variable $f_{i,j}$ and a penalty item $\varphi(f_{i,j})$, and thus Eq. (5) is rewritten to:

$$\min_{\mathbf{S},\mathbf{U},\mathbf{F}} \frac{1}{2}\sum_{i=1}^{n}\|\mathbf{x}_i-\mathbf{u}_i\|_2^2 + \frac{\alpha}{2}\sum_{i,j=1}^{n} s_{i,j}(f_{i,j}\|\mathbf{u}_i-\mathbf{u}_j\|_2^2 + \varphi(f_{i,j})) + \beta\sum_{i=1}^{n}\|\mathbf{s}_i\|_2^2 \;\; s.t., \forall i, s_{i,j} \geq 0, \boldsymbol{s}_i^T\boldsymbol{e} = 1 \tag{7}$$

where $\varphi(f_{i,j}) = \mu\left(\sqrt{f_{i,j}}-1\right)^2, i, j = 1 \dots n$

This objective function is still challenging to solve. An iterative optimization process is adopted to tackle this challenge. In the next section, we will show how iterative optimization is utilized to solving the problem.

C. Optimization

Eq. (7) is not jointly convex on **F**, **U**, and **S**, but is convex on each variable while fixing the rest. To solving the Eq. (7), the alternating optimization strategy is applied. We optimize each variable while fixing the rest until the algorithm converges. The pseudo-code of KIS clustering algorithm is given in Algorithm 1.

*1) Update **F** while fixing **S** and **U***

While **S** and **U** are fixed, the objective function can be rewritten in a simplified matrix form to optimize **F**:

$$\mathrm{Min}_{\mathbf{F}} \frac{\alpha}{2}\sum_{i,j=1}^{n} s_{i,j}(f_{i,j}\|\mathbf{u}_i-\mathbf{u}_j\|_2^2 + \mu\left(\sqrt{f_{i,j}}-1\right)^2) \tag{8}$$

Since the optimization of $f_{i,j}$ is independent of the optimization of other $f_{p,q}, i \neq p, j \neq q$, the $f_{i,j}$ is optimized first as shown in following Eq. (9)

$$\mathrm{Min}_{f_{i,j}} \frac{\alpha}{2}(s_{i,j}f_{i,j}\|\mathbf{u}_i-\mathbf{u}_j\|_2^2 + s_{i,j}\left(\mu\left(f_{i,j}-2\sqrt{f_{i,j}}+1\right)\right) \tag{9}$$

By conducting a derivative on Eq. (9) with respect to $f_{i,j}$, we get Eq. (10).

$$\frac{\alpha}{2}(s_{i,j}\|\mathbf{u}_i-\mathbf{u}_j\|_2^2 + s_{i,j}\mu - s_{i,j}\mu f_{i,j}^{-\frac{1}{2}}) = 0 \tag{10}$$

$$\Rightarrow \frac{\alpha}{2}\, s_{i,j}\|\mathbf{u}_i-\mathbf{u}_j\|_2^2 + \frac{\alpha}{2}\, s_{i,j}\mu - \frac{\alpha}{2}s_{i,j}\mu f_{i,j}^{-\frac{1}{2}} = 0 \tag{11}$$

$$\Rightarrow f_{i,j} = \left(\frac{\mu}{\mu+\|\mathbf{u}_i-\mathbf{u}_j\|_2^2}\right)^2 \tag{12}$$

*2) Update **S** while fixing **U** and **F***

While fixing **U** and **F**, the objective function Eq. (7) with respect to **S** is:

$$\mathrm{Min}_{\mathbf{S}} \frac{\alpha}{2}\sum_{i,j=1}^{n}(s_{i,j}f_{i,j}\|\mathbf{u}_i-\mathbf{u}_j\|_2^2 + s_{i,j}(\mu\left(\sqrt{f_{i,j}}-1\right)^2)) + \beta\sum_{i=1}^{n}\|\mathbf{s}_i\|_2^2, s.t., \forall i, s_{i,j} \geq 0, \mathrm{s}_i^T\mathrm{e} = \mathrm{I} \tag{13}$$

Since the optimization of $\mathbf{s}_\mathrm{I}$ is independent of the optimization of other $\mathbf{s}_j, i \neq \mathrm{j}, i, j = 1, \dots, \mathrm{n}$, the $\mathbf{s}_i$ is optimized first as shown in following:

$$\mathrm{Min}_{\mathbf{s}_i} \frac{\alpha}{2}\sum_{j=1}^{n} s_{i,j}(f_{i,j}\|\mathbf{u}_i-\mathbf{u}_j\|_2^2 + \mu\left(\sqrt{f_{i,j}}-1\right)^2) + \beta\|\mathbf{s}_i\|_2^2, \quad s.t., \forall i, s_{i,j} \geq 0, \mathrm{s}_i^T\mathrm{e} = 1 \tag{14}$$

Let $b_{i,j} = f_{i,j}\|\mathbf{u}_i-\mathbf{u}_j\|_2^2$ and $c_{i,j} = \mu\left(\sqrt{f_{i,j}}-1\right)^2$, Eq. (14) is equivalent to:

$$\mathrm{Min}_{\mathbf{s}_i} \frac{\alpha}{2}\sum_{j=1}^{n} s_{i,j}b_{i,j} + \frac{\alpha}{2}\sum_{j=1}^{n} s_{i,j}c_{i,j} + \beta\|\mathbf{s}_i\|_2^2, s.t., \forall i, s_{i,j} \geq 0, \boldsymbol{s}_i^T e = 1 \tag{15}$$

$$\Rightarrow \min_{\mathbf{s}_i} \frac{\alpha}{2}\mathbf{s}_i^T\mathbf{b}_i + \frac{\alpha}{2}\mathbf{s}_i^T\mathbf{c}_i + \beta\|\mathbf{s}_i\|_2^2, \quad s.t., \forall i, s_{i,j} \geq 0, \mathbf{s}_i^T e = 1 \tag{16}$$

$$\Rightarrow \min_{\mathbf{s}_i} \mathbf{s}_i^T\mathbf{s}_i + 2\mathbf{s}_i\frac{\alpha}{4\beta}\mathbf{s}_i^T(\mathbf{b}_i+\mathbf{c}_i) + \frac{\alpha}{4\beta}\mathbf{s}_i^T(\mathbf{b}_i+\mathbf{c}_i)^T(\mathbf{b}_i+\mathbf{c}_i) - \frac{\alpha}{4\beta}\mathrm{s}_i^T(\mathbf{b}_i+\mathbf{c}_i)^T(\mathbf{b}_i+\mathbf{c}_i), s.t., \forall i, s_{i,j} \geq 0, \mathbf{s}_i^T e = 1 \tag{17}$$

$$\Rightarrow \min_{\mathbf{s}_i} \left\|\mathbf{s}_i + \frac{\alpha}{4\beta}(\mathbf{b}_i+\mathbf{c}_i)\right\|_2^2, s.t., \forall i, s_{i,j} \geq 0, \mathbf{s}_i^T\mathrm{e} = 1 \tag{18}$$

According to Karush-Kuhn-Tucker (KKT) [29], the optimal solution $\mathbf{s}_i$ should be

$$S_{i,j} = \max\{-\frac{\alpha}{4\beta}(b_{i,j} + c_{i,j})\} + \theta, 0\}, j = 1, \dots, n \quad (19)$$

where $\theta = \frac{1}{\rho}\sum_{j=1}^{\rho}\left(\frac{\alpha}{4\beta}(b_{i,j} + c_{i,j}) + 1\right)$, and ω is the descending order of $\frac{\alpha}{4\beta}(b_{i,j} + c_{i,j})$. and $\rho = {}^{max}_{j}\{\boldsymbol{\omega}_j - \frac{1}{j}\left(\sum_{r=1}^{j}\boldsymbol{\omega}_r - 1\right), 0\}$.

Algorithm 1. The pseudo code for KIS clustering algorithm

Input: $\mathbf{X} \in \mathbb{R}^{n \times d}$
Output: a set of *K* clusters

Initialization: **U** = **X**;
Repeat:
- Update **F** using Eq. (12)
- Update **S** using Eq. (19)
- Update **U** using Eq. (24)

Until U converges

*3) Update **U** while fixing **S** and **F***

While **S** and **F** are fixed, the objective function can be rewritten in a simplified form to optimize **U**:

$$\text{Min}_{\mathbf{U}} \ \frac{1}{2}\sum_{i,j=1}^{n}\|\mathbf{x}_i - \mathbf{u}_i\|_2^2 + \frac{\alpha}{2}\sum_{i,j=1}^{n} s_{i,j} f_{i,j}\left\|\mathbf{u}_i - \mathbf{u}_j\right\|_2^2 \quad (20)$$

Let $h_{i,j} = s_{i,j} f_{i,j}$. Eq. (3.23) is equivalent to:

$$\text{Min}_{\mathbf{U}} \ \frac{1}{2}\|\mathbf{X} - \mathbf{U}\|_F^2 \ + \frac{\alpha}{2}\sum_{i,j=1}^{n} h_{i,j}\left\|\mathbf{u}_i - \mathbf{u}_j\right\|_2^2 \quad (21)$$

$$\Rightarrow \text{min}_{\mathbf{U}} \ \frac{1}{2}tr(\mathbf{X}^T\mathbf{X} - 2\mathbf{U}^T\mathbf{X} + \ \mathbf{U}^T\mathbf{U}) \ + \frac{\alpha}{2}tr(\mathbf{U}^T\mathbf{L}\mathbf{U}) \quad (22)$$

After conducting a derivative on Eq. (22) with respect to **U**, we get Eq. (23).

$$\Rightarrow \frac{1}{2}(-2\mathbf{X} + 2\mathbf{U}) \ + \frac{\alpha}{2}(\mathbf{L}\mathbf{U} \ + \ \mathbf{L}^T\mathbf{U}) = 0 \quad (23)$$

$$\Rightarrow \mathbf{U} = (\mathrm{I} + \alpha\mathbf{L})^{-1}\mathbf{X} \quad (24)$$

D. Convergence Analysis

In this section, we prove the convergence of the proposed KIS clustering algorithm in order to prove the proposed algorithm can reach an optimal solution, so we apply Theorem 1.

Theorem 1. KIS clustering algorithm decreases the objective function value of Eq. (7) until it converges.

Proof.

By denoting $\mathbf{F}^{(t)}$, $\mathbf{S}^{(t)}$, and $\mathbf{U}^{(t)}$, the results of the t-th iteration of **F**, **S**, and **U** respectively, we further denote the objective function value of Eq. (7) in the t-th iteration as $\mathcal{L}(\mathbf{F}^{(t)}, \mathbf{S}^{(t)}, \mathbf{U}^{(t)})$.

According to Eq. (12), **F** has a closed-form solution, thus we have the following inequality:

$$\mathcal{L}(\mathbf{F}^{(t)}, \mathbf{S}^{(t)}, \mathbf{U}^{(t)}) \geq \mathcal{L}(\mathbf{F}^{(t+1)}, \mathbf{S}^{(t)}, \mathbf{U}^{(t)}) \quad (25)$$

According to Eq. (19), **S** has a closed-form solution, thus we have the following inequality:

$$\mathcal{L}(\mathbf{F}^{(t+1)}, \mathbf{S}^{(t)}, \mathbf{U}^{(t)}) \geq \mathcal{L}(\mathbf{F}^{(t+1)}, \mathbf{S}^{(t+1)}, \mathbf{U}^{(t)}) \quad (26)$$

According to Eq. (24), **U** has a closed-form solution, thus we have the following inequality:

$$\mathcal{L}(\mathbf{F}^{(t+1)}, \mathbf{S}^{(t+1)}, \mathbf{U}^{(t)}) \geq \mathcal{L}(\mathbf{F}^{(t+1)}, \mathbf{S}^{(t+1)}, \mathbf{U}^{(t+1)}) \quad (27)$$

Finally, based on above three inequalities, we get

$$\mathcal{L}(\mathbf{F}^{(t)}, \mathbf{S}^{(t)}, \mathbf{U}^{(t)}) \geq \mathcal{L}(\mathbf{F}^{(t+1)}, \mathbf{S}^{(t+1)}, \mathbf{U}^{(t+1)}) \quad (28)$$

Equation. (28) indicates that the objective function value in Eq. (7) decreases after each iteration of Algorithm 1. This concludes the proof of Theorem 1.

IV. EXPERIMENTS

In this section, we evaluated the performance of the proposed K-Initialization-Similarity (KIS) algorithm, by comparing it with two benchmark algorithms on ten real UCI data sets, in terms of evaluation metric Accuracy.

A. Data Sets

We used ten UCI data sets in the experiments [30] including the standard data sets for email spam. wine quality, website fishing, and chess game data sets, etc. The details are summarized in Table II.

B. Comparison Algorithms

The five comparison algorithms are summarized below:

- K-means clustering algorithm randomly initializes the cluster center, then (re)assigns data points to their nearest cluster center and recalculates cluster centers iteratively until converge.
- Spectral clustering algorithm constructs the similarity matrix, and then defines the feature vectors. Finally, it runs K-means clustering algorithm.
- ROCK clustering algorithm randomly selects a number of data points from the original data and uses the number of links as the similarity measure.
- LKG constructs graph and low-rank kernel matrix by exploiting the similarity of the kernel matrix and an optimal kernel from the neighboring candidate kernels. The graph and kernel are iteratively enhanced by each other.
- U-SPEC constructs a sparse affinity sub-matrix by using a hybrid representative selection strategy and a K-nearest representatives approximation method.

C. Evaluation Measure

To assess the performance of the proposed algorithms with related algorithms, we adopted accuracy (ACC) which is a popular evaluation metric for clustering algorithms. ACC measures the percentage of samples correctly clustered [31].

TABLE II
DESCRIPTION OF TEN BENCHMARK DATA SETS

Dataset	Sample	Feature	Class
Isolet	7797	617	2
SpamBase	4601	57	2
Chess	3196	36	2
Banknote	1372	5	2
Diabetes	1151	19	2
Yeast	1484	8	10
Website	1353	9	2
Wine	1599	11	6

The definition of ACC is given below.

$$\mathrm{ACC} = \frac{N_{correct}}{N} \quad (29)$$

where $N_{correct}$ represents the number of correct clustered samples, and N represents total number of samples.

To rank the performance of different algorithms, we used dense ranking which the highest accuracy rate receives number 1, and the next accuracy rate receives the immediately following ranking number. Same accuracy rates receive the same ranking number. Thus if A ranks ahead of B and C (which compare equal) which are both ranked ahead of D, then A gets ranking number 1 ("first"), B gets ranking number 2 ("joint second"), C also gets ranking number 2 ("joint second") and D gets ranking number 3 ("Third").

D. Experiment Setup

In the experiments, first, we tested the robustness of the proposed KIS clustering algorithm by comparing it with K-means, Spectral, ROCK, LKG, and U-SPEC clustering algorithms using real data sets in terms of evaluation metric widely used for clustering research. Second, we investigated the parameters' sensitivity of the proposed KIS clustering algorithm (i.e. α and β in Eq. (7)) via varying their values to observe the variations of clustering performance. Third, we demonstrated the convergence of Algorithm 1 by checking the value of the proposed objective function Eq. (7) via iteration times.

E. Experimental Results Analysis

The performances of all algorithms are listed in Table III, which shows that the KIS algorithm achieved the best overall performance on each of the eight data sets in terms of ACC. More specifically, on the average ACC results of all eight data sets, the KIS algorithm increased it by 33.42%, 28.03%, 26.58%, 30.25%, and 25.93% respectively, compared to K-means clustering result, Spectral, ROCK, U-SPEC, and LPG. Other observations are listed below.

First, KIS, LKG, U-SPCE, and Spectral clustering algorithm outperformed K-means clustering algorithm. This implied that constructing the graph or learning a new representation of original data points improves the clustering performance. The reason could be that original data generally contains some noise or redundant information, which is often true in real data set and the noise and redundancy may corrupt the performance of clustering methods. In contrast, the non-prototype graph-based algorithms construct the new representation to conduct clustering, which can relieve the affection of noise and redundancy from original data, so the clustering performance can be improved.

Second, clustering algorithms using adaptive similarity measure, e.g. KIS clustering algorithm, performed better than nonadaptive clustering algorithms, e.g. K-means, Spectral, ROCK, U-SPCE that use the fixed similarity measurement to measure the similarity, our KIS employed an adaptive learning strategy to dynamically update the similarity matrix. In this way, our KIS can more accurately capture the intrinsic correlation of original data. This explains why our KIS easily outputs better clustering results than nonadaptive similarity clustering algorithms. This proves that the adaptive learning similarity leads to optimal clustering results, whereas the nonadaptive similarity measure clustering algorithms achieves sub-optimal results.

Third, the proposed KIS clustering algorithm use the unified framework to simultaneously address the major issues of clustering algorithms. Addressing these issues in a unified way achieves one global goal leading to optimal clustering results, whereas the multi-stage clustering algorithms with separate goals in each stage achieve sub-optimal results. When a clustering algorithm addresses those problems separately, it is easily to be trapped into the sub-optimal results, for example, even if a best initial value or the best similarity matrix is found, the final optimal results may not be obtained. Because the results of the previous steps are not obtained according to the requirements of the final step.

F. Parameters' Sensitivity

We varied parameters α and β in the range of $[10^{-3}, \ldots 10^{3}]$, and recorded the values of ACC of eight data sets clustering results for KIS clustering algorithm in Fig. 1. The parameter α is used to tune the auxiliary variable F. The parameter β is used to tradeoff the importance of similarity matrix S.

The different data sets needed different ranges of parameters to achieve the best performance. For example, KIS clustering algorithm achieved the best ACC (85.79%) on data set Isolet when both parameters α is 10^{2} and β were 10^{3}. But for the data set Wine, KIS clustering algorithm achieved the best ACC (92.94%) when both $\beta = 10^{-3}$ and $\alpha = 10^{-3}$. This indicated that KIS clustering algorithm was data driven.

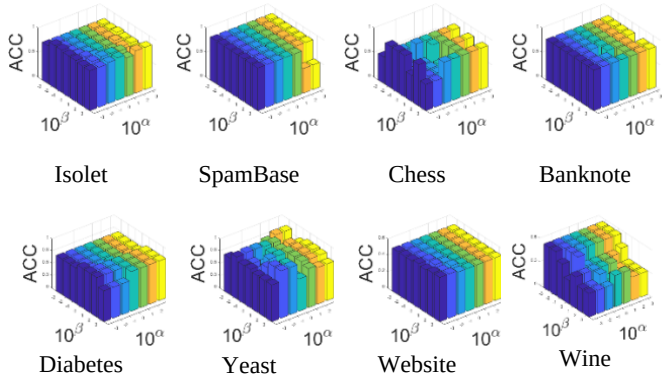

Fig. 1. ACC results of KIS algorithm with respect to different parameter settings

G.Convergence

Fig. 2 showed the trend of objective values generated by the proposed algorithm 1 with respect to iterations. The convergence curve indicates the change of the objective function value during the iteration process. From Fig. 2, we can see that the algorithm 1 monotonically decreased the objective function value until it converged, when applying it to optimize the proposed objective function in Eq. (7). That means that the value of the objective function stop changing or only change in a small range e.g. $\left|obj_{(t+1)} - obj_{(t)}\right| / obj_{(t)} \leq 10^{-9}$, where $obj_{(t)}$ represents the objection function value of Eq. (7) after the t-th iteration. In our proposed optimization algorithm, we have employed an alternating optimization strategy to optimize our objective function, i.e., iteratively updating each parameter until the algorithm converges. Thus, the optimal solution can be worked out by multiple iterations until the demand of minimizing the objective values is satisfied, which means the objective values decline to stable, as shown as the convergence lines. It is worth noting that the convergence rate of the algorithm 1 was relatively fast, converging to the optimal value within 20 iterations on all the data sets used. In other words, we can complete the optimization of our model in a fast speed.

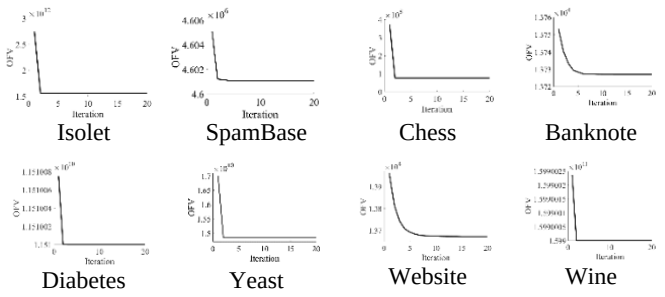

Fig. 2. Objective function values (OFVs) versus iterations for KIS algorithm

V. Conclusion

In this research we have proposed a new algorithm named K-Initialization-Similarity (KIS) which aims to solving the cluster number K determination, initialization, similarity measure issues of K-means clustering algorithm in a unified way. Specifically, we fixed the initialization by using the sum-of-norms regularization which outputted the new representation of original data points. The similarity matrix learning is based on the data distribution. The robust loss function is applied to automatically generate the cluster number *K*. The optimal performance is achieved when the separated issues are solved in a unified way. Experiment results on eight real-world benchmark data sets show that KIS outperforms the comparison clustering algorithms in terms of accuracy (ACC), the popular evaluation metric for clustering algorithm.

Although the proposed KIS clustering algorithm achieved good clustering results, we haven't considered imbalanced data sets. Hence, future research needs to improve our KIS clustering algorithm to automatically determine the clustering number K, fix the initialization, learn similarity in a unified way and have capability of handling imbalanced data.

TABLE III
ACC results of KIS algorithm on ten benchmark data sets (The highest score of each evaluation metric for each data set is highlighted in bold font)

Datasets	K-means	Spectral	ROCK	U-SPEC	LKG	KIS
Isolet	0.5065	0.5067	0.5103	0.5910	0.5410	**0.8579**
SpamBase	0.5915	0.5965	0.5947	0.6001	0.6062	**0.8660**
Chess	0.4959	0.5976	0.5207	0.5864	0.6677	**0.8403**
Banknote	0.4776	0.5904	0.4439	0.605	0.6276	**0.8216**
Diabetes	0.5130	0.5130	0.5317	0.530	0.5317	**0.6419**
Yeast	0.2353	0.3134	0.3383	0.2709	0.3127	**0.7663**
Website	0.4808	0.5212	0.5203	0.5795	0.5203	**0.5824**
Wine	0.3309	0.424	0.4259	0.4165	0.424	**0.9294**
Rank	5	4	4	3	2	**1**

References

[1] Hartigan, J.A. and M.A. Wong, *Algorithm AS 136: A k-means clustering algorithm.* Journal of the Royal Statistical Society. Series C (Applied Statistics), 1979. 28(1): p. 100-108.

[2] Jeong, Y., et al., K-means data clustering with memristor networks. Nano letters, 2018. 18(7), p. 4447-4453.

[3] Femi, P.S. and S.G. Vaidyanathan. Comparative Study of Outlier Detection Approaches. in ICIRCA. 2018. IEEE. p. 366-371.

[4] Buczkowska, S., N. Coulombel, and M. de Lapparent, *A comparison of euclidean distance, travel times, and network distances in location choice mixture models.* Networks and spatial economics, 2019: p. 1-34.

[5] Doad, P.K. and M.B. Mahip, *Survey on Clustering Algorithm & Diagnosing Unsupervised Anomalies for Network Security.* International Journal of Current Engineering and Technology ISSN, 2013: p. 2277-410.

[6] Saradha, T.M.P.D.A., *An Improved K-means Cluster algorithm using Map Reduce Techniques to mining of inter and intra cluster datain Big Data analytics.* International Journal of Pure and Applied Mathematics, 2018. **119**(7): p. 679-690.

[7] Shah, S.A. and V. Koltun, *Robust continuous clustering.* Proceedings of the National Academy of Sciences, 2017. **114**(37): p. 9814-9819.

[8] Lakshmi, M.A., G.V. Daniel, and D.S. Rao, *Initial Centroids for K-Means Using Nearest Neighbors and Feature Means*, in *Soft Computing and Signal Processing*. 2019, Springer. p. 27-34.

[9] Motwani, M., N. Arora, and A. Gupta, *A Study on Initial Centroids Selection for Partitional Clustering Algorithms*, in *Software Engineering*. 2019, Springer. p. 211-220.

[10] Duan, Y., Q. Liu, and S. Xia. *An improved initialization center k-means clustering algorithm based on distance and density*. in *AIP*, https://doi.org/10.1063/1.5033710

[11] Yan, Q., et al., *A discriminated similarity matrix construction based on sparse subspace clustering algorithm for hyperspectral imagery.* Cognitive Systems Research, 2019. **53**: p. 98-110.

[12] Bian, Z., H. Ishibuchi, and S. Wang, *Joint Learning of Spectral Clustering Structure and Fuzzy Similarity Matrix of Data.* IEEE Transactions on Fuzzy Systems, 2019. **27**(1): p. 31-44.

[13] Rong, H., et al., *A novel subgraph K^+-isomorphism method in social network based on graph similarity detection.* Soft Computing, 2018. **22**(8): p. 2583-2601.

[14] Fränti, P. and S. Sieranoja, *How much can k-means be improved by using better initialization and repeats?* Pattern Recognition, 2019. **93**: p. 95-112.

[15] Xu, D. and Y. Tian, *A comprehensive survey of clustering algorithms.* Annals of Data Science, 2015. **2**(2): p. 165-193.

[16] Song, J., F. Li, and R. Li. *Improved K-means Algorithm Based on Threshold Value Radius*. in *IOP Conference Series: Earth and Environmental Science*. 2020. IOP Publishing. doi:10.1088/1755-1315/428/1/012001

[17] Saraswathi, S. and M.I. Sheela, *A comparative study of various clustering algorithms in data mining.* International Journal of Computer Science and Mobile Computing, 2014. **11**(11): p. 422-428.

[18] Rezaei, M., *Improving a Centroid-Based Clustering by Using Suitable Centroids from Another Clustering.* Journal of Classification, 2019: p. 1-14.

[19] Lu, Y., et al., *A Tabu Search based clustering algorithm and its parallel implementation on Spark.* Applied Soft Computing, 2018. **63**: p. 97-109.

[20] Guha, S., R. Rastogi, and K. Shim, *ROCK: A robust clustering algorithm for categor*ical attributes. Information systems, 2000. 25(5): p. 345-366.

[21] Kang, Z., et al., Robust Graph Learning from Noisy Data. IEEE transactions on cybernetics, 2020. 50 (5): p. 1833-1843.
[22] Togninalli, M., et al. Wasserstein weisfeiler-lehman graph kern*els*. in *NIPS*. 2019. p. 6436-6446.
[23] Zhu, X., et al., *Low-rank sparse subspace for spectral clustering.* IEEE Transactions on Knowledge and Data Engineering, 2019. **31**(8): p. 1532-1543.
[24] Liu, G., et al., *Robust recovery of subspace structures by low-rank representation.* IEEE transactions on pattern analysis and machine intelligence, 2013. **35**(1): p. 171-184.
[25] Kang, Z., et al., *Low-rank kernel learning for graph-based clustering.* Knowledge-Based Systems, 2019. **163**: p. 510-517.
[26] Zheng, W., et al., *Unsupervised feature selection by self-paced learning regularization.* Pattern Recognition Letters, 2018: p. 438-446
[27] Geman, S. and D.E. McClure, *Statistical Methods for Tomographic Image Reconstruction.* Bulletin of the International statistical Institute, 1987. **52**(4): p. 5-21.
[28] Nikolova, M. and R.H. Chan, *The equivalence of half-quadratic minimization and the gradient linearization iteration.* IEEE Transactions on Image Processing, 2007. **16**(6): p. 1623-1627.
[29] Voloshinov, V.V., *A generalization of the Karush–Kuhn–Tucker theorem for approximate solutions of mathematical programming problems based on quadratic approximation.* Computational Mathematics and Mathematical Physics, 2018. **58**(3): p. 364-377.
[30] Dua, D. and C. Graff, *UCI Machine Learning Repository, University of California, Irvine, School of Information and Computer Sciences. 2019*. 2019.
[31] Zhu, X., et al., *One-step multi-view spectral clustering.* IEEE Transactions on Knowledge and Data Engineering, 2018. **31**(10): p.2022-2034.

Social Media Data Schemas for Crowdsourced Topics Observational Analytics

Ilias Dimitriadis, Vasileios G. Psomiadis, and Athena Vakali

***Abstract* - Crowdsourcing offers an invaluable toolkit for obtaining dynamic trends and insights from social media data analytics, enabling the capture of the wisdom of the crowds. The plethora of available platforms requires the appropriate definition of data schemas and techniques to allow for efficient knowledge extraction from unstructured social media user generated content and users' multilevel interactions. The present work addresses such challenges by designing an effective and flexible document-based data model that supports heterogeneous social media data integrations. This model is then exploited under a crowdsourced topics observatory that involves interactive visualization modules and advanced topic modelling methods. The proposed framework is implemented and demonstrated on a social innovation platform aiming to promote awareness on plastic waste revaluation and empower stakeholders of the plastics value chain.**

***Index Terms* - Information systems, Crowdsourcing, RESTful web services, Computing methodologies, Topic modeling, Social media analytics, Thematic detection, Heterogeneous data sources, Dynamics and trends discovery, Wisdom of the crowds, Visualization.**

II. INTRODUCTION

Social media data are constantly produced and shared, offering the ground for knowledge extraction and crowd's trends and opinions detection. Social media analytics has become a valuable approach to harvest such knowledge from openly circulated data over multiple popular platforms (such as Twitter, Facebook, Instagram, Flickr). The knowledge derived from the evolving crowd-driven ideas, identified as "the wisdom of the crowds", is heavily dependent on multiple data sources which should be integrated under appropriate and adaptive data schemas. Data produced in abundance in online social media sources offer a fertile ground for harvesting most popular topics expressed openly in the form of opinions and views of users about key issues.

This work is motivated by the need to define appropriate approaches which will support effective social media data schemas designs, resulting in social media data topics analytics. Dealing with multiple data types in social media is a challenging task since each of the online platforms produces different data types and formats and has a different scope and origin. For example, data produced in Twitter include a source (tweet text) and a metadata (retweet, time, user-id, geo-location, etc.) part, data in Facebook include such parts but also a so called social graph (with users friendship information), data in Flickr or Instagram focus on mostly multimedia type of User Generated Content (UGC), etc.

Capturing the actual wisdom of the crowds, ideally demands social media data integration from multiple platforms, thus harvesting knowledge over multiple crowdsourced data threads and types requires novel approaches and solutions. The exploitation of social media resources leads to the delivery of innovative and more human-centred services by leveraging the collective intelligence of the crowd. In this context, social media data mining as a collective intelligence approach, has evolved tremendously during the last decade. It involves extracting latent information and insights from: (i) unstructured social media user generated content (UGC); and (ii) users' interaction in social media, such as communities or groups with similar interests and behaviour.

The availability of these large volume and heterogenous data streams pose significant challenges to typical data mining algorithms. Existing relevant approaches mainly focus on general categorization of social media content [Zubiaga15]. NLP approaches have been utilized for assessing linguistic features [Duan12] and performing sentiment analysis [Kanavos14], [Santos14]. Although a lot of work has focused on the prediction of emerging topics and the identification of current trends [Xie16], real-time topic and trend detection still remains an issue. However, in several tasks a more fine-grained context and location-specific categorization were deemed necessary [Dong15], [Unankard15]. Identifying experts on a certain field using crowdsourcing can further improve the efficiency of these tasks and has also been addressed extensively [Ghosh12], [Brem15], [Bozzon13]. Users classified as experts are expected to provide cleaner data input and thus allow the crowdsourcing analytics process to produce more solid results [Rjab16]. While using crowdsourcing for data-mining is quite popular, the quality control of its outcome still has some drawbacks [Xintong14].

A fine-tuned combination of the sources above under a new framework, that can provide an improved semantics-aware crowdsourced observatory for specific thematic related terms and qualitative identification of current trends and topics, is utilized in the present work. This approach places emphasis on identifying the proper data schemas which will sustain the components required for offering a complete crowdsourced topics observatory. The data schemas proposed support an effective topic observatory design and implementation with a set of inter-linked components. These components complement

a framework which can be easily accessed and which is human-friendly in terms of interpreting the derived knowledge.

The proposed crowdsourced topics observatory highlights patterns relevant with a given thematic and it is adaptive and customizable depending on a given set of terms and hashtags. Emphasis is also placed on appropriate visualization of social media driven 'wisdom of the crowds' with highlighting topics perception as it is expressed in social media interactions and content threads. Advanced data visualizations are proposed in this observatory (such as word clouds, geolocations maps, etc.) to allow users to gain an evidence of the social data topics correlations providing zoom in and filtering capabilities, topic-level details, etc. This topic observatory allows users to easily comprehend thematic relationships and topics inter-dependencies.

Particular technology mature solutions, in Machine Learning (ML) and Natural Language Processing (NLP) areas, are exploited and advanced to deliver the required components that support the proposed crowdsourced topics observatory. The developed framework aims at offering a global and open interoperable environment among multiple stakeholders since it provides access both to topics summaries and to an open API, enabling various users' interactions. At the same time, it is easily customizable per thematic case and it follows a robust social media content processing pipeline to enable spotting of simple and sophisticated thematic correlations, trends and phenomena. By identifying qualitative content (in terms of readability and interestingness) and further classify it in a set of context-specific thematic categories, the topics correlations are evident and well aligned with their intensity and dynamics in the underlying collected crowdsourced dataset.

The remaining of this paper is organized as follows: Section 2 discusses the related work with emphasis on social media data driven topics and thematic detection and analysis. Section 3 summarizes the proposed social data schemas and objects design, while Section 4 discusses the proposed components and their data schemas functionality. Section 5 demonstrates the results of the proposed topics observatory testing with emphasis on a thematic related with social innovation. Finally, Section 6 includes the conclusions of the article and future potentials.

III. Related work

The challenges posed by the massive data sizes and low-quality content (e.g. poorly formatted, short textual entities, etc.) in social media platforms require advanced data modelling and analysis approaches. Existing relevant approaches focus on content classification (such as Twitter data threads) into categories such as substance, status style, social [Rampage10]. NLP approaches have been developed for assessing the expressed sentiment which may be more engaging for readers and can indicate their engagement capacity [Hoang13]. Moreover, the differences in privacy perceptions between users of different OSNs were studied and both Facebook and MySpace OSNs are concerned about privacy, yet this does not prevent them from sharing information online [Zhang15]. It is believed that often the perceived benefits of users outweigh the risk of personal privacy [Li14], [Debatin09]. Since users expose their views in open social platforms and in free forms, a more fine-grained and possible context-specific topics and thematic categorization is required. Such an approach requires the utilization of external domain specific knowledge provided by online sources (such as DBpedia and Wikipedia) or ontologies. For example, in [Gattani13] a Wikipedia-based global "real-time" knowledge base is utilized to automatically classify streaming data from Twitter. In addition, there are quite a few efforts for defining ontologies about products / brands that are used by such Linked Open Data efforts, such as The Product Types Ontology, the GoodRelations ontology, schema.org, and even the DBpedia and Freebase ontologies / folksonomies. The Product Types Ontology provides high-precision identifiers (ca. 300,000) for product or service types based on Wikipedia, extending the schema.org and GoodRelations standards for e-commerce mark-up. Therefore, social media topics related either to the company, the products and/or product types can be linked to these open classification standards to provide an improved semantics-aware repository for a thematic set of related terms.

Topic modelling techniques applied on social media content have also been researched in crowdsourcing applications. A two-stage hierarchical topic modelling system to address the clustering of noisy and sparse content from tweets [Wang2017], while topic modelling on Instagram hashtags was utilized to predict the subject of related images and improve image annotation [Argyrou2018].

The proposed work leverages on a terms-based crowdsourcing framework by building on advanced data modelling and mining approaches, to establish a robust social media content processing pipeline that will be able to identify qualitative content, and further classify topics in a set of context-specific categories. Analysis of the metadata and information around the collected UGC via topic modelling approaches are accompanied by ontology-based classification approaches to extract context along with features that will enable categorization in appropriate, predefined classes (such as crowdsourced ideas, thematic taxonomies, etc.). The proposed methodology for automatically collecting and identifying content relevant to specific topic areas acquires coarse-grained streams of content (such as posts and annotated multimedia and text content, as well as user reactions to them) via appropriate data collection mechanisms. Then, the identification and characterization of qualitative relevant information and the fine-tuning of the data collection parameters is followed exploiting and advancing the state of the art in the areas of qualitative crowdsourcing and on dynamic topic categorization. The proposed approach establishes a closed loop between a topics observatory taxonomy and a dynamic topic categorization model of social media content along with semantics annotations from existing available open data.

IV. Social Data schemas and objects design

The volume and velocity of data produced in Social

Networks increase at a very high pace and appropriate data objects and data models design are required. Due to the unstructured nature of social media posts, NoSQL database schemas are typically proposed. This work builds on a data schema which favors a document-based NoSQL database, since such technologies use an effective data model. The proposed data model considers that each record and its associated data forms an entity called "document". This document entity encapsulates all information related with the database object and it thus offers a generic model for multiple social media data integrations.

Data encoding should also follow an appropriate approach to match with this generic data model. Already some of the popular social media platforms have opened information about their preferred data models. For example, the Twitter API offers tweet data encoded by using Object Notation based on a particular technology (JavaScript – JSON, n.d.). JSON is based on key-value pairs, with named attributes and associated values. These attributes, and their state are used to describe objects. Such data object encodings match well with the NoSQL databases which perform very well in the handling of JSON encoded objects. The proposed social media knowledge observatory enables streaming data collection from Twitter and other social media and thus, the data schema proposed temporarily stores streaming data posts in accordance to such official Twitter schemas (e.g. as with Fig. 1 visualizing a Twitter object).

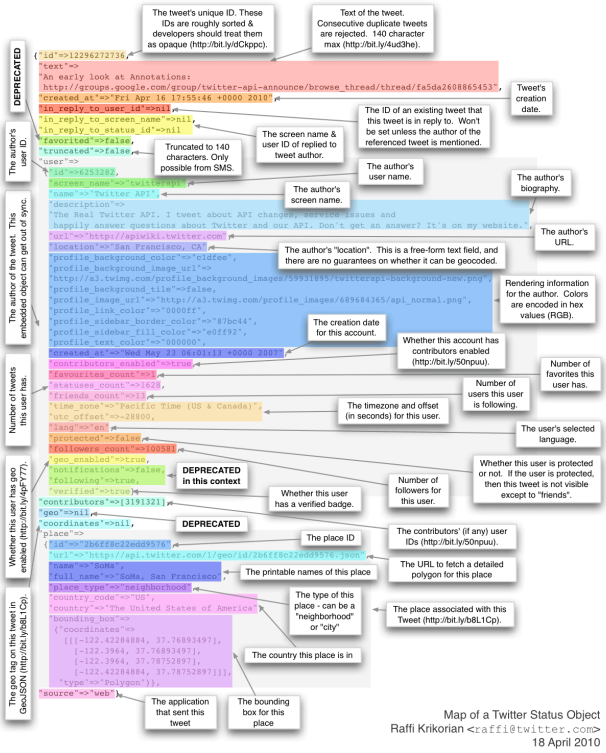

Figure 1: A sample of Twitter's JSON data object

The Twitter's JSON object contains (more than 150 attributes. However, only a small subset of these attributes is typically useful for any filtering process. In the proposed work we follow a flexible approach by using simple attributes with emphasis on those which embed information that could lead to topics detection. As depicted in Figure 2, social media data streams received by popular social media platforms are utilized by harvesting selected few and important attributes (such as geolocation information - Geo, Tweet timestamp, ids, hashtags etc.). These attributes are then used to support a filtering process which enables connections with specific and focused topic related components in the proposed observatory. The proposed observatory captures multiple users and content attributes to result in adaptive and effective data schemas which can support all of the proposed components.

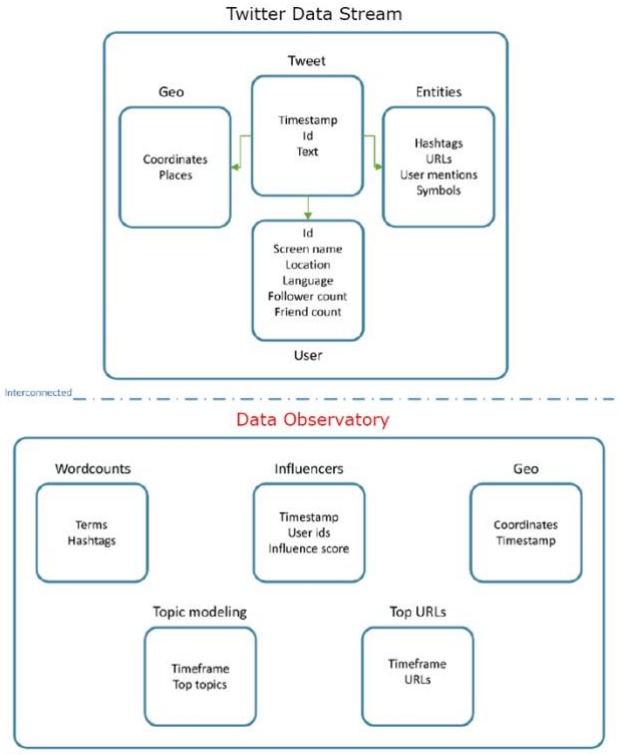

Figure 2: Data streams and their selected attributes for the Topics Observatory

V. Data Driven Components for a Crowdsourced Topics Observatory

The proposed design for a crowdsourced topics observatory is based on a flexible data schema model which involves specific sub components that are inter-connected but each with its own data schema used for storing its data filtering process results. All of the available components are detailed in the next subsection.

A. Topic Observatory components

The topics observatory components focus on content visualization (Wordclouds, Locations, Topic Modelling), influential content identification (Top Posts, Top URLs, Influencers) and discovery of crowdsourced open-licensed media related with a specific thematic (Repositories).

Wordclouds

The Wordclouds component is proposed and designed to visualize the most frequent hashtags and terms in the collected

data. Word clouds for both terms and hashtags are calculated on a frequent (e.g. daily) basis, for each keyword category. The results are stored in a JSON document, which contains only limited and necessary attributes. The data schema of the word clouds results is presented in Fig. 3 and includes the following attributes:

Figure 3: Wordclouds data schema

- Timestamp from [tweets with timestamp equal or higher than]
- Timestamp to [tweets with timestamp lower than]
- Language
- Keyword category (which category of keywords is responsible for collecting this tweet)
- All terms in this period and their frequency
- All hashtags in this period and their frequency

The terms and hashtags attribute actually represent a key – value pair, where the key is the term or hashtag and the value is the number of appearances of the specific term/attribute in this date range. The "timestamp from" – "timestamp to" period specify the duration of the analysis (e.g. they would vary 1 day if all word clouds are calculated in a daily basis). When a user performs a query for a certain date range, the system returns an aggregation of all the documents within this period. For example, if a user requests for the word clouds in the date range of July 1st – July 15th, the hashtags and terms word clouds included in the documents that are within this date range sum up and a final word cloud with all the hashtags and terms is generated. In general, the proposed schema of these attributes is depicted in Fig. 4.

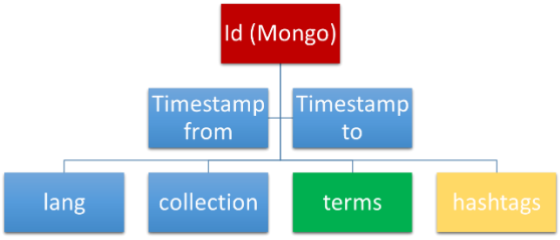

Figure 4: JSON schema for Wordclouds

Locations

The Locations component covers two tasks providing: a) a map visualization of the places with the highest crowd posting activity (such as tweeting) regarding the topic of interest and b) information about the terms and hashtags used more frequently in an area selected by the observatory user. For these tasks the data schema requires the following attributes:

- Geographical coordinates (Latitude, Longitude)
- Hashtags included in the post (if available)
- Terms included in the post's text
- Keyword category (which category of keywords is responsible for collecting this post)
- Timestamp
- Language

In task a) the user may have the option of selecting a keyword category, a certain time range and a specific language, while in task b) the user may also select a certain area in the map and retrieve the terms and hashtags with the highest frequency in it. The data schema used to store the results for this component has been designed in such a way, so that a single response to task a) also fulfills the requirements of task b) and vice versa. It is obvious that the data schema remains the same, compared to the one used in the Wordclouds component, with an addition of an array that contains the geographical coordinates of the tweet (Latitude, Longitude). The Locations data schema is described in Fig. 5.

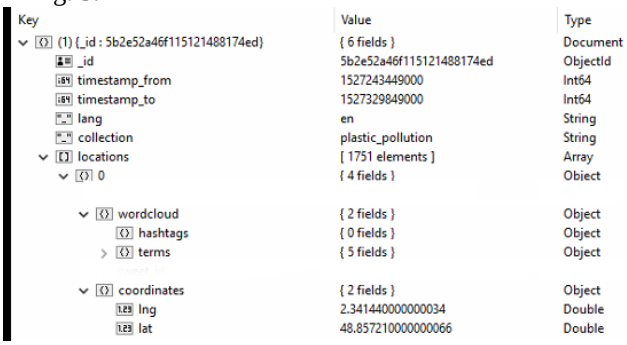

Figure 5: Locations data schema

Topic Modelling

Topic modelling is used to extract related sub-topics under the umbrella of a main theme. In our approach we utilized the probabilistic topic modelling method called Latent Dirichlet Allocation (LDA)[Blei2003] such that the data schemas do not store anything more than the time period in which the training has been done, the language, the keyword category and the path to the model that has been produced. LDA uses a bag of words approach to transform user's corpus into a vector of word counts. After the post's text is pre-processed for cleanup, a stemming process is applied to reduce the words in their base root. Then the training of the model occurs, while the next and final step is the evaluation of the results. A topics-to-words matrix for each collection is generated and the end-user can examine the words with the highest weight inside a topic. The component's data schema is provided in Fig. 6.

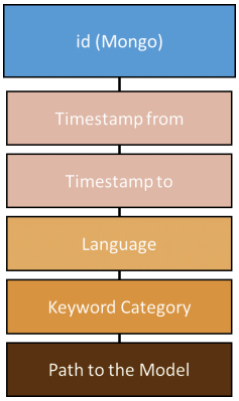

Figure 6: Topic Modelling data schema

Top Posts and Top URLs

The proposed topic observatory also includes a component which produces a list with the most propagated URLs (included in social media posts) and a list of the top posts (based on the number of retweets in the case of Twitter). The official Twitter JSON document contains an object called "entities" which includes potential URLs added to the original tweet. A process running in the back-end keeps track of the URLs posted each day. Following a similar approach to the one demonstrated in the Wordclouds component, the results of the Top URLs use the data schema depicted in Fig. 7.

Key	Value	Type
(4) {_id : 5b6ac4956f1151319c2a2ddb}	{ 6 fields }	Document
_id	5b6ac4956f1151319c2a2ddb	ObjectId
timestamp_from	1521139464000	Int64
timestamp_to	1521225864000	Int64
lang	en	String
collection	innovations	String
top_urls	[0 elements]	Array

Figure 7: Top URLs data schema

The *top_urls* is supported by a key-value pairs array, where key is the URL and the value is the number of appearances in the time period defined by the timestamp from to the timestamp to values. The official Twitter JSON document also contains a *retweeted_status* object containing, amongst others, an attribute that counts the number of times the original post has been retweeted. Following the same methodology described earlier in the Top URLs component we end up with a similar schema for the Top Posts, with the only difference in the case of Twitter being that the "number of retweets" is calculated in a cumulative manner. Thus, instead of storing each occurrence we only store the highest to avoid duplicates.

Influencers

The Influencers component is designed to identify the top k (top 100 for example) influencers in a given crowdsourced dataset. To discover these users, a social graph model is proposed based on several metadata (such as the retweets, mentions and replies) among all users for whose posts been collected. To generate a social graph that can produce reliable results regarding the detection of expert users, a large number of nodes and edges is enabled. Thus, the influencer detection algorithm runs for a period of time (e.g. one month) and in an accumulative manner (e.g. two months, three months, etc.). This process runs periodically in the back-end part of the designed crowdsourcing observatory. The results of the influencer identification process follow the next data schema (see Fig. 8):

Key	Value	Type
(14) {_id : 5b6acfbc6f11510d587d5e96}	{ 6 fields }	Document
_id	5b6acfbc6f11510d587d5e96	ObjectId
timestamp_from	1532108664000	Int64
timestamp_to	1532972664000	Int64
lang	en	String
collection	innovations	String
influencers	[12 elements]	Array
0	2327035620	String
1	122030887	String
2	462152975	String
3	2438302135	String
4	3004320047	String
5	626261903	String
6	2327277326	String
7	824984548677185536	String
8	258710092	String
9	1205042994	String
10	16900850	String
11	932594007489826816	String

Figure 8: Influencers data schema

The data schema includes the following attributes:

- Timestamp from
- Timestamp to
- Language
- Keyword category
- List of identified influencers

Repositories

In this component various media streams, associated with the specific thematic under examination and knowledge extracted previously, are captured and presented to the end-users in real time utilizing the official APIs of the respective services. Social media platforms such as Thingiverse (hosting open 3D printer designs) and Flickr (photos sharing) are utilized. Additionally, an Instagram crawler locates posts associated with predefined hashtags. These hashtags are associated with activities that are immediately linked with the thematic of interest for the observatory offering handpicked high-quality content to the user.

B. Topic Observatory Open API

To further facilitate the use of the proposed topic observatory and strengthen its interoperability, an open API is designed and implemented exploiting REST API technologies. Registered users of the observatory are able to make calls to this open crowdsourcing API and receive the information in response to their request. The user is enabled to make API calls to certain endpoints of the API and retrieve a JSON object, according to the called endpoint. Figure 9 presents all available endpoints and the description of the retrieved JSON object response provided back to the caller.

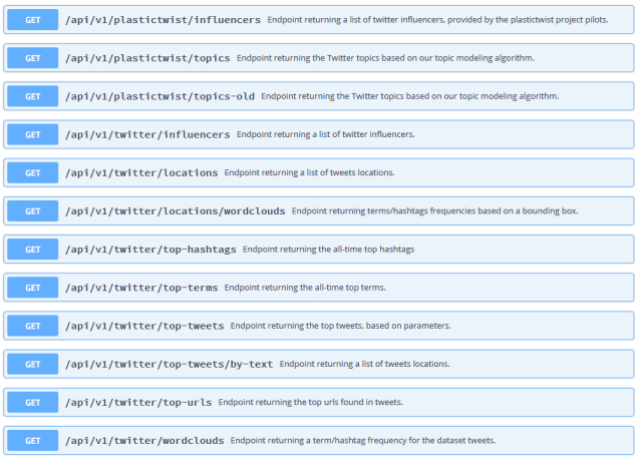

Figure 9: Crowdsourcing API endpoints

Our endpoints were semantically mapped based on the social media data that users would like access to. All the endpoints follow a structure that can be easily matched by the user to the corresponding developed components of the observatory. For example:

- /api/v1/twitter/top-urls
- /api/v1/flickr/top-photos
- /api/v1/thingiverse/top-things

Since the present work handles social media data which are semi-structured and not model-centric, there was not a need for complex frameworks with various drivers and tools built-in. Instead we opted for a lightweight solution ensuring high performance and flexibility. Therefore, the development of the API was completed using Python's Flask with the Flask-Restful extension.

VI. Experimentation on a Social Innovation Platform

The proposed components have been developed and tested under a topics observatory platform which focuses on the sensitive issue of plastic waste revaluation, aiming to promote collective awareness and social innovation. Indicative datasets outcomes are presented next with a focus on the topics related with plastic pollution and reuse.

Wordclouds

Figures 10-11 showcase some results from the wordclouds visualizations regarding plastic innovations and plastic pollution from prosts in Twiteer in a two month period.

Figure 10: Tagcloud for the plastic innovations category

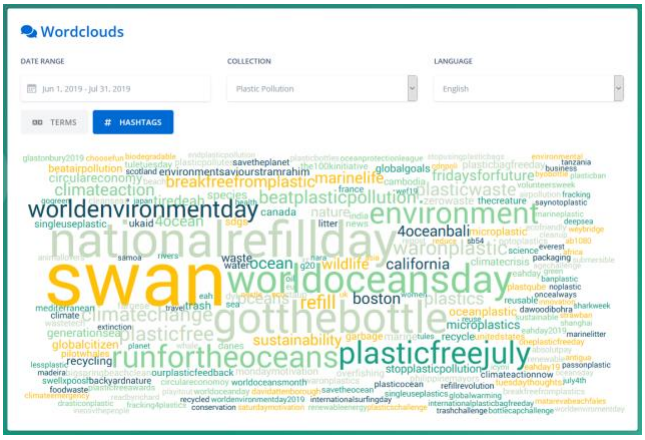

Figure 11: Tagcloud for the plastic pollution category

The observatory allows users to chose all the desired parameters (date range, collections, language) present in the data schema of the Wordclouds component to output the requested visualizations. Moreover, users can proceed with selecting words of their interest that are present in the generated cloud to explore the top social media posts that contain those terms (see Fig. 12). Thus, exploiting an interlink between various components, Wordclouds and Top Posts in this case, that encourages further user engagement and awareness.

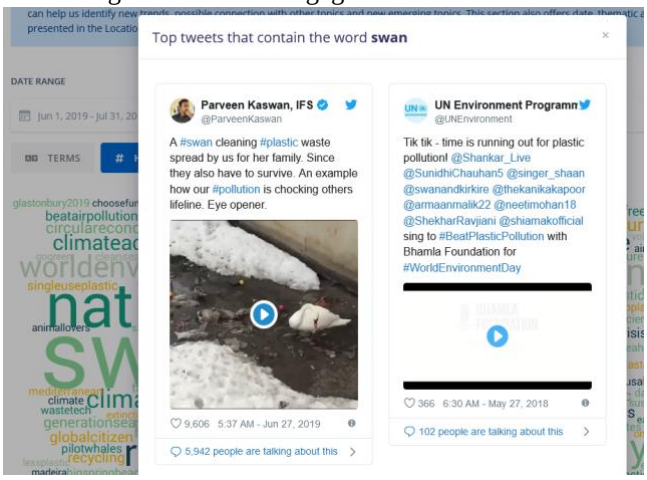

Figure 12: Interconnection between Wordclouds & Top Posts

Locations

Social media posts collected in the Locations sub component originate from users that share either their exact or approximate location or they state their location in their social network profile. In our case, where Twitter is utilized as data source, tweets that can not be associated with any geographical information are discarded. The total number of tweets collected, along with their geo-information grouped at city level, is aggregated and the results are displayed over a world map where users can pan and zoom in real time (Fig. 13). Again, users can define all the parameters (date range, collections, language) present in the data schema of the Locations component to generate the desired output.

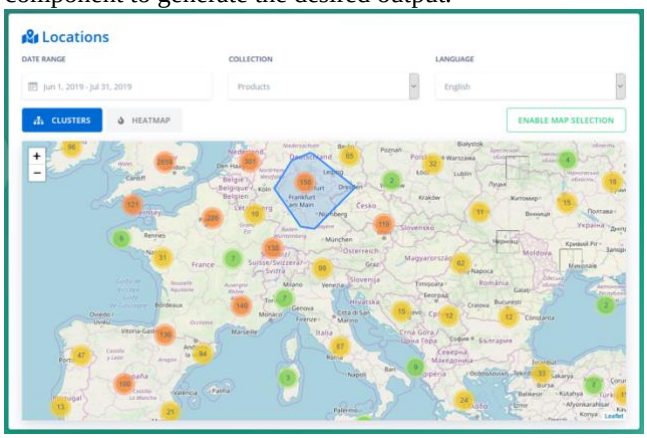

Figure 13: Dynamic map displaying terms density over geographical regions

This module can also interact with the Worldclouds component when a user selects a specific geographical region on the map to view the popular terms and hashtags of that location (Fig. 14).

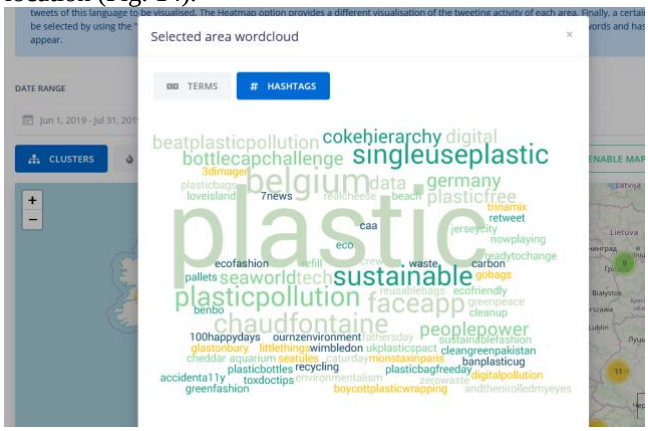

Figure 14: Tag cloud generated for a user selected geographical region

Topic Modelling

In this component users will gain an overview of the social media content and what they should expect to find in each of the available collections. After selecting the collection, they are interested in, they can explore the discovered topics in an interactive graph that also displays information regarding the frequency of the terms for each topic structured in a histogram (Figures 15-16). Topic Modelling covers the entire time span of the collected social media data.

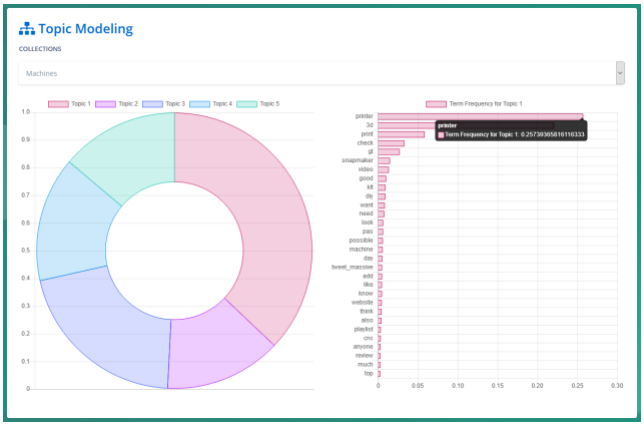

Figure 15: Identified topics from the plastic machines collection

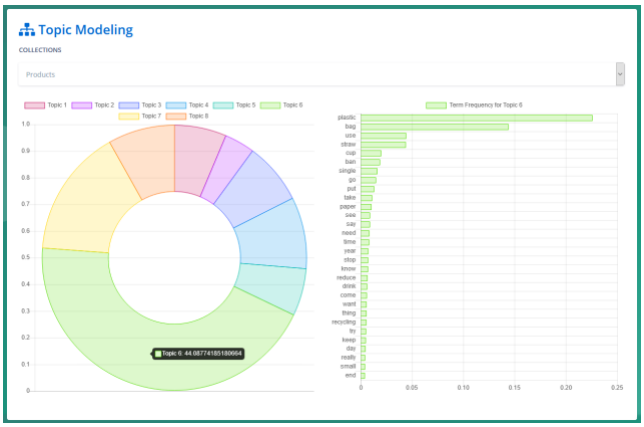

Figure 16: Identified topics from the plastic products collection

In Fig. 15 the prevailing identified topic of the *plastic machines* collection includes terms related with 3D printers and Snapmaker[1], popular among the makers community super-compact 3-in-1 machine supporting 3D printing, laser-engraving and CNC carving. These results are highly correlated with the discovered topic since 3D printing is one of the most used technology in plastic upcycling.For the *plastic products* collection the most significant topic is related with single-use products like plastic bags, straws and cups (Fig. 16). Single-use plastics, or disposable plastics, are a major concern in the discussion surrounding plastic pollution since their useful lifecycle is very short and they end-up in the environment as plastic waste really quick.

Top Posts

The quality of social media content can be tracked by filtering out posts based on other users' interactions. For Twitter, these interactions refer to the number of times a post has been retweeted, replied or marked as favorite. To select the top collected tweets, a ranking by retweets, favorites and replies count is performed and presented to users (Fig. 17).

[1] https://snapmaker.com/

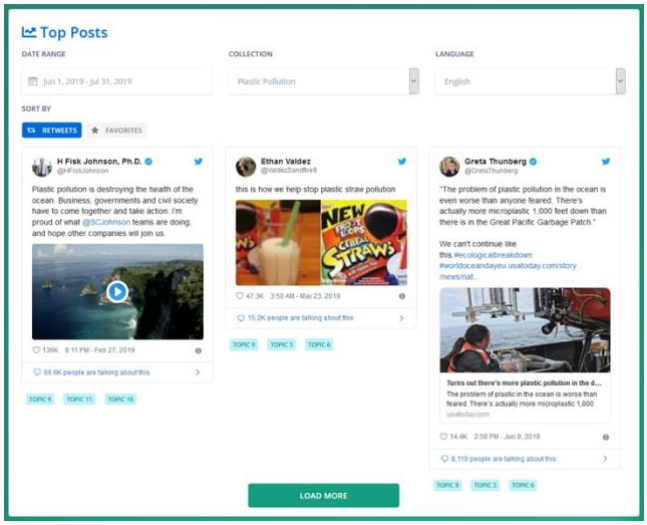

Figure 17: Top posts by retweets in the plastic pollution collection

The displayed top posts are also associated with the discovered topics from the Topic Modelling component. Pressing the buttons under each post will transfer the user to the respective identified topic for further exploration of the correlated terms under the specific topic.

Top URLs

Since social media posts often contain web links, these are also filtered by calculating their appearance frequency in the collections of every month and then sorted in descending order to determine the top URLs (Fig. 18).

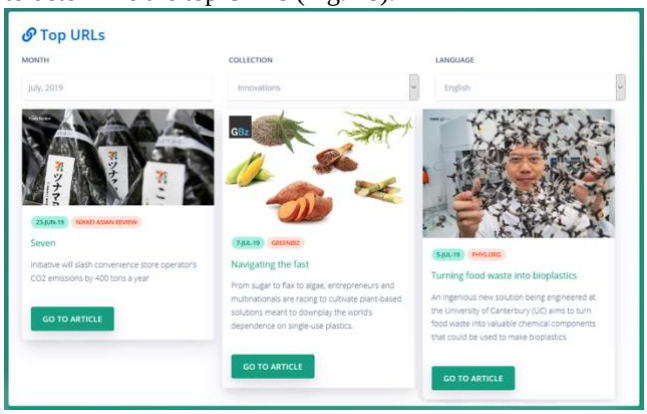

Figure 18: Top URLs in the plastic innovations collection

Influencers

This component supports the detection of influencers (users that diffuse information related to the subject of interest more efficiently throughout the network) among the Twitter social network using either the NetShield algorithm [Tong10] or the betweenness centrality metric [Kitsak10]. Users select the criteria they prefer (time period monthly based, collection, language and algorithm) and are presented with a list of the most influencing Twitter accounts (Fig. 19).

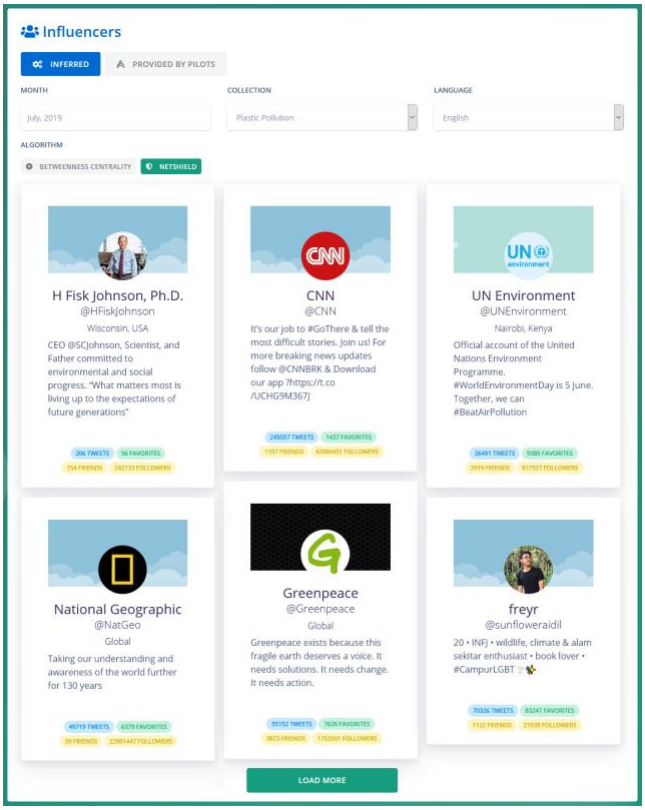

Figure 19: Top influencers regarding plastic pollution by NetShield

Repositories

The following Figures 20-23, contain examples of related external media content that is discovered and presented to the users via the crowdsourced observatory. For Thingiverse poplar, featured and new open 3d printer designs can be displayed while for Flickr users are asked to select one of the top hashtags, as determined by the observatory, or provide their own to discover related photos.

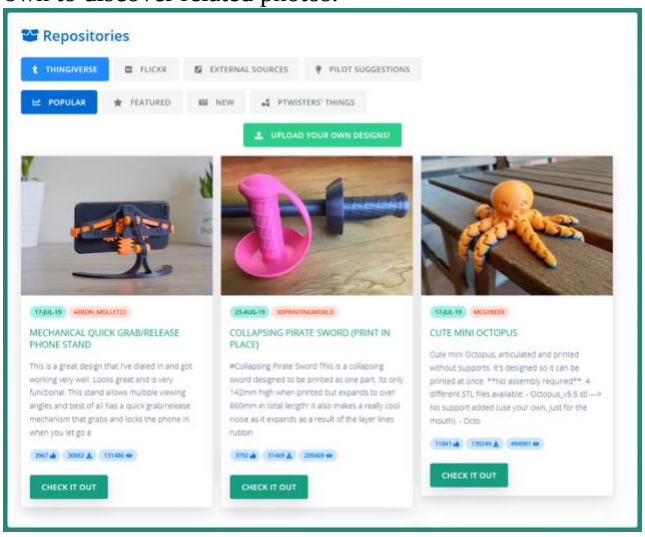

Figure 20: Popular open 3d printer designs from Thingiverse

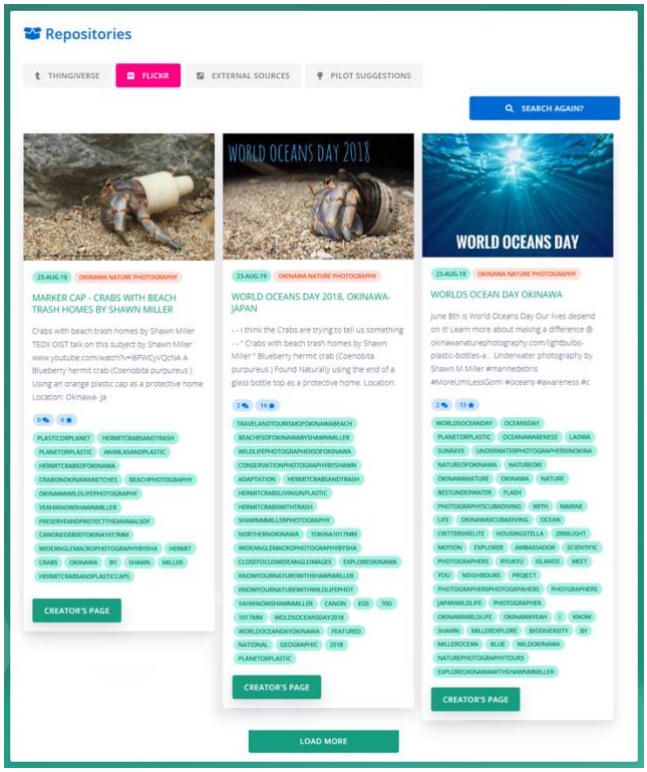

Figure 21: Flickr photos related with the #worldoceansday hashtag

Posts with predefined hashtags related with the observatory's thematic are collected and visualized through the Instagram feed (Fig. 22).

VII. Conclusions and Future Work

In the present work appropriate data schemas and a flexible and effective data model have been specified in order to successfully incorporate crowdsourced content from heterogenous social media sources. The presented document-based data model has allowed us to address challenges due to the high volume and velocity of social media data and has been exploited successfully towards the developing of a crowdsourced topics observatory. This dynamic observatory, that utilizes interactive visualization and advanced topic modelling methods, supports effectively multiple social media data schemas under a single thematic. Emphasis was placed to the design of the data schemas to successfully describe the unstructured social media content and users' interaction in social networks while allowing for efficient interoperability with the observatory's visualization applications. The fine-tuned combination of the available social media resources under various appropriately adapted approaches allows to efficiently capture the dynamics, trends and patterns relevant with the given thematic, depending on an initial set of keywords and a given hierarchy. The implementation of a streamlined user interface with advanced visualizations offers a high-level experience and customization capabilities where the user is able to explore and delve into the collective intelligence of the crowd offered.

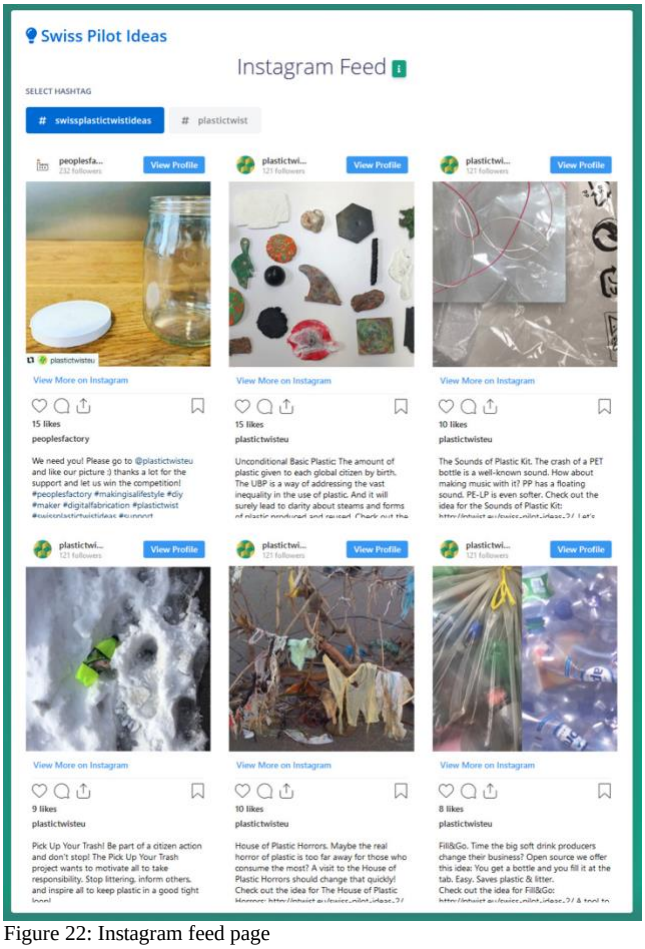

Figure 22: Instagram feed page

ACKNOWLEDGMENTS

Parts of this work have been supported by the European Union's Horizon 2020 Research and Innovation Programme under Grant Agreements No. 780121 and No. 871403.

REFERENCES

[1] Argyrou, A., Giannoulakis, S. and Tsapatsoulis, N. Topic modelling on Instagram hashtags: An alternative way to Automatic Image Annotation? *13th International Workshop on Semantic and Social Media Adaptation and Personalization (SMAP)*, Zaragoza, 2018, pp. 61-67.

[2] Blei, David M., Ng, Andrew Y. and Jordan, Michael I. 2003. Latent dirichlet allocation. J. Mach. Learn. Res. 3, (3/1/2003), 993–1022.

[3] Bozzon, A., Brambilla, M., Ceri, S., Silvestri, M., and Vesci, G. 2013. Choosing the right crowd: expert finding in social networks. *In Proceedings of the 16th International Conference on Extending Database Technology*, ACM, March 2013, pp. 637-648.

[4] Brem, A., and Volker B. 2015. The search for innovative partners in co-creation: Identifying lead users in social media through netnography and crowdsourcing. *Journal of Engineering and Technology Management* 37 (2015): 40-51.

[5] Debatin, B., Lovejoy, J. P., Horn, A. K., and Hughes, B. N. 2009. Facebook and Online Privacy: Attitudes, Behaviors, and Unintended Consequences. *Journal of Computer-Mediated Communication* 15:83–108.

[6] Dong, X., Mavroeidis, D., Calabrese, F., and Frossard, P. 2015. Multiscale event detection in social media. *Data Mining and Knowledge Discovery*, 29(5), 1374-1405.

[7] Duan, Y, Chen, Z., Wei, F., Zhou, M., and Shum, H.-Y. 2012. Twitter Topic Summarization by Ranking Tweets using Social Influence and Content Quality. *Proceedings of COLING 2012*: 763-780.

[8] Gattani, A., Lamba, D. S., Garera, N., Tiwari, M., Chai, X., Das, S., Subramaniam, S., Rajaraman, A., Harinarayan V., and Doan, A.H. 2013. Entity extraction, linking, classification, and tagging for social media: a wikipedia-based approach. *Proc. VLDB Endow*. 6, 11 (August 2013), 1126-1137.

[9] Ghosh, S., Sharma, N.K., Benevenuto, F., Ganguly, N., and Gummadi, K.P. (2012). Cognos: crowdsourcing search for topic experts in microblogs. SIGIR.Proceedings of the 35th international ACM conference on research and development in information retrieval.

[10] Hoang, T.-A., Cohen, W.W., Lim, E.-P., Pierce, D., and Redlawsk D. P. 2013. Politics, sharing and emotion in microblogs. *In Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* (ASONAM '13). ACM, New York, NY, USA, 282-289.

[11] Kanavos, A., Perikos, I., Vikatos, P., Hatzilygeroudis, I., Makris, C. and Tsakalidis, A. 2014. Modeling ReTweet Diffusion using Emotional Content. *IFIP International Conference on Artificial Intelligence Applications and Innovations* (AIAI 2014), Rodos, Greece.

[12] Kitsak, M., Gallos, L.K., Havlin, S., Liljeros, F., Muchnik, L., Stanley, H.E., and Makse, H.A. 2010. Identification of influential spreaders in complex networks. *Nature Physics*, 6(11), 888–893.

[13] Li, Y. and Smeaton, A.F. 2014. From Smart Cities to Smart Neighborhoods: Detecting Local Events from Social Media, *ECIR'14 Information Access in Smart Cities Workshop*, (i-ASC).

[14] Rampage, D., Dumais, S., and Liebling, D. 2010. Characterizing microblogs with topic models. *In Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media*, 2010.

[15] Rjab, A. B., Kharoune, M., Miklos, Z., and Martin, A. 2016. Characterization of experts in crowdsourcing platforms. *In International Conference on Belief Functions*, September 2016, pp. 97-104. Springer, Cham.

[16] Santos, dos C., and Gatti, M. 2014. Deep convolutional neural networks for sentiment analysis of short texts. *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, August 2014, 69–78.

[17] Tong, H., Prakash, B. A., Tsourakakis, C., Eliassi-Rad, T., Faloutsos, C. and Chau, D. H. 2010. On the vulnerability of large graphs. *In Proceedings - IEEE International Conference on Data Mining*, ICDM, 1091-1096.

[18] Unankard, S., Li, X., and Sharaf, M. A. (2015). Emerging event detection in social networks with location sensitivity. *World Wide Web*, 18(5), 1393-1417.

[19] Wang, B., Liakata, M., Zubiaga, A., Procter, R. 2017. A Hierarchical Topic Modelling Approach for Tweet Clustering. *In: Ciampaglia G., Mashhadi A., Yasseri T. (eds) Social Informatics. SocInfo 2017*. Lecture Notes in Computer Science, vol 10540. Springer, Cham

[20] Xie, W., Zhu, F., Jiang, J., Lim, E. P., and Wang, K. 2016. Topicsketch: Real-time bursty topic detection from twitter. *IEEE Transactions on Knowledge and Data Engineering*, 28(8), 2216-2229.

[21] Xintong, G., Hongzhi, W., Song, Y., and Hong, G. 2014. Brief survey of crowdsourcing for data mining. *Expert Systems with Applications*, 41(17), 7987-7994.

[22] Zhang, J. 2015. Voluntary information disclosure on social media. *Decision Support Systems* 73: 28-36.

[23] Zubiaga, A., Spina, D., Martínez, R., and Fresno, V. 2015. Real-Time Classification of Twitter Trends. *Journal of the Association for Information Science and Technology, 66(3), 462–473.*

Causality Learning: A New Perspective for Interpretable Machine Learning

Guandong Xu, Tri Dung Duong, Qian Li, Shaowu Liu and Xianzhi Wang

***Abstract*—Recent years have witnessed the rapid growth of machine learning in a wide range of fields such as image recognition, text classification, credit scoring prediction, recommendation system, etc. In spite of their great performance in different sectors, researchers still concern about the mechanism under any machine learning (ML) techniques that are inherently black-box and becoming more complex to achieve higher accuracy. Therefore, interpreting machine learning model is currently a mainstream topic in the research community. However, the traditional interpretable machine learning focuses on the association instead of the causality. This paper provides an overview of causal analysis with the fundamental background and key concepts, and then summarizes most recent causal approaches for interpretable machine learning. The evaluation techniques for assessing method quality, and open problems in causal interpretability are also discussed in this paper.**

***Index Terms*—Interpretable Machine Learning, Causal Inference, Counterfactual Explanation, Causal Feature, Causal Interpretability**

I. Introduction

In the past decades, machine learning has achieved the impressive performance in diverse tasks, and is increasingly applied in science, society and business. However, most of state-of-the-art models remained incomprehensible for both researchers, users and engineers, causing difficulties when deploying in real world. Specifically, there are several high-stake decision-making domains such as self-driving cars, crime prediction or personalized medicine in which the lack of transparency in machine learning prevents themselves from being adopted. Take for instance, in the healthcare sector where each decision can affect the people's survival, physicians are frequently concerned about the safety and trust of any deployed models. They do not likely trust the model's prediction if they can not understanding the rationales behind it. Consequently, interpretability in machine learning plays a significantly important role in generating trust-worthy models. This furthermore allows researchers, data scientists and engineers to ensure the models following the human understanding, ethnic codes, fairness and security. We as human have an insatiable curious nature; thus, our goal is not only to understand models' mechanism but also to generate and extract new knowledge of the world.

In view of the time of explanation generation shown in Figure 1, interpretable machine learning can be divided into two branches: ad-hoc and post-hoc methods. The evolutionary history of noticeable traditional interpretable machine learning techniques is briefly described in the Figure 1. The ad-hoc type focuses on building the model architecture, algorithms or mechanisms that are self-explainable and transparent. Intrinsically interpretable models are the central research in the early years of artificial intelligence with the dominance of symbolism methods, followed by more advanced approaches such as decision sets [1], generalized linear regression, generalized additive model [2], [3], [4], Bayesian probabilistic model [5], [6], rule-based model [7], [8], attention mechanism [9], fuzzy inference systems [10], [11], [12], TabNet [9], etc. With the rapid growth of deep learning in recent decades, machine learning model is gradually evolved into complicated and incomprehensible form, which leads to the increasing attention on post-hoc interpretations. Several prominent approaches in this category include Local surrogate models (LIME [13], SHAP [14], LORE [15], etc), influence functions [16] and feature importance estimation [17], [18] have been introduced.

However, traditional interpretable machine learning focuses on the association instead of the causality. With the emergence of causal inference, an increasing number of causality-oriented methods have been proposed in interpretable machine learning. In comparison with traditional methods, causal approaches can be utilized to identify causes and effects of models architecture or conduct the reasoning over its decisions and behaviors. This article examines the overview of interpretable machine learning, presents the causal analysis in machine learning interpretability and finally discusses the future research directions. More specifically, we first present the background of causal analysis with key concepts, models and evaluation metrics. We then provide an overview of state-of-the-art works on causal interpretability. We also illustrate the potential evaluation metrics used in interpretable machine learning.

II. Causality Analysis

Causality analysis can exploit the causality mechanisms underlying the data-generating process, which is more advanced than the predictive or descriptive capability in machine learning techniques. Causal inference and causal discovery are two main research topics for causality analysis. The goal of causal inference is to estimate the causal effect of treatment (i.e., a decision made or action taken) on the outcome (i.e., the result of treatment). Causal discovery examines whether a set of causal relationships exists among the variables. This paper would primarily focus on causal effect, which is more correlated to machine learning interpretability.

School of Computer Science, University of Technology Sydney, Australia
[Guandong.Xu,Qian.Li,Shaowu.Liu,Xianzhi.Wang]uts.edu.au,
TriDung.Duong@student.uts.edu.au

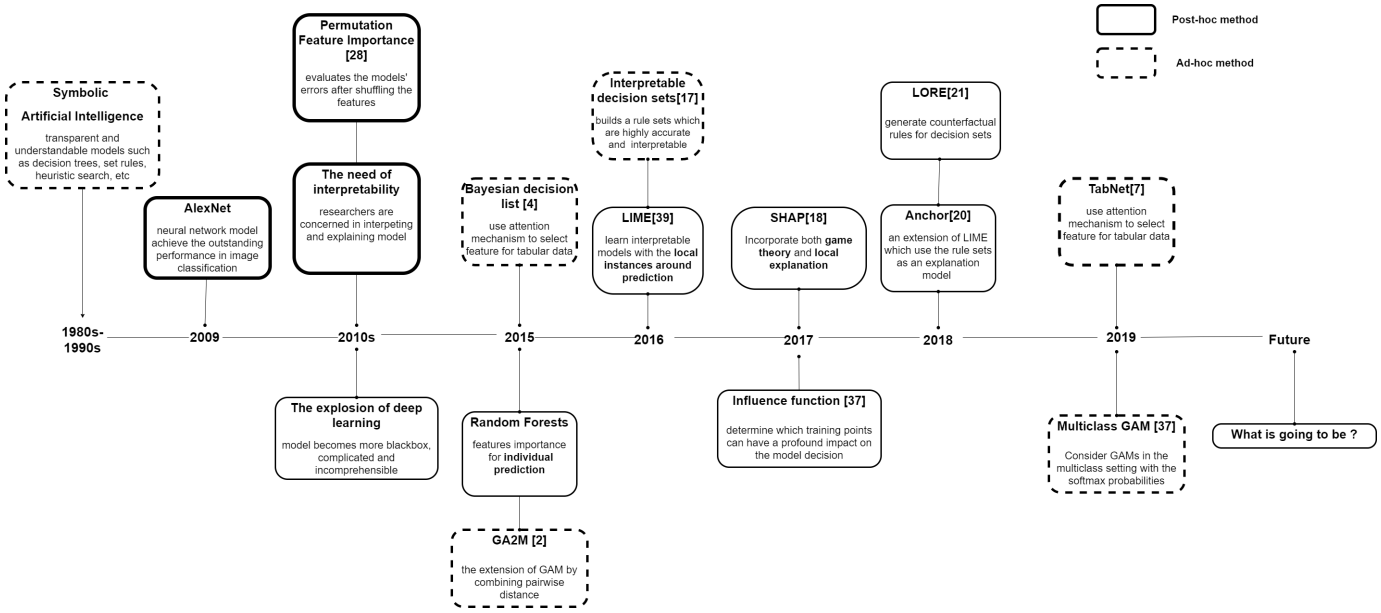

Fig. 1. The evolution of interpretable machine learning

A. Causal Inference

Causal inference has been widely applied in econometric, social science and medicine fields for evaluating the policy's effect or the drugs' side effect. Effect estimation is tied to the outcome caused by the treatment applied to an instance. An instance is the atomic research object, which can be a physical object or an individual person. Treatment and outcome are terms that denote a decision made or action taken and its result, respectively. We first introduce the essential concepts for learning treatment effect followed by the causal models.

- Covariates X refers to the background variables or features of the instance.
- Treatment T refers to the action (manipulation or intervention) that applies to a instance.
- Outcome Y is the result of the treatment applied on a instance.
- Confounder Z is a variable which causally affects both treatment and outcome.

To better understand causal inference, we give the following example combined with the notations defined above. To prove the efficiency of the medication on the disease, the scientist needs to assess its positive effect into the patients' recovery rate. Figure 2 depicts the corresponding causal relationships among the essential variables. The treatment T is whether the drugs are applied or not, and the observed features X are the patients' condition such as the level of insulin and cholesterol, heart rate, etc. Outcome Y is the recovery rate and age is the confounder Z. This is simply because age firstly determined the need of applying medication into patients, since the young people may not necessarily take the medicine. Age also affects to the recovery rate: the youth has a higher probability to recover than the elderly.

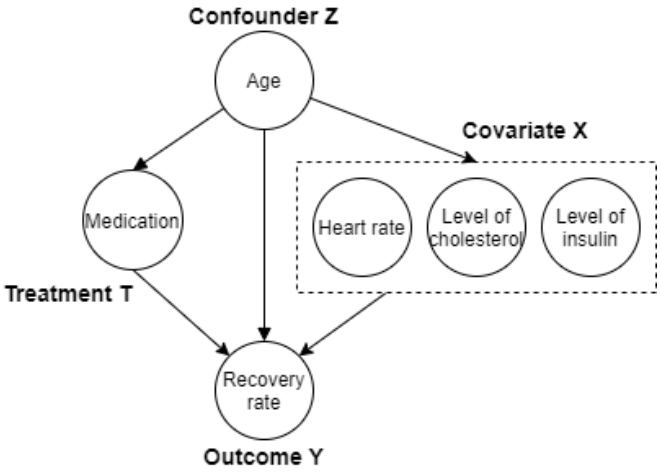

Fig. 2. The causal graph for recovery rate problem

B. Causal Models

We now introduce the two most important formal frameworks used for causal inference, namely the structural causal models and the potential outcome framework.

Structural causal model[19] consists of two main components: the causal graph and structural equations. Causal graph is the probabilistic graphical model which is used to represent the assumption about prior knowledge and data generating process. A causal graph is defined as $\mathcal{G} = \langle \mathcal{V}, \mathcal{E} \rangle$ where $\mathcal{V}$ is the set of nodes and $\mathcal{E}$ is the set of edges. Structural equation is a set of equations Eq. (1) which are used to represent the causal effect illustrated by the edge in the causal graph.

$$\begin{aligned} X &= f_X(E_X), \\ T &= f_T(X, E_T). \\ Y &= f_Y(X, D, E_Y) \end{aligned} \tag{1}$$

where E_X, E_T, E_Y are exogenous variables, which are in-

dependent from other models' variable, and are determined outside the model.

Potential outcome framework is proposed by Neyman and Rubin [20]. Considering binary treatments for a set of units, there are two possible outcomes for each unit. The unit will be assigned to the control treatment if $T = 0$, or to the treated treatment if $T = 1$. As a result, we denote two potential outcomes Y_0 and Y_1 as the results caused by $T = 0$ and $T = 1$, respectively. Importantly, only one potential outcome is observed corresponds to the assigned treatment T, and we call this as the observed (factual) outcome Y. The unobserved potential outcome refers to the counterfactual outcome. Given the treatment T_i, the relationship between the observed outcome Y and two potential outcomes are

$$Y_i = T_i Y_1 + (1 - T_i) Y_0 \tag{2}$$

C. Treatment Effect Metric

With the key concepts and causal models, the treatment effect can be measured at the population, treated group, subgroup, and individual level. For simplicity, we discuss the treatment effect under the binary treatment, and it can be easily extended to multiple treatments by considering multiple potential outcomes.

The individual treatment effect (ITE) is defined as the change of Y_0 and Y_1, while keeping the covariates X unchanged (i.e., condition on those covariates). For an instance i with covariates X_i, its corresponding ITE is

$$ITE(\boldsymbol{X}_i) = E[Y_1|X_i] - E[Y_0|X_i] \tag{3}$$

As only one potential outcome is observed, it is nearly impossible to estimate the effect at the individual level. A more feasible way is to measure treatment effect at the average level.

The average treatment effect (ATE) measures the treatment effect at the whole population level as

$$ATE = E[Y_1 - Y_0] \tag{4}$$

The average treatment effect (ATT) is for the group of instances with the treatment equal to 1, i.e., the treated group.

$$ATT = E[Y_1 - Y_0|T = 1] \tag{5}$$

Conditional average treatment effect (CATE) known as heterogeneous treatment effect is defined on the subgroup with the particular covariate $X = x$.

$$CATE = E[Y_1 - Y_0|X = x] \tag{6}$$

D. Tools for Causal Analysis

Several libraries or tools are available for causal inference. Examples including *Double Machine Learning* [21], *Meta-learners* [22], *Orthogonal Learning* [23], [24] have been supported by EconML, CausalML, DoWhy and CausalNex, whereas causal discovery methods including graph inference and pairwise inference are provided in Causal Discovery Toolbox. Meanwhile, TIGRAMITE is a novel framework for causal discovery in time series. We summarize the existing toolboxes in Table I.

III. Interpretable machine learning with causality

Pearl [25] argues that causal reasoning is indispensable for machine learning to reach human-level artificial intelligence, since it is the basic mechanism for humans to be aware of the world. As a result, causal methodology is gradually becoming a vital component in explainable and interpretable machine learning. However, most of current interpretability techniques pay attention to solving the correlation statistic rather than the causation. Therefore, the causal approaches should be emphasized to achieve a higher degree of interpretability.

A. Model-Agnostic Causality for Deep Neural Neworks

The traditional way to analyze Deep Neural Network is to build several models with different architectures and make a comparison between their performances. The problem is that re-training DNNs is computationally expensive, and infeasible when it comes to the complicated architecture. Inspired by causal model, several methods have been proposed to interpret neural network model.

Chattopadhyay et al. [26] define $\text{ACE}^{y}_{do(x_i=\alpha)}$ as the causal attribution of neuron x_i to the output neuron y_i, and $\mathbb{E}[y|do(x_i = \alpha)]$ as the interventional expectation Eq. (7). The polynomial function is selected to estimate this value.

$$\mathbb{E}[y|do(x_i = \alpha)] = \int yp(y|do(x_i = \alpha))dy \tag{7}$$

Narendra et al. [27] propose to construct a modified structural causal model as an abstraction of a DNN to make an reasoning over its elements. Thereafter, they rank each component based on their contribution to the final prediction for evaluation.

Based on TCAV [28] which generates a high-level concept-based explanation such as gender, race, background, others, the study in [29] evaluates the *causal concept effect* on a neural network prediction. They overcome the problem of do-operator by using Variational AutoEncoder (VAE).

Regarding Generative Adversarial Networks (GANs) interpretability, Bau et al. [30] proposes an approach for visualization and understanding at unit-, object-, and scene- level by estimating the causal effect of the models' interpretable components. There are two main steps in their approach: dissection and intervention. In the dissection step, the classes with the explicit representation are firstly identified. Thereafter, they make an intervention by forcing the units to be appeared and disappeared, and calculate its causal effect. Meanwhile, the authors [31] propose a causal framework to explore the intervention effect for proving that the components in images generated by GAN can be modified independently.

In terms of reinforcement learning, action influence model [32] is introduced for explaining the behavior of RL agents. They construct a modified structural causal model, learn the causal equation as the regression model during training the agent, and finally generate the contrastive explanation to answer the counterfactual question "Why does the agent choose action A instead of action B?".

B. Post-hoc Interpretability

Model-Agnostic explanations are particular challenging when the models' parameters have more complex relationships. To further aid the intepretability, the practitioners propose a variety of post-hoc interpretability methods to exploit what a trained model has learned, without changing the underlying model. Most widely useful post-hoc interpretation methods fall into two main categories: causal feature learning and counterfactual explanations, respectively.

1) Casual Feature Learning: Recent work on feature learning derives the subset of features that have causal contributions to the models' prediction. Early causal feature learning is to find the Markov Blanket (MB) containing a set of features which makes the target (T) independent from other features given MB(T). In the study [33], the authors firstly use the HITON algorithm [34] to derive the Markov Blanket, and thereafter deploy Max-Min Hill-Climbing (MMHC) algorithm to identify the causes and effects of the target variable. Given the number of transfer learning tasks D, Peters et al. [35] assume that there exists a subset of features X_{S^*} such that the conditional distribution $Y_k|X_{S*}$ is the same for different tasks k, and other settings Eq. (8). They propose an algorithm called subset search which samples the subset features, and then adopt the Levene test to assess the assumption.

$$Y_k|X^k_{S*} \approx Y'_k|X^{k'}_{S*} \quad \forall k, k' \in 1, ..., D \tag{8}$$

CXPlain [18] is the causal framework that can explain more complex machine learning models by estimating the feature importance. Granger-causal objective is introduced to quantify how much the exclusion of a single feature reduces model performance. Particularly, CXPlain trains a separate explanation model to any predictor f by optimizing a Granger-causal objective. CXPlain can also estimate the uncertainty of features importance by calculating confidence interval (CI).

2) Counterfactual Explanation: Counterfactual explanation is the example-based model-agnostic method which generates new instances that would change the models' prediction. The prominent example [36] in this research is that a person x with an annual income a and a current balance b has been rejected a loan by the financial institution, so how can they change their income and balance to a' and b' to receive the loan. Given the set of points P, to generate the set of counterfactual samples F, the objective function of counterfactual explanation [37] is to optimize the following function:

$$\begin{aligned} &\arg\min_{x} \max_{\lambda} (\lambda \cdot (\hat{f}(x') - y')^2 + d(x, x')) \\ &d(x_i, x') = \sum_{k \in F} \frac{\mid x_k - x'_k \mid}{MAD_k} \\ &MAD_k = \underset{(j \in \mathrm{P})}{\mathrm{median}}(|X_{j,k} - \underset{(l \in \mathrm{P})}{\mathrm{median}}(X_{l,k})|) \end{aligned} \tag{9}$$

where x is an original instance, x' is the counterfactual instance which close to x, y' is the target class label for x', λ is the regularized parameter, $d(x, x')$ denotes the distance between the original instance and the counterfactual samples, MAD_k is the median absolute deviation for feature k.

Grath et al. [36] extend $d(x, x')$ in Eq. (9) by adding a weight vector Θ. The vector Θ is used to evaluate models' feature importance, and can be obtained by many algorithms such as K-Nearest Neighbors or global feature evaluation. Dhurandhar et al. [38] combine the loss function generated from Convolutional AutoEncoder, while Arnaud [39] uses the prototypes function to ensure that the generated perturbation falls into the same distribution with the original data as well as increasing the computational speed without tuning too many parameters. Additionally, the counterfactual samples should be as diverse as possible; the study [40] proposes to use determinant of kernel matrix to illustrate this property.

To empower the capability of counterfactual explanations, constraints are considered in optimization problem of counterfactual explanation. Take for example, a person cannot decrease his age, or change his race and skin color. Recent work [41], [42] adopt Mixed Integer Programming (MIP) formulation to deal with categorical, numeric and mixed data type. Meanwhile, Artelt et al. [43] propose convex density constraints to generate counterfactual located in a region of the data space. Specifically, the density constraint $\hat{p}_y \geq \delta$ denoted by a kernel density estimator or a Gaussian mixture model is added into the distance function $d(x, x')$.

CERTIFAI [44] proposed by Sharma et al. as a novel and flexible approach which can be used in any type of data. CERTIFAI uses the customized genetic algorithm to choose individuals that have the best fitness scores defined as follows.

$$\begin{aligned} &fitness = \frac{1}{d(x, x')} \\ &d(x, x') = \begin{cases} \frac{n_{x'on}}{n} l_1(\mathbf{x}, \mathbf{x}') + \frac{n_{cat}}{n} simp(\mathbf{x}, \mathbf{x}') & \text{tabular data} \\ \frac{1}{SSIM(x, x')} & \text{image data} \end{cases} \end{aligned} \tag{10}$$

For tabular data, CERTIFAI chooses l_1 norm for continuous features and a simple matching distance for categorical features (simp). For image data, Structural Similarity Index Measure (SSIM) [45] measures the similarity of what humans consider. n_{con} and n_{cat} are the number of continuous features and categorical features, respectively.

Instead of identifying the minimum changes leading to the desired outcome, a new line of counterfactual explanations provides feasible paths to transform a selected instance into one that meets a certain goal. FACE [46] proposed by Poyiadzi et al. constructs a graph over the data points with the weights illustrating the feasible degree to transit between two vertices. FACE thereafter can be solved by the *Disjstra* algorithm to find the shortest path from the original instance to the counterfactual one.

C. Visualization of Causal Effect

Visualization-based method is another commonplace approach for quick understanding what the models have learned. Partial dependence plot (PDP) [47] depicts the marginal effect of features into the predicted outcomes. The partial dependence function is defined as:

$$\hat{f}_{x_S}(x_S) = E_{x_C}\left[\hat{f}(x_S, x_C)\right] = \int \hat{f}(x_S, x_C)p(x_C)dx_C \tag{11}$$

Zhao et al.[48] use Partial dependence plot (PDP) an its extension called Individual Conditional Expectation (ICE) to extract the causal information from machine learning model. These visualization tools allow to measure the predictions' change after making an intervention, which can help to discover the features' causal relationship.

IV. Evaluation

Evaluation in causal interpretability is an extremely difficult task currently, since there are nearly no groundtruth data to evaluate the methods' performance. Evaluation for traditional interpretable machine learning evaluation can be classified into three categories [49]: application-based, human-based and function-based. We apply the same category and focus on evaluations that can be used in causal interpretability.

A. Application-based

In real-world scenario where the machine learning model is deployed to assist experts, application-based evaluation illustrates how well the models provide explanations to human experts for improving their performance in specific tasks. Take for example, a randomized experiment [50] is conducted among a group of learner to solve the problems. They then rate the explanation generated by the machine learning models. With the assistance of models, the performance of people in different tasks is proved to be improved.

B. Human-based

Human-based evaluation methods refer to evaluate the performance of interpretable models with the assistance of human. Madumal et al. [32] generate explanation for the reinforcement learning. They implement an RL agent, and conduct an experiment running on StarCraft II, a strategic game, with 120 participants. *Explanation Satisfactory Scale* [51] is defined as the degree of human understanding of the AI system to measure the quality of generated explanations.

C. Function-based

Functional-based evaluation methods can be carried out without the assistance of human to evaluate the performance of the explanation model. There are some evaluation procedures for different techniques in Section IV:

1) Causal Interpretability for DNN: The lack of ground truth for feature effect makes it challenging to evaluate the performance of causal effect estimation. Chattopadhyay et al. [26] compare the salient map [52] generated by causal attribution method with Integrated Gradient [53]. Harradon etc al. [54] identify the components having the significant causal effects into the individual prediction. Specifically, they conduct the experiments in three different architectures VGG 19 in Birds200, VGG 16 and 6-layer cov network applied in Inria dataset. Thereafter, they make a query for an individual input, and then visualize top k variables according to their causal effect.

2) Counterfactual Explanations: A previous research [40] suggests that there are three main metrics to evaluate the counterfactual explanation: *proximity*, *diversity* and *sparsity*. The *proximity* is to reflect the similarity between the CF examples and the original one which was calculated as the mean proximity all over the examples. Meanwhile, the *diversity* measures the mean of the distances between the pairs of samples, ensuring that the generated instances should be as diverse as possible. Finally, the *sparsity* is the average number of changes converting CF examples to the original one.

$$\begin{aligned} proximity &= \frac{-1}{k}\sum_{i=1}^{k} dist(x_{cf_i}, x) \\ diversity &= \frac{1}{C_2^k}\sum_{i=1}^{k-1}\sum_{j=i+1}^{k} dist(x_{cf_i}, x_{cf_j}) \\ sparsity &= 1 - \frac{1}{k \cdot d}\sum_{i=1}^{k}\sum_{l=1}^{d} 1[x_{cf_i}^l \not\equiv x_i^l] \end{aligned} \tag{12}$$

with x_{cf} and x are the counterfactual samples and original instance, respectively, $dist(x_{cf_i}, x_{cf_j})$ illustrates the distance between two generated counterfactual instances, d is the number of input features, k is the number of counterfactual samples to be generated.

V. Open questions and discussions

The need of explaining and interpreting models becomes highly critical along with the growing popularity of deep learning and automated machine learning. Although, there are currently several studies in this field, several open problems still remain unresolved.

1) Counterfactual explanation in classification tasks. There are a plethora of constraints, especially features' causal relationship, should be taken into consideration when adopting counterfactual explanation. Take for example, the counterfactual explanation cannot recommend the users to change sensitive and discriminative features such as race and gender in order to be accepted by the system. Therefore, its reasonability and feasibility should be discovered and investigated more strictly.

2) Counterfactual explanation in recommendation system and time series data. Although recommendation system gains the immense popularity these days, there are not many studies working on counterfactual explanation for such system. How we can make an intervention into human actions to enable the system to change their recommended items still remains an open question. Meanwhile, regarding time series data, it is also interesting to discover that what the model would change its prediction if we change something in the past.

3) Causal reasoning in knowledge graph. Knowledge graph is recently utilized as an effective tool in several tasks such as recommendation system, knowledge extraction, classification, etc. Instead of embedding the knowledge graph as the latent features, Xian et al. [68] state that the true intelligent recommendation systems have to own the ability to recommend their items based on their causal reasoning.

4) Explanation understandable by non-experts. A number of recent methods frequently provide the explanations to

TABLE I
TOOLBOX FOR CAUSAL ANALYSIS

Library	Feature	Algorithms	License
DoWhy [55]	Individual treatment effect estimation	Prospensity score matching [56]	
Stratification [57]	Microsoft		
EconML [58]	Individual treatment effect estimation		
Interpreter of the causal model	Double Machine Learning [21]		
Orthogonal Random Forests [23], [24]			
Meta-Learners [22]			
Deep Instrumental Variables	Microsoft		
Causal ML [59]	Individual treatment effect estimation	Meta-learners	
Uplift modeling [60], [61]	Uber		
Causal discovery toolbox [62]	Causal relationship discovery	Graph Inference	
Pairwise inference	ElementAI		
CausalNex	Learn causal structures		
Estimate the effects of potential interventions using data.	Using Bayesian Networks for Causal Inference	QuantumBlack Labs	
TIGRAMITE	Causal discovery for time series datasets	PCMCI [63], Generally [64], CMIknn [65], Mediation class [66], [67]	GNU General Public

experts and researchers rather than the end-users. Therefore, another challenge is to generate explanation under the form such as rules, natural language, images, etc which can allow nonprofessional people to catch up with machine learning model behaviors.

VI. CONCLUSION

Interpretable machine learning is expected to become a mainstream topic in the foreseeable future. This paper provides the desiderata and brief overview of causal inference, followed by the causality based interpretable machine learning. We present two main causal approaches for interpretable machine learning including feature importance estimation, causal effects of model components, and counterfactual explanation. Finally, we has discussed several potentially unresolved problems in this field which open opportunities for researchers to work in.

In machine learning, the more data the better. However, in causal inference, the more data alone is not yet enough. Having more data only helps to get more precise estimates, but it cannot make sure these estimates are correct and unbiased. Machine learning methods enhance the development of causal inference, meanwhile, causal inference also helps machine learning methods. The simple pursuit of predictive accuracy is insufficient for modern machine learning research, and correctness and interpretability are also the targets of machine learning methods. Causal inference is starting to help to improve machine learning, such as recommender systems or reinforcement learning.

REFERENCES

[1] H. Lakkaraju, S. H. Bach, and J. Leskovec, "Interpretable decision sets: A joint framework for description and prediction," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, pp. 1675–1684.

[2] X. Zhang, S. Tan, P. Koch, Y. Lou, U. Chajewska, and R. Caruana, "Axiomatic interpretability for multiclass additive models," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2019, pp. 226–234.

[3] R. Caruana, Y. Lou, J. Gehrke, P. Koch, M. Sturm, and N. Elhadad, "Intelligible Models for HealthCare: Predicting Pneumonia Risk and Hospital 30-day Readmission," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '15*. Sydney, NSW, Australia: ACM Press, 2015, pp. 1721–1730.

[4] X. Zhang, S. Tan, P. Koch, Y. Lou, U. Chajewska, and R. Caruana, "Axiomatic Interpretability for Multiclass Additive Models," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining - KDD '19*. Anchorage, AK, USA: ACM Press, 2019, pp. 226–234.

[5] P. J. Darwen, "Bayesian model averaging for river flow prediction," *Applied Intelligence*, vol. 49, no. 1, pp. 103–111, 2019.

[6] B. Letham, C. Rudin, T. H. McCormick, D. Madigan *et al.*, "Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model," *The Annals of Applied Statistics*, vol. 9, no. 3, pp. 1350–1371, 2015.

[7] T. Wang, "Multi-Value Rule Sets," *arXiv:1710.05257 [cs], NIPS*, Oct. 2017, arXiv: 1710.05257.

[8] B. Letham, C. Rudin, T. H. McCormick, and D. Madigan, "Interpretable classifiers using rules and Bayesian analysis: Building a better stroke prediction model," *The Annals of Applied Statistics*, vol. 9, no. 3, pp. 1350–1371, Sep. 2015, arXiv: 1511.01644.

[9] S. O. Arik and T. Pfister, "TabNet: Attentive Interpretable Tabular Learning," *arXiv preprint arXiv:1908.07442*, 2019.

[10] J.-S. Jang and C.-T. Sun, "Functional equivalence between radial basis function networks and fuzzy inference systems," *IEEE transactions on Neural Networks*, vol. 4, no. 1, pp. 156–159, 1993.

[11] S. Guillaume, "Designing fuzzy inference systems from data: An interpretability-oriented review," *IEEE Transactions on fuzzy systems*, vol. 9, no. 3, pp. 426–443, 2001.

[12] J.-S. Wang and C. G. Lee, "Self-adaptive neuro-fuzzy inference systems for classification applications," *IEEE Transactions on Fuzzy Systems*, vol. 10, no. 6, pp. 790–802, 2002.

[13] M. T. Ribeiro, S. Singh, and C. Guestrin, "" Why should i trust you?" Explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 2016, pp. 1135–1144.

[14] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in neural information processing systems*, 2017, pp. 4765–4774.

[15] R. Guidotti, A. Monreale, S. Ruggieri, D. Pedreschi, F. Turini, and F. Giannotti, "Local rule-based explanations of black box decision systems," *arXiv preprint arXiv:1805.10820*, 2018.

[16] P. W. Koh and P. Liang, "Understanding black-box predictions via influence functions," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2017, pp. 1885–1894.

[17] P. Schwab, D. Miladinovic, and W. Karlen, "Granger-Causal Attentive Mixtures of Experts: Learning Important Features with Neural Networks," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, pp. 4846–4853, Jul. 2019.

[18] P. Schwab and W. Karlen, "CXPlain: Causal Explanations for Model Interpretation under Uncertainty," *arXiv:1910.12336 [cs, stat]*, Oct. 2019, arXiv: 1910.12336.

[19] J. Pearl, "Causal inference in statistics: An overview," *Statistics Surveys*, vol. 3, pp. 96–146, 01 2009.

[20] D. B. Rubin, "Estimating causal effects of treatments in randomized and nonrandomized studies." *Journal of educational Psychology*, vol. 66, no. 5, p. 688, 1974.

[21] V. Chernozhukov, D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins, "Double/debiased machine learning for treatment and causal parameters," *arXiv preprint arXiv:1608.00060*, 2016.

[22] S. R. Künzel, J. S. Sekhon, P. J. Bickel, and B. Yu, "Metalearners for estimating heterogeneous treatment effects using machine learning," *Proceedings of the National Academy of Sciences*, vol. 116, pp. 4156–4165, 2019.

[23] M. Oprescu, V. Syrgkanis, and Z. S. Wu, "Orthogonal random forest for causal inference," *arXiv preprint arXiv:1806.03467*, 2018.

[24] D. J. Foster and V. Syrgkanis, "Orthogonal statistical learning," *arXiv preprint arXiv:1901.09036*, 2019.

[25] J. Pearl and D. Mackenzie, *The Book of Why: The New Science of Cause and Effect*, 1st ed. USA: Basic Books, Inc., 2018.

[26] A. Chattopadhyay, P. Manupriya, A. Sarkar, and V. N. Balasubramanian, "Neural network attributions: A causal perspective," *arXiv preprint arXiv:1902.02302*, 2019.

[27] T. Narendra, A. Sankaran, D. Vijaykeerthy, and S. Mani, "Explaining deep learning models using causal inference," *arXiv preprint arXiv:1811.04376*, 2018.

[28] B. Kim, M. Wattenberg, J. Gilmer, C. J. Cai, J. Wexler, F. B. Viégas, and R. Sayres, "Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav)," in *ICML*, 2017.

[29] Y. Goyal, A. Feder, U. Shalit, and B. Kim, "Explaining Classifiers with Causal Concept Effect (CaCE)," *arXiv:1907.07165 [cs, stat]*, Feb. 2020, arXiv: 1907.07165.

[30] D. Bau, J.-Y. Zhu, H. Strobelt, B. Zhou, J. B. Tenenbaum, W. T. Freeman, and A. Torralba, "Gan dissection: Visualizing and understanding generative adversarial networks," *arXiv preprint arXiv:1811.10597*, 2018.

[31] M. Besserve, A. Mehrjou, R. Sun, and B. Schölkopf, "Counterfactuals uncover the modular structure of deep generative models," *arXiv preprint arXiv:1812.03253*, 2018.

[32] P. Madumal, T. Miller, L. Sonenberg, and F. Vetere, "Explainable reinforcement learning through a causal lens," *arXiv preprint arXiv:1905.10958*, 2019.

[33] G. C. Cawley, "Causal & non-causal feature selection for ridge regression," in *Proceedings of the Workshop on the Causation and Prediction Challenge at WCCI 2008*, ser. Proceedings of Machine Learning Research, I. Guyon, C. Aliferis, G. Cooper, A. Elisseeff, J.-P. Pellet, P. Spirtes, and A. Statnikov, Eds., vol. 3. Hong Kong: PMLR, 03–04 Jun 2008, pp. 107–128.

[34] C. F. Aliferis, I. Tsamardinos, and A. Statnikov, "Hiton: a novel markov blanket algorithm for optimal variable selection," in *AMIA annual symposium proceedings*, vol. 2003. American Medical Informatics Association, 2003, p. 21.

[35] J. Peters, P. Bühlmann, and N. Meinshausen, "Causal inference using invariant prediction: identification and confidence intervals," *arXiv e-prints*, p. arXiv:1501.01332, Jan. 2015.

[36] R. M. Grath, L. Costabello, C. L. Van, P. Sweeney, F. Kamiab, Z. Shen, and F. Lecue, "Interpretable Credit Application Predictions With Counterfactual Explanations," *arXiv:1811.05245 [cs]*, Nov. 2018, arXiv: 1811.05245.

[37] S. Wachter, B. Mittelstadt, and C. Russell, "Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR," *arXiv:1711.00399 [cs]*, Mar. 2018, arXiv: 1711.00399.

[38] A. Dhurandhar, P.-Y. Chen, R. Luss, C.-C. Tu, P. Ting, K. Shanmugam, and P. Das, "Explanations based on the missing: Towards contrastive explanations with pertinent negatives," in *Advances in neural information processing systems*, 2018, pp. 592–603.

[39] A. Van Looveren and J. Klaise, "Interpretable Counterfactual Explanations Guided by Prototypes," *arXiv:1907.02584 [cs, stat]*, Feb. 2020, arXiv: 1907.02584.

[40] R. K. Mothilal, A. Sharma, and C. Tan, "Explaining machine learning classifiers through diverse counterfactual explanations," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. Barcelona Spain: ACM, Jan. 2020, pp. 607–617.

[41] B. Ustun, A. Spangher, and Y. Liu, "Actionable recourse in linear classification," in *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 2019, pp. 10–19.

[42] C. Russell, "Efficient Search for Diverse Coherent Explanations," in *Proceedings of the Conference on Fairness, Accountability, and Transparency - FAT* '19*. Atlanta, GA, USA: ACM Press, 2019, pp. 20–28.

[43] A. Artelt and B. Hammer, "Convex density constraints for computing plausible counterfactual explanations," *arXiv preprint arXiv:2002.04862*, 2020.

[44] S. Sharma, J. Henderson, and J. Ghosh, "Certifai: Counterfactual explanations for robustness, transparency, interpretability, and fairness of artificial intelligence models," *arXiv preprint arXiv:1905.07857*, 2019.

[45] Zhou Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[46] R. Poyiadzi, K. Sokol, R. Santos-Rodriguez, T. De Bie, and P. Flach, "Face: Feasible and actionable counterfactual explanations," in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 2020, pp. 344–350.

[47] A. Goldstein, A. Kapelner, J. Bleich, and E. Pitkin, "Peeking inside the black box: Visualizing statistical learning with plots of individual conditional expectation," *Journal of Computational and Graphical Statistics*, vol. 24, no. 1, pp. 44–65, 2015.

[48] Q. Zhao and T. Hastie, "Causal interpretations of black-box models," *Journal of Business & Economic Statistics*, vol. 0, no. 0, pp. 1–10, 2019. [Online]. Available: https://doi.org/10.1080/07350015.2019.1624293

[49] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," *arXiv preprint arXiv:1702.08608*, 2017.

[50] J. J. Williams, J. Kim, A. Rafferty, S. Maldonado, K. Z. Gajos, W. S. Lasecki, and N. Heffernan, "Axis: Generating explanations at scale with learnersourcing and machine learning," in *Proceedings of the Third (2016) ACM Conference on Learning @ Scale*, ser. L@S '16. New York, NY, USA: Association for Computing Machinery, 2016, p. 379–388. [Online]. Available: https://doi.org/10.1145/2876034.2876042

[51] R. R. Hoffman, S. T. Mueller, G. Klein, and J. Litman, "Metrics for explainable ai: Challenges and prospects," *arXiv preprint arXiv:1812.04608*, 2018.

[52] T. Kadir and M. Brady, "Saliency, scale and image description," *Int. J. Comput. Vision*, vol. 45, no. 2, p. 83–105, Nov. 2001. [Online]. Available: https://doi.org/10.1023/A:1012460413855

[53] M. Sundararajan, A. Taly, and Q. Yan, "Axiomatic attribution for deep networks," in *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 2017, pp. 3319–3328.

[54] M. Harradon, J. Druce, and B. Ruttenberg, "Causal learning and explanation of deep neural networks via autoencoded activations," *arXiv preprint arXiv:1802.00541*, 2018.

[55] "DoWhy: A Python package for causal inference," https://github.com/microsoft/dowhy.

[56] R. H. Dehejia and S. Wahba, "Propensity score-matching methods for nonexperimental causal studies," *Review of Economics and statistics*, vol. 84, no. 1, pp. 151–161, 2002.

[57] C. E. Frangakis and D. B. Rubin, "Principal stratification in causal inference," *Biometrics*, vol. 58, no. 1, pp. 21–29, 2002.

[58] M. Research, "EconML: A Python Package for ML-Based Heterogeneous Treatment Effects Estimation," https://github.com/microsoft/EconML, 2019, version 0.x.

[59] H. Chen, T. Harinen, J.-Y. Lee, M. Yung, and Z. Zhao, "Causalml: Python package for causal machine learning," 2020.

[60] Y. Zhao, X. Fang, and D. Simchi-Levi, "Uplift Modeling with Multiple Treatments and General Response Types," *arXiv e-prints*, p. arXiv:1705.08492, May 2017.

[61] N. J. Radcliffe and P. D. Surry, "Real-world uplift modelling with significance-based uplift trees," *White Paper TR-2011-1, Stochastic Solutions*, pp. 1–33, 2011.

[62] D. Kalainathan and O. Goudet, "Causal Discovery Toolbox: Uncover causal relationships in Python," *arXiv e-prints*, p. arXiv:1903.02278, Mar. 2019.

[63] J. Runge, P. Nowack, M. Kretschmer, S. Flaxman, and D. Sejdinovic, "Detecting and quantifying causal associations in large nonlinear time series datasets," *Science Advances*, vol. 5, no. 11, 2019. [Online]. Available: https://advances.sciencemag.org/content/5/11/eaau4996

[64] J. Runge, "Causal network reconstruction from time series: From theoretical assumptions to practical estimation," *Chaos: An Interdisciplinary Journal of Nonlinear Science*, vol. 28, no. 7, p. 075310, 2018. [Online]. Available: https://doi.org/10.1063/1.5025050

[65] ——, "Conditional independence testing based on a nearest-neighbor estimator of conditional mutual information," *arXiv preprint arXiv:1709.01447*, 2017.

[66] J. Runge, V. Petoukhov, J. F. Donges, J. Hlinka, N. Jajcay, M. Vejmelka, D. Hartman, N. Marwan, M. Paluš, and J. Kurths, "Identifying causal gateways and mediators in complex spatio-temporal systems," *Nature communications*, vol. 6, no. 1, pp. 1–10, 2015.

[67] J. Runge, "Quantifying information transfer and mediation along causal pathways in complex systems," *Physical Review E*, vol. 92, no. 6, p. 062829, 2015.

[68] Y. Xian, Z. Fu, S. Muthukrishnan, G. De Melo, and Y. Zhang, "Reinforcement knowledge graph reasoning for explainable recommendation," in *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2019, pp. 285–294.

Koreisha: Web Platform to Measure Healthcare System Coverage in Chile

Rodrigo Pérez*, Víctor Hernández M†, Fernando Henríquez‡, Paulina Arriagada§, Pedro Zitko ¶, Andrea Slachevsky ‖ and Juan D. Velásquez **

Web Intelligence Centre. Department of Industrial Engineering, University of Chile.* † **

Neuropsychology and Clinical Neuroscience Laboratory (LANNEC), Physiopathology Department - ICBM, Neurocience and East Neuroscience Departments, Faculty of Medicine, University of Chile.‡ ‖

Neurology Unit, Regional Hospital Coyhaique, Aysén Health Service.§

Health Service & Population Research, IoPNN. King's College London.¶

Email: rodrperez@ing.uchile.cl*, victor.hernandez@wic.uchile.cl†, fehech@gmail.com‡, arriagadapau@saludaysen.cl§, pedrozitko@gmail.com¶, andrea.slachevsky@uchile.cl‖, jvelasqu@dii.uchile.cl**

ABSTRACT

To assess the performance of the healthcare system, decision makers require tools to quantify coverage, ideally in real time. We implement a methodology to measure system coverage by software, which is applied as a prototype in the city of Coyhaique, southern Chile. The current implementation use a survey that gathers sociodemographic data about the patients, and at the same time, healthcare system performance information which is related to an adaptation of Tanahashi's coverage model. Results are deployed to decision makers by means of a web dashboard that delivers system coverage measurements obtained from the survey and an analysis of the reasons that justify the failure in treatment success. The platform includes a Predictive Module which estimates system coverage for a treatment, given the patient information.

I. INTRODUCTION

The healthcare system involves multiple components and stakeholders, which creates a complex structure and interactions that difficult the measurement of performance metrics (Smith *et al.*, 2010; Zhang *et al.*, 2015). In order to tackle this issue, we present Koreisha, which is a web platform whose objective is to measure health system coverage for a set of healthcare needs (normative needs). In each case, we analyze the performance of the treatments related to the normative need, describing the reason of failure in treatment success. So far, this structure allows us to connect patients, health system personnel and decision makers.

One of the instruments that Chile use for decision making on public health management is the National Health Survey (Margozzini & Passii, 2018). The latest comprised the period 2016-2017 and included 6233 people. Koreisha aims to improve the process of data gathering and processing in order to increase the rate of health system information. The methodology implements a digital survey, which results could be processed in real time, and be used to train a predictive model for classifying the expected system coverage for a patient. The survey measures several aspects of the patients: sociodemographic and labour history, medical variables to assess the normative need and the reasons that explains why a patient has failed in a treatment. This information allows us to create a profile that could be used, for example, to analyze equitative access to health system attention (Frenz & Vega, 2010).

The survey is implemented with the help of the Center of Geroscience, Mental Health and Metabolism (Gero)[1]. They developed a platform with the ability to implement complex surveys through a graphical user interface or even by coding it with the Scheme language. The agile implementation of the instrument made possible to apply the methodology as a prototype study in the city of Coyhaique (Aysén region, southern Chile). Together with the Health Service Management of the Aysén Region, we developed instructions to apply the Koreisha survey through the Gero platform, which allowed the health service personnel to interview the people.

II. METHODS

A. Interviews

Koreisha was applied in the city of Coyhaique with focus on senior people able to answer the survey. Besides the sociodemografic data, the survey comprises 15 health modules, each related to a normative need. This results in a long instrument whose completion took around 10 hours distributed in two sessions. The local Health Service suggested 440 potential candidates for the interview, from which 137 were actually interviewed. Finally, a group of 81 people answered the full survey.

B. Data processing

Each normative need has a set of rules that determines if the interviewee has the need or not. For example, the cognitive module comprises the results of the Montreal Cognitive Assesment test, Pfeffer Functional Assessment Questionnaire and AD8 Informant Questionnire. If someone turns out positive in all three

[1] http://www.gerochile.cl/es/

tests, then it is said that the person has the normative need.

Treatment coverage is measured by relating its value to the reason that an interviewee asigned to the question of why the treatment was dropped. That said, if the person answered that the treatment wasn't received because there were no professionals to treat the disease, then healthcare system failed in availability. Similarly, the stages of accessibility, acceptability and effectivity are related to other reasons available in the survey. The system provided proper care if the answer is not related to any of the previously mentioned stages.

Coverage results are used to train a supervised model, to estimate the coverage output for a treatment given the patient data. The model will give an idea to health care personnel of the stage in which the system will fail. So, preventive measures could be taken to assist the patient and increase coverage.

C. Information system

The software implements an Extract, Transform and Loading (ETL) process. We read the Gero database related to the project, which has the answers to the survey in JSON format and the questions in Scheme language representation. We parse the questions and match them with the answers which is saved as a MongoDB document. Then we compute (transform) the answers of a survey to normative need indicators and health system coverage measurements. This data is loaded to be available through a Django based Application Programming Interface (API) and to be processed by the Predictive Module. The latter results are also available through the API. A diagram of the structure of the Koreisha information system is presented in Fig. (1).

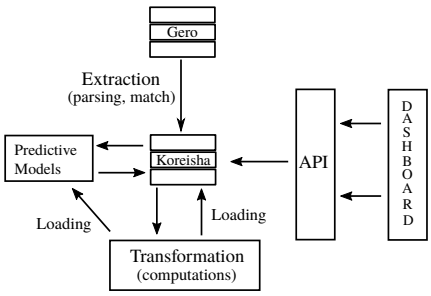

Fig. 1. Information system diagram. Extraction, transformation and loading process describe the interaction between the Gero and Koreisha databases.

III. Results

Fig.(2) shows the main view of the web dashboard. It summarizes the results grouped by health module and provides tools for data and project related documents downloading. Also, the dashboard is designed to display the Predictive Module results, which full implementation to the end user is currently under development.

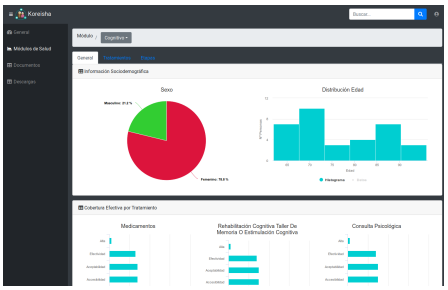

Fig. 2. General view of the Koreisha dashboard.

The information system displays the interviewees age histogram which range from 60 to 96 years old, and 75% of them were females, (Fig. 3). From the medical exams, it was measured that 9 of them had the cognitive normative need.

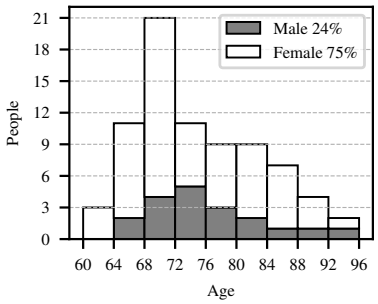

Fig. 3. Age histogram of the interviewed people differentiated by sex. Females interviewees were predominant reaching the 75% of the sample.

Fig. (4a) shows the survey coverage results for the cognitive rehabilitation treatment in the cognitive health module. Fig (4b) show the reasons for treatment abandonment. Both charts are a representation of the information displayed by Koreisha for every treatment in a health module.

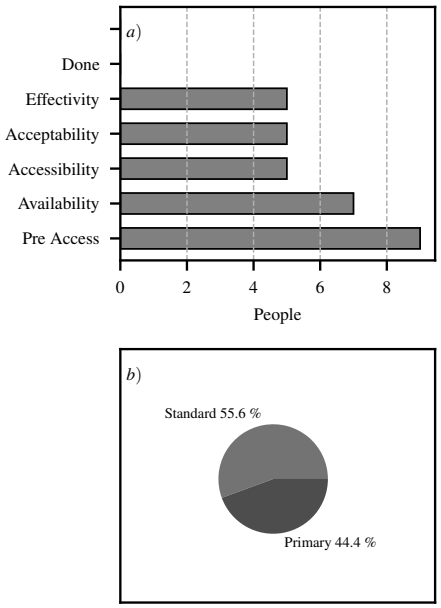

Fig. 4. Coverage results for the cognitive rehabilitation treatment in the cognitive health module: a) Coverage diagram based on Tanahashi's proposal b) Distribution of the treatment abandonment reasons.

IV. Discussion

The Koreisha Platform successfully implements a methodology that transfers patient data to the decision makers and health care personnel. We acknowledge the fact that our database is still small and that a protocol for continuos data gathering should be implemented in order to advance towards real time. Eventually, the system should develop a structure like the proposed in Zhang *et al.* (2015), allowing real time interaction between multiple stakeholders and a big data environment. So far Koreisha measures health system coverage based on Tanahashi (1978). Nevertheless, we are aware that effective coverage is a more robust indicator of health system performance (Shengelia *et al.*, 2005; Ng *et al.*, 2014). Future work will be focused on improving the survey in order to make it shorter and easier to apply. Also, a relation between resources allocation and coverage indicators will be performed. This information will be useful to perform simulations whose input is based on sociodemographic parameters for a given population group and with those results, assess if the system is ready to satisfy population needs.

REFERENCES

Frenz, P., & Vega, J. 2010. Universal health coverage with equity: what we know, don't know and need to know. *First Global Symposium on Health Systems Research*, **1**.

Margozzini, P., & Passii, A. 2018. Encuesta Nacional de Salud, ENS 2016-2017: Un aporte a la planificación sanitaria y políticas públicas de Chile. *Ars Medica Revista de Ciencias Médicas*, **43**(1).

Ng, M., Fullman, N., Dieleman, J., Flaxman, A., Murray, C., & Lim, S. 2014. Effective Coverage: A Metric for Monitoring Universal Health Coverage. *PLOS Medicine*, **11**(9).

Shengelia, B., Tandon, A., Adams, O., & Murray, C. 2005. Access, utilization, quality, and effective coverage: An integrated conceptual framework and measurement strategy. *Social Science & Medicine*, **61**, 97–109.

Smith, P., Mossialos, E., Papanicolas, I., & Leatherman, S. 2010. *Performance measurement for health system improvement*. Cambridge University Press.

Tanahashi, T. 1978. Health service coverage and its evaluation. *Bulletin of the World Health Organization*, **56**(2), 295–303.

Zhang, Y., Qiu, M., Tsai, C., Mehedi, M., & Alamri, A. 2015. Health-CPS: Healthcare Cyber-Physical System Assisted by Cloud and Big Data. *IEEE Systems Journal*, 295–303.

Prediction of Public Health System Coverage for Senior Adults in Chile using Machine Learning Tools

FOUNDATIONS FOR A MONITORING PLATFORM FOR PUBLIC HEALTH SYSTEM COVERAGE IN CHILE

Víctor Hernández M*, Rodrigo Pérez†, Fernando Henríquez‡, Paulina Arriagada§, Pedro Zitko ¶, Andrea Slachevsky ‖ and Juan D. Velásquez **

Web Intelligence Centre. Department of Industrial Engineering, University of Chile.* † **

Neuropsychology and Clinical Neuroscience Laboratory (LANNEC), Physiopathology Department - ICBM, Neurocience and East Neuroscience Departments, Faculty of Medicine, University of Chile.‡ ‖

Neurology Unit, Regional Hospital Coyhaique, Aysén Health Service.§

Health Service & Population Research, IoPNN. King's College London.¶

Email: victor.hernandez@wic.uchile.cl*, rodrperez@ing.uchile.cl†, fehech@gmail.com‡, arriagadapau@saludaysen.cl§, pedrozitko@gmail.com¶, andrea.slachevsky@uchile.cl‖, jvelasqu@dii.uchile.cl**

ABSTRACT

Chile is an aging country, so monitoring the health coverage of the public system towards its elderly is of interest. We propose a novel approach to eventually predict the level of health coverage that an elderly person could receive, based on a profile based on sociodemographic and health variables, through the use of Machine Learning. Data is collected through a specialized survey and tests are carried out through different supervised classification algorithms. Our results suggest that this approach could complement the information that will be available to the decision makers of the public health system, nevertheless, the collection of data required to be improved and our approach need to be evaluated in future surveys.

I. INTRODUCTION

Chile is a country that has aged rapidly over the last decade, from having 15% of the total population over 60 years in 2009, to 19.3% in 2017. In addition, that same year, 84.9% of the population over 60 years of age was part of the public health system (FONASA). On the other hand, there are health problems of high prevalence for these age segments, such as dementia disorders or neurocognitive disorders, which are an influential factor in the inequality of the population, representing high monetary and social costs in the family of those affected (Ministerio de Desarrollo Social, 2017). This background shows that it is important to have efficient methods to monitor the health coverage of senior adults in Chile, to support decision-making and an efficient allocation of limited resources (Hojman *et al.*, 2015).

Currently, the data that exists in Chile regarding these factors is static and scarce. The closest to the collection of data of this nature is the National Health Survey (Margozzini & Passii, 2018), carried out periodically and on a limited sample of people. The last edition was carried out in the period 2016-2017 covering 6233 people. It is worth mentioning that the processing of its results is a slow process, carried out by expert personnel, which leads to a report. In other words, static results are obtained for each edition of the survey.

Here, we provide an approach to collect information in an expeditious manner, which allows us to monitor the public health system coverage for older adults in Chile. The main idea is to make predictions of the potential health coverage for a elderly, using Data Mining and Machine Learning tools, to generate information that supports the decision maker. This is a work developed jointly between the Web Intelligence Centre of the Faculty of Engineering and the Faculty of Medicine of the University of Chile.

II. METHODOLOGY

A. Data collection

In order to gather the patients information, we designed a survey composed of two sections for each health problem: a module that evaluates if the person has a health care need (normative need) (Bradshaw, 1972), and a module that assesses the health system coverage. The latter, considers the primary care level (care centers), the secondary care level (specialized clinics) and the multiple treatments that could be indicated for a given health care need. Along with this, there are attached instruments that assess the socioeconomic status of the respondents, their level of accessibility to healthcare centers, their occupational career, their level of literacy in health, among others.

The survey includes 15 health modules, associated with problems prevalent in the elderly population. Some of them are: cognitive, affective, cardiometabolic and musculoskeletal. To obtain the normative need, a standardized instrument is used for each corresponding health module. To model the health system coverage, an adaptation of the Tanahashi coverage model (Tanahashi, 1978) has been used. The selected population is located in the city of Coyhaique (Aysén region, southern Chile).

The survey was carried out with the help of a digital platform developed by the Center of Geroscience, Mental Health and Metabolism (Gero)[1] and trained interviewers. 2754 addresses were visited in a period of two months, from which 440 subjects were found eligible for the study, and 137 agreed to participate. 81 surveys were answered in their entirety, beign the length of the survey and the inability of some people to respond due to their health status, between the main causes of uncompleted surveys.

B. Data preprocessing

Once the 81 completed surveys were identified, a preprocessing is performed based on the cleaning of null values, conversion of numerical values with respect to formats and units of measurement, and semi-automatic processing of unstructured text is included in some answers (e.g., how long does it take to get to a health care center from your home?).

The majority of the survey consists of categorical variables, so in addition, these questions were preprocessed as binary variables (or dummy variables) to facilitate their interpretation by the classifier.

C. Data transformation

This stage consists of two steps. First, for each health module the calculation of the normative need is made, to establish whether the person has the problem or not, which is represented with a binary variable. For this analysis, we evaluate the rules for each module, which were defined based on the standards associated with each instrument. For example, cognitive health module uses Montreal Cognitive Assessment (Nasreddine & Phillips, 2005), Pfeffer Functional Activities Questionnaire (Fuentes-García, 2014) and AD8 Informant Questionnaire (Galvin *et al.*, 2005; Muñoz *et al.*, 2010) instruments together to determine the normative need.

The second step computes the health system coverage for each of the treatments associated with the module and seeks to capture whether at each level of care the patient was diagnosed and received the right treatment. It also captures the patient's response to the treatment and the reasons of treatment abandonment. The transformation interprets the response flow and links it with the level of health system coverage, which is saved in a vector of categorical variables. It is worth mentioning that for most health modules, if the person does not have the health problem, the person immediately falls into the first level of health coverage, implying that the person did not require the service. In those health modules where the normative need may be affected by an ongoing treatment, the flow of responses is analyzed anyway.

[1] http://www.gerochile.cl/es/

D. Classification Process

Up to this point there are three types of data: the variables associated with all the answers given by the respondents, the computed normative needs and the level of health system coverage for each treatment, for each health module of the survey.

Because we aim to predict the health system coverage based on a patient's profile, a supervised classification process is applied, where the dependent variable corresponds to the treatment coverage levels found, and the independent variables are the computed normative needs, together with the variables that allowed us to determine that value. In addition, the sociodemographic information, accessibility and occupational trajectory are included within the independent variables to further complement the profile of each person.

Tests were done with Multinomial Naive Bayes, K-Nearest Neighbors and Random Forests algorithms for supervised classification. These were implemented from the library scikit-learn (Pedregosa *et al.*, 2011). Each classification includes a 3-fold cross validation process for the evaluation.

III. Results

Below are the best 5 average results for each classification process.

TABLE I: Multinomial Naive Bayes results

Health module	Precision	Recall	F-measure
Respiratory	0,7802	0,3502	0,4562
Cognitive	0,8099	0,3270	0,4362
Sleep disorder	0,5727	0,3304	0,4064
Vision	0,5168	0,2484	0,2926
Affective	0,4091	0,2338	0,2800

TABLE II: K-Nearest Neighbors results

Health module	Precision	Recall	F-measure
Respiratory	0,8142	0,9020	0,8557
Cognitive	0,7997	0,8924	0,8426
Sleep disorder	0,7027	0,7348	0,6590
Diabetes	0,4579	0,4658	0,4503
Overweight	0,4286	0,4878	0,4458

Random Forests results:

TABLE III: Random Forests results

Health module	Precision	Recall	F-measure
Cognitive	0,8368	0,9042	0,8659
Respiratory	0,8142	0,9020	0,8557
Sleep disorder	0,5388	0,7052	0,6094
Vision	0,4731	0,6329	0,5167
Overweight	0,4865	0,6321	0,5134

IV. Discussion

The results show that the best performance is obtained with the Random Forests algorithm. However, the process should continue to improve given that performance is only high for some health modules. The survey turned out to be an impractical experience for the respondents, so an application outside the experimental environment would be unfeasible. The next step is to identify, through the results obtained, which are the health modules and annexed instruments that add more information, to preserve only those relevant questions in a future instance of data collection. This will allow a greater representativeness of the population to be achieved, through a more complete data set and a larger scale. However, this work reflects a novel contribution to complement the information available to decision makers of the public health system in Chile and that laid a basis for directing the availability of this information towards a dynamic approach, fed by various data sources.

References

Bradshaw, Jonathan. 1972. Taxonomy of social need.

Fuentes-García, Alejandra. 2014. *Pfeffer Functional Activities Questionnaire*. Dordrecht: Springer Netherlands.

Galvin, JE, Roe, CM, & Powlishta. 2005. The AD8: a brief informant interview to detect dementia. *Neurology*, **65**(4), 559–564.

Hojman, D, Duarte, F, Ruiz-Tagle, J, Nuñez-Huasaf, J, Budinich, M, & Slachevsky, A. 2015. The cost of dementia: The case of chile. Results of the cuideme study. *Journal of the Neurological Sciences*, **357**, e11.

Margozzini, P., & Passii, A. 2018. Encuesta Nacional de Salud, ENS 2016-2017: Un aporte a la planificación sanitaria y políticas públicas de Chile. *Ars Medica Revista de Ciencias Médicas*.

Ministerio de Desarrollo Social, Chile. 2017. *Encuesta CASEN 2017 Adultos Mayores - Sintesis de Resultados*.

Muñoz, Carlos, Nunez, Javier, Flores, Patricia, Behrens, P MI, & Slachevsky, Andrea. 2010. Usefulness of a brief informant interview to detect dementia, translated into Spanish (AD8-Ch). *Revista medica de Chile*, **138**(8), 1063.

Nasreddine, Ziad S, & Phillips. 2005. The Montreal Cognitive Assessment, MoCA: a brief screening tool for mild cognitive impairment. *Journal of the American Geriatrics Society*.

Pedregosa, Fabian, Varoquaux, Gaël, Gramfort, Alexandre, & Michel. 2011. Scikit-learn: Machine learning in Python. *Journal of machine learning research*.

Tanahashi, T. 1978. Health service coverage and its evaluation. *Bulletin of the World Health Organization*, **56**(2), 295–303.

Selected Ph.D. Thesis Abstracts

This Ph.D thesis abstracts section presents theses defended in 2019 and 2020. These submissions cover a range of research topics and themes under intelligent informatics, such as crowdsourced data, transfer learning, dynamic networks, content-based language learning, intelligent transportation systems, k-means clustering, deep learning and data mining.

COMPLEX TASK ALLOCATION FOR CROWDSOURCING IN SOCIAL NETWORK CONTEXT

Jiuchuan Jiang
jcjiang@nufe.edu.cn
Nanjing University of Finance and Economics, China

ALLOCATION of complex tasks has attracted significant attention in crowdsourcing area recently, which can be categorized into decomposition and monolithic allocations. Decomposition allocation means that each complex task will first be decomposed into a flow of simple subtasks and then the subtasks will be allocated to individual workers; monolithic allocation means that each complex task will be allocated as a whole, which includes individual-oriented and team formation-based approaches. However, those existing approaches have some problems for real crowdsourcing markets. On the other hand, workers are often connected through social networks, which can significantly facilitate crowdsourcing of complex tasks. Therefore, this thesis investigates crowdsourcing in social network context and presents models to address the typical problems in complex task allocation. The main contributions of this thesis are shown as follows.

First, traditional decomposition allocation for complex tasks has the following typical problems: 1) decomposing complex tasks into a set of subtasks requires the decomposition capability of the requesters; and 2) reliability may not be ensured when there are many malicious workers in the crowd. To this end, this thesis investigates the context- aware reliable crowdsourcing in social networks. In our approach, when a requester wishes to outsource a task, a worker candidate's self- situation and contextual-situation in the social network are considered. Complex tasks can be performed through autonomous coordination between the assigned worker and his contextual workers in the social network; thus, requesters can be exempt from decomposing complex tasks into subtasks. Moreover, the reliability of a worker is determined not only by the reputation of the worker himself but also by the reputations of the contextual workers, which can effectively address the unreliability of transient or malicious workers.

Second, traditional individual-oriented monolithic allocation for complex tasks often allocate tasks independently, which has the following typical problems: 1) the execution of one task seldom utilize the results of other tasks and the requester must pay in full for the task; and 2) many workers only undertake a very small number of tasks contemporaneously, thus the workers' skills and time may not be fully utilized. To this end, this thesis investigates the batch allocation for tasks with overlapping skill requirements. Then, two approaches are designed: layered batch allocation and core-based batch allocation. The former approach utilizes the hierarchy pattern to form all possible batches, which can achieve better performance but may require higher computational cost; the latter approach selects core tasks to form batches, which can achieve suboptimal performance with significantly reducing computational cost. If the assigned worker cannot complete a batch of tasks alone, he/she will cooperate with the contextual workers in the social network. Through the batch allocation, requesters' real payment can be discounted because the real execution cost of tasks can be reduced, and each worker's real earnings may increase because he/she can undertake more tasks contemporaneously.

Third, traditional team formation-based monolithic allocation for complex tasks has the following typical problems: 1) each team is created for only one task, which may be costly and cannot accommodate crowdsourcing markets with a large number of tasks; and 2) most existing studies form teams in a centralized manner, which may place a heavy burden on requesters. To this end, this thesis investigates the distributed team formation for a batch of tasks, in which similar tasks can be addressed in a batch to reduce computational costs and workers can self-organize through their social networks to form teams. In the presented team formation model, the requester only needs to select the first initiator worker and other team members are selected in a distributed manner, which avoids imposing all team formation computation loads on the requester. Then, two heuristic approaches are designed: one is to form a fixed team for all tasks in the batch, which has lower computational complexity; the other is to form a basic team that can be dynamically adjusted for each task in the batch, which performs better in reducing the total payments by requesters.

Fourth, current workers are often naturally organized into groups through social networks. To address such common problem, this thesis investigates a new group-oriented crowdsourcing paradigm in which the task allocation targets are naturally existing worker groups but not individual workers or artificially-formed teams as before. An assigned group often needs to coordinate with other groups in the social network contexts for performing a complex task since such natural group might not possess all of the required skills to complete the task. Therefore, a concept of contextual crowdsourcing value is presented to measure a group's capacity to complete a task by coordinating with its contextual groups, which determines the probability that the group is assigned the task;

then the task allocation algorithms, including the allocations of groups and the workers actually participating in executing the task, are designed.

In summary, this thesis develops new models to cover the shortages of previous complex task allocation works and designs efficient algorithms to solve the corresponding problems by considering the social network contexts. Experimental results conducted on real- world datasets collected from some representative crowdsourcing platforms show that the presented approaches outperform existing benchmark approaches in previous studies. The future work mainly includes designing a mechanism to identify malicious workers in the task allocation process and considering the factor of non-cooperation between workers for the complex task allocation.

Link Prediction via Transfer Learning across Multiple Domains

Paolo Mignone
paolo.mignone@uniba.it
University of Bari, Italy

THE data currently generated by modern society far exceeds our ability to analyze them manually without the help of automated analytical techniques. In this regard, the research field of knowledge discovery from data (KDD) aims at developing methods and techniques for the automatic analysis and the discovery of knowledge from data. The most important phase of the KDD process, called data mining, consists in the identification of relationships and delivers results that can be exploited both by other systems or by human experts.

In recent years, the data mining research community has spent much effort on data represented as networks, since they allow a natural description of many social, biological and information systems. In particular, individuals/objects/items are represented as nodes in the network and relationships or interactions among them are represented as links.

Among the possible mining tasks on network data, many works focused on the identification of previously unknown links among nodes. This task, which is called *link prediction*, aims to identify previously unknown links among nodes, on the basis of other known links and on the basis of the attributes/features of nodes. The link prediction task could be performed as a binary classification task when positive links (i.e. the existing links) and negative links (i.e. the non-existing links) are present. However, many real contexts are described according to the existing links among the nodes of the networks by lacking the negative links. This problem could be overcome by resorting to *transfer learning* methods.

In a classical machine learning setting, the learning is performed by considering a single target domain in order to learn a single task. Differently, in the transfer learning setting at least two domains are necessary. The goal is to exploit the knowledge of the source domain to improve the task in the target domain. Therefore, in this work three different and novel transfer learning methods are proposed which are able to fruitfully exploit the link knowledge of external and related networks in order to overcome the gap of the target network links' knowledge. Specifically, the proposed methods are able to originally combine the predictive model learned on the source network with the predictive model learned on the target network by constructing a hybrid predictive model that is more accurate to identify unknown links of the considered networks.

Experiments were performed by considering the mouse GRN as the source network and the human GRN as the target network. Quantitative and qualitative results showed that the proposed transfer learning methods are able to fruitfully exploit the knowledge acquired from the source network by outperforming state-of-the-art transfer learning methods. As a second contribution, the cross-organism importance of the organs for the network reconstruction is investigated by emphasizing that the skin and heart of the mouse are crucial to identify unknown gene regulation activities. Moreover, the proposed methods suggested gene regulations, which were not detected by other tools, that are identified as biologically relevant by experts in the biological domain.

As future work, a distributed version of the proposed methods will be developed to handle the whole set of all the possible GRNs' connections. Moreover, the methods will be enabled to work also with multiple source domains by catching the cross-domain homologies.

Pattern-based Change Detection in Large Dynamic Networks

Angelo Impedovo
angelo.impedovo@uniba.it
Universita' degli Studi di Bari Aldo Moro, Italy

DYNAMIC networks are made of interconnected nodes of different types whose topology continuously evolves. In such an evolving scenario, traditional mining algorithms designed for static networks become inadequate. This is due to the concept drift problem, which manifests every time the observed data begins to diverge from the ordinary situation, thus quickly degrading the performance of previously learned models. The problem is tackled by change detection algorithms that track a quality measure of learned models to i) quickly react to drifts in data and ii) undertake the necessary actions to adjust previously learned models. However, traditional algorithms are not designed for dynamic networks, and their adaptation poses different challenges. Firstly, it is not clear what feature-set of the network to consider while performing change detection. Secondly, the notion of change is ascribed to quantitative measures, while no clear definition exists for dynamic networks since they are subject to different types of changes. Thirdly, traditional algorithms only quantify changes without characterizing them. To address these issues, this thesis proposes pattern-based change detection algorithms (PBCDs), a novel class of symbolic, unsupervised, and non-parametric change detection algorithms for simultaneously detecting and characterizing changes exhibited by large dynamic networks over the time. In particular, PBCDs search for changes on a symbolic model of the network, typically frequent patterns denoting the stable features of the network over the time discovered by pattern mining algorithms, rather than on raw data. The symbolic model is learned in an

unsupervised fashion without making any prior assumption on the data distribution.

The thesis collects different contributions concerning PBCDs. Specifically, the KARMA algorithm (networK streAm macRoscopic Microscopic chAnge) is proposed as the first PBCD algorithm adopting automatically sized time windows to seek changes as variations in the sets of frequent connected subgraphs over time. Then, KARMA is generalized into a general PBCD architecture in which to accommodate the definition of new PBCDs. Such architecture alternates the execution of a pat- tern mining step in which the symbolic model of the network is learned from incoming snapshots, a change detection step in which variations in the symbolic model are measured, and a change characterization step which characterizes the detected changes.

The architecture, implemented in an open-source framework for disseminating existing PBCDs and promoting the development of new ones, is leveraged to empirically evaluate a wide collection of PBCDs on real-world and synthetic networks. Results show that PBCDs are more accurate and efficient change detection approaches, and offer more accurate and more complete change characterizations than state-of-the-art methodologies. Moreover, the effectiveness of PBCDs in real-world applications is shown by two applications in communication network analysis and process mining, respectively. Future directions of research may concern the development of more elegant PBCDs with alternative pattern mining, change detection, and characterization steps.

AUTOMATING VOCABULARY TESTS AND ENRICHING ONLINE COURSES FOR LANGUAGE LEARNERS

Jemma König
jemma.konig@waikato.ac.nz
The University of Waikato, New Zealand

THE past decade has seen massive growth in online academic courses, most of which are offered in the English language. However, although more people speak English as their second language than as their first, online course providers do not offer language assistance. This thesis aims to remedy that by integrating domain-specific language resources into online content, taking advantage of Massive Open Online Courses for "content based language learning". Content-based language learning is the dual concept of learning a subject through a foreign language, and learning the foreign language by studying the subject. This type of content-based approach fits well with the idea of integrating language resources into existing online courses. However, doing so raises several challenges, three of which are addressed in this thesis.

First, courses teach subjects in particular domains, but supporting domain-specific language requires knowledge of specialized vocabulary. This thesis develops an automated approach to generating domain-specific corpora and wordlists, extracting domain-specific vocabulary in a way that can be applied to any online course. This has resulted in a set of automated applications that collect spoken and written content from online courses, build and annotate domain-specific corpora, and extract domain-specific wordlists based on the criteria used by the Academic Word List.

Second, acquiring and measuring language come hand-in-hand. Tools that help learners acquire new language should also include methods for testing it. This thesis takes an existing general-purpose vocabulary test – the EFL Vocabulary Test – and automates it for domain-specific language. EFL uses a combination of real and imaginary words (pseudowords) to test learners' receptive vocabulary, the automation of which has resulted in two application. The first can be used to generate domain-specific pseudowords from domain-specific wordlists, using character-grams of a specified length; while the second is used in conjunction with the first to generate domain-specific vocabulary tests.

Third, for content-based language learning to be used successfully, the language components must be smoothly integrated into courses without disturbing the original content. Moreover, our first two challenges focused on single domain-specific words, yet vocabulary support should include not just single words, but also multi-word lexical items such as collocations and lexical bundles. The culmination of the work in this thesis has resulted in the creation of F-Lingo, a Chrome extension that works on top of FutureLearn MOOCs to provide online learners with language resources for domain-specific words, phrases (collocations and lexical bundles), and concepts. It is completely automated, though would also lend itself to selective teacher intervention.

Finally, a learner-based evaluation has been conducted, where 109 participants were tracked using the F-Lingo Chrome extension. This evaluation provided insight into the way in which learners interact with F-Lingo, showing, for example, that they spend more time looking at additional lexical information for concepts than they do for words or phrases. The next step would be to conduct an extensive longitudinal study, measuring learner's vocabulary before, during, and after using the Chrome extension, in turn measuring the effectiveness of F-Lingo as a language resource or for language acquisition. (https://researchcommons.waikato.ac.nz/handle/10289/12929)

DEVELOPING NEW TECHNIQUES TO IMPROVE LICENCE PLATE DETECTION SYSTEMS FOR COMPLICATED AND LOW QUALITY VEHICLE IMAGES

Meeras Salman Juwad AL-Shemarry
Meerassalmanjuwad.al-shemarry@usq.edu.au
University of Southern Queensland, Australia

INTELLIGENT transportation systems (ITSs) play a very important role in people's lives in many respects. One of the most important ITS applications is for automatic number plate recognition systems. Over the years, many algorithms have been developed for detecting licence plates (LPs) from vehicle images or from a sequence of images in a video. Many existing ITSs work only under good conditions or normal environments.

It is still challenging to find effective techniques to identify LPs under difficult conditions, such as low/high contrast, bad illumination, foggy, dusty, or distorted by high speed or bad

weather. New techniques are needed to improve the performance of existing detection systems. The main contributions of this research as follows:

1) Effective methods were developed for detecting LPs under complicated conditions, such as low/high contrast, bad illumination, foggy, dusty, and distorted by high speed and bad weather. They improved the detection system performance with less execution time and a low false-positive rate.
2) Presenting new preprocessing and extraction techniques that can improve the classification accuracy.
3) Investigating which method is better to achieve the main requirements of an LPD system under difficult conditions.

The research significance is that the proposed methods can improve the performance of the existing ANPR systems under complicated conditions. In addition, the outcomes will contribute to increasing the quality of transport systems with better efficiency and safety.

The motivation of this research is to select good software components for detecting complicated LP problems. Those components play an important role in the quality of the LPD system. Therefore, the preprocessing and feature extraction techniques should be selected and developed carefully to improve system weaknesses.

In this thesis, novel methods are developed for licence plate detection (LPD) systems to extract key features, and classify the LP region from complicated vehicle images based on preprocessing methods and machine learning algorithms with several types of texture descriptors.

In order to identify LPs from complicated vehicles images, four LPD methods were developed in this research. The first, is a three-level local binary pattern operator based on an ensemble of Adaboost cascades classifiers. The second method, introduces a new texture descriptor based on a multi-level preprocessing stage with extended local binary pattern descriptor using an extreme learning machine classifier. The third, develops learning-based preprocessing methods using a local binary pattern and a median filter histogram of the oriented gradient with support vector machine classifier for detecting complicated LPs. Finally, for identifying distorted LPs using hybird features, median robust extended local binary pattern and speeded-up robust with an extreme learning machine classifier. The experimental results show that both of the third and fourth algorithms perform very well in LP detection accuracy rate compared with first and second algorithms. Also, the false positive rate (FPR) for both methods is better than those algorithms. The second and fourth methods carry out significant classification of different types of LP key features. The first approach takes much less execution time and produces high FPR compared to the three other methods. But it was a good technique for selecting suitable preprocessing and extraction methods, for detecting LPs from low quality vehicleimages.

The experimental results proved the efficiency of the proposed approaches for detecting difficult regions of the LP inside a vehicle image. The findings suggest that the outcomes of this study can improve the performances of existing LPD systems. They can assist in law enforcement with an ITS system. Also, it can be effectively used to detect LPs in real-time applications under difficult conditions. Method 1: The overall performance evaluation for detection, precision, and F-measure rates are 98.56%, 95.9%, and 97.19%, respectively, with an FPR of 5.6%. The average detection time for the whole system per vehicle image was 2.001ms. Method 2: The detection accuracy and FPR compare with Method 1 were improved by 0.54% and 0.56%, respectively. The classification and detection rates are 99.78% and 99.10%, respectively, with an FPR of 5%. The average execution time for the whole detection system per vehicle image was 2.4530ms. Method 3: This method yielded an excellent improvement over existing methods, a 4% improvement for the FPR, and 1.50% for accuracy with execution time. The overall performance evaluation for the object localization metrics of the detection or recall rate is 99.62%, with an FPR of 1.675%. The average of the runtime for the whole detection system per vehicle image was 2.2187ms. Method 4: The accuracy and detection rates are 97.92% and 99.71, respectively, with the FPR of 2.24%. The average runtime for the whole detection system per vehicle image was 2.108 ms. The method was superior in the performance and execution time over the existing proposed methods in this research.

The future work will investigate the possibility of using those methods to improve ANPR applications. To facilitate the further development of this work, a few key areas below have been explored.

Concerning the first and second algorithms, they can be improved to further reduce the false positive rate and extraction time using the preprocessing techniques.

One future improvement could be to eliminate those LP objects that look like the LP and have the same characteristics as LP regions, such as texts or commercial signs and logo objects. This step would decrease the processing time as well as the memory required to process the LP detection task.

In addition, using a combination of several supervised machine learning algorithms instead of a single one is very efficient. This is a preferable solution for capturing more information about the LP area and increase the detection system and classification accuracies. Those methods can be applied to different types of LP datasets, such as Australian car LPs, Arabic car LPs, and so on. More generally the proposed methods could be used by other fields that are related to objects detection subjects. Due to using supervised learning techniques, there is no limitation in those methods which are associated with objects shape, color, and edge and so on.

Further study is required to take account of other challenges and to enhance this work for dealing with other difficult conditions, such as licence plates with difficult tilt, rain, and snow in images. The detected LPs are normally stored as images in the memory and used by transportation systems to complete their tasks. This needs more storage devices, therefore, the LP recognition stage is required. This stage works to recognize the LP number as a text using deep learning algorithms and template matching techniques with optical character recognition (OCR).

This thesis studied offline detection methods, but it is

desirable for this work to be applied to real online LPD systems to see the impact of this research. This will require more work. Therefore, all of the proposed methods need to be employed for online detection. This would be a significant achievement in the field of transport systems for work under difficult conditions.

IMPROVED K-MEANS CLUSTERING ALGORITHMS

Tong Liu
t.liu@massey.ac.nz
Massey University, New Zealand

K-MEANS clustering algorithm is designed to divide the data points into subsets with the goal that maximizes the intra-subset similarity and inter-subset dissimilarity where the similarity measures the relationship between two data points. As one of the most popular and widely used unsupervised machine learning techniques, K-means clustering algorithm has been applied in a variety of areas such as artificial intelligence, data mining, biology, psychology, marketing, medicine, etc.

The result of K-means clustering algorithm depends on the initialization, the similarity measure, and the predefined cluster number. Previous research focused on solving a part of these issues but has not focused on solving them in a unified framework. However, fixing one of these issues does not guarantee the best performance, so that it is significant to conduct further research to improve it. This thesis conducts an extensive research on K-means clustering algorithm aiming to solve the issues of the initialization, the similarity measure, and the determination of cluster number simultaneously.

First, the Initialization-Similarity (IS) clustering algorithm is developed to solve the issues of the initialization and the similarity measure of K-means clustering algorithm in a unified way. Specifically, the initialization of the clustering is fixed by using sum-of-norms (SON) which outputs the new representation of the original dataset and the similarity matrix is learnt based on the data distribution. Furthermore, the derived new representation is used to conduct K-means clustering.

Second, a Joint Feature Selection with Dynamic Spectral (FSDS) clustering algorithm is developed to solve the issues of the cluster number determination, the similarity measure, and the robustness of the clustering by selecting effective features and reducing the influence of outliers simultaneously. Specifically, the similarity matrix is learnt based on the data distribution as well as adding the ranked constraint on the Laplacian matrix of the learned similarity matrix to automatically output the cluster number. Furthermore, the L2,1-norm is employed as the sparse constraints on the regularization term and the loss function to remove the redundant features and reduce the influence of outliers respectively.

Third, a Joint Robust Multi-view (JRM) spectral clustering algorithm is developed. JRM considers information from all views of a multi-view dataset to conduct clustering while solving initialization, similarity measure, cluster number determination, feature selection, and outlier reduction issues for multi-view data in a unified way. Extensive experiments have been carried out to evaluate the performance of all the proposed algorithms on real-world data sets from UCI machine learning repository. The results obtained and presented in the thesis show that the proposed algorithms outperformed the state-of-the-art comparison clustering algorithms. More specifically, the proposed IS clustering algorithm increased average ACC by 6.4% and 3.5% compared to K-means and Spectral clustering algorithm on data sets Digital, MSRA, Segment, Solar, USPS, USPST, Waveform, Wine, Wireless, and Yale. The proposed FSDS clustering algorithm increased average ACC by 12.56%, 4.43%, 5.79%, and 11.68% respectively compared to K-means clustering algorithm, spectral clustering algorithm, clustering with adaptive neighbors algorithm, and robust continuous clustering algorithm on datasets Cardiotocography, Diabetic Retinopathy, Parkinson Speech, German Credit, Australian Credit Approval, Balance Scale, Credit Approval, and Musk. The proposed JRM algorithm increased average ACC by 41.95%, 33.49%, 40.01%, 34.38%, and 39.32% respectively, compared to best K-means clustering, concatenation-based K-means clustering, graph-based system algorithm, adaptively weighted Procrustes algorithm, and multi-view low-rank sparse subspace clustering algorithm on datasets 3Source, Washington, Flowers, Texas, Wisconsin, and Cornell.

The proposed clustering algorithms in this thesis solve the determination of the cluster number K, initialization, similarity measure and robustness issues of K-means clustering algorithm in a unified way. In addition, the convergences of the proposed optimization methods for the proposed objective functions are theoretically proved. The proposed algorithms can be used in a wide range of applications such as customer behavioral segmentation, anomalies detection, cyber security, sensor measurements sorting, inventory categorization, etc.

CONTRIBUTION TO THE DEVELOPMENT OF ALGORITHMS BASED DEEP LEARNING ARCHITECTURES FOR MOBILE ROBOTIC'S APPLICATIONS

Nabila ZRIRA
nabilazrira@gmail.com
University Mohammed V, Rabat, Morocco

MOBILE robotic encounters many problems as robots move into dynamic environments. Particularly, the large amount of variety faced in real-world environments is extremely difficult for existing robotic applications to handle. This requires the use of machine learning methods, which can learn models for each task. Most of these methods require significant hand-designed representations to learn classification models. However, designing good features is crucial to the success of a machine learning algorithm for a specific problem, such features are often unintuitive and need considerable effort to design.

Recently, deep learning can solve these problems by learning features directly from data without any human intervention. In such architectures, the inference consists of a series of matrix multiplications to weight inputs followed by element-wise non-linear operations, and thus justify that the inference does not require optimization. Due to these advantages, we

propose different approaches in the context of mobile robotic applications using the existing deep learning methods.

This dissertation presents a direct application of deep learning in different mobile robotic tasks including object and scene classification as well as topological navigation and is encompassed in three major parts. In the object classification part, we propose several approaches using 2D/3D descriptors and Deep Belief Network (DBN). In the first contribution, we propose many local and global approaches for classifying both 2D and 3D objects using 2D/3D Bag of Words as well as our new global descriptor Viewpoint Features Histogram- Color (VFH-Color). VFH-Color combines both the color information and the geometric features extracted from the previous version of Viewpoint Features Histogram (VFH). In the second contribution, we extract geometric features from the segmented 3D point clouds using the VFH descriptor and then we learn these features with both generative and discriminative DBNs to evaluate their performance in the context of 3D object categorization.

The second part tackles the scene classification including two main contributions. The first one is centered on biologically inspired methods for representation and classification of indoor environments. It combines gist features and discriminative DBN, which showed previously its performance in object classification. The second contribution provides a new multimodal feature fusion for robust RGBD indoor scene classification. This approach consists of two separate Convolution Neural Networks (CNNs) trained on RGB and depth images, then combined with a late fusion network.

The last part presents our contributions in the topological navigation field. First, we propose a new method of exploring indoor environments by an autonomous mobile robot, as well as building topological maps. In this contribution, we define a new topological map building concept using global visual attributes that are extracted from omnidirectional images. Second, we extend our previous work by using Convolution Long Short-Term Memory (C-LSTM) to perform scene recognition-based topological mapping and localization. The C-LSTM involves CNN layers to extract features from the input data combined with LSTM to consider the information of the previous frames, thus learning the temporal dependencies of the robot movement.

The results obtained during this thesis are globally very promoting and encouraging. However, we focus on the limitation of some approaches and the problems encountered throughout these works, and, at the same time, the possible solutions to overcome them. Training deep neural networks with large datasets requires an increasing amount of computation resources. This might take from hours to weeks depending on the dataset, the computational power, and the algorithms being used for the training. However, the common limitation of all our approaches depends on the hardware used to learn our data. In our experiments, we used only the CPU device because of the limited graphic memory of our GPU card. Therefore, we fixed a limited number of epochs and the small size of image datasets, which may influence the obtained results. In future work, we will implement our approaches on the GPU card to be massively parallelized and thus sped up.

Since our 2D/3D object classification approaches in the first part showed good results compared with the state-of-the-art, we will extend our work to object grasping which constitutes an essential component in an autonomous robotic manipulation system operating in human environments.

In the second part, we will exploit the object classification results to perform indoor scene recognition through the objects present in the scene. We will assign a probability to each object class, then count all the object probabilities to predict the scene class. In this way, the object and scene classification will be two dependent tasks that can be used in mobile robotic navigation.

In the last part, we will propose a semantic navigation approach based on the sequence to sequence learning. Such an approach will provide high-level communication between robots and humans. Besides omnidirectional and RGB images, in the next work, we will integrate depth information to perform navigation with RGBD sensors.

Data Mining for Personalised Clinical Decision Support Systems

Wee Pheng Goh
weepheng.goh@usq.edu.au
University of Southern Queensland, Australia

With the addition of new drugs in the market each year, the number of drugs in drug databases is constantly expanding, posing a problem when prescribing medications for patients, especially elderly patients with multiple chronic diseases who often take a large variety of medications. Besides the issue of polypharmacy, the need to handle the rapid increase in the volume and variety of drugs and the associated information exert further pressure on the healthcare professional to make the right decision at point-of-care. Hence, a robust decision support system will enable users of such systems to make decisions on drug prescription quickly and accurately.

Although there are many systems which predict drug interactions, they are not customised to the medical profile of the patient. The work in this study considers the drugs that the patient is taking and the drugs that the patient is allergic to before deciding if a specific drug is safe to be prescribed. To exploit the vast amount of biomedical corpus available, the system uses data mining methods to evaluate the likelihood of a drug interaction of a drug pair based on the textual description that describes the drug pair. These methods lie within the prediction layer of the conceptual three-layer framework proposed in the thesis. The other two layers are the knowledge layer and the presentation layer. The knowledge layer comprises information on drug properties from drug databases such as DrugBank. The presentation layer presents the results via a user-friendly interface. This layer also obtains information from the user the drug to be prescribed and the medical profile of patients. Models used in these data mining methods include the network approach and the word embedding approach. A survey conducted on dentists found positive response in the use of such a system in helping them in drug prescription which result in a better treatment outcome.

One possible extension to the current work include the leveraging of Semantic Web technology with alternative data repositories such as PubMed and compare the results to evaluate if it is more efficient. Performance of the experiment can also be further evaluated by having the models amalgamated to form an ensemble model. The research has made the novel discovery that drug interactions are associated with similarities derived from their feature vectors. Similarity ratio of a drug-pair can be obtained from the paths that link the common drugs within the set of interacting drugs of the respective drug-pair. This results in a significant contribution relating to the design of personalised clinical decision support systems for use in healthcare institutions, transforming the clinical workflow at point-of-care within the healthcare domain.

EVENTS/CONFERENCES SPONSORED BY TCII

Past Events/Conferences

IEEE ICKG 2020
The 11th IEEE International Conference on Knowledge Graph (ICKG-2020)
Nanjing, China (Virtual Conference)
August 9-11, 2020
http://ickg2020.bigke.org

Knowledge Graph deals with fragmented knowledge from heterogeneous, autonomous information sources for complex and evolving relationships, in addition to domain expertise. The IEEE International Conference on Knowledge Graph (ICKG), previously entitled ICBK until 2019, provides a premier international forum for presentation of original research results in Knowledge Graph opportunities and challenges, as well as exchange and dissemination of innovative, practical development experiences. The conference covers all aspects of Knowledge Graph, including algorithms, software, platforms, and applications for knowledge graph construction, maintenance, inference and applications.

ICKG 2020 drew researchers and application developers from a wide range of Knowledge Graph related areas such as knowledge engineering, big data analytics, statistics, machine learning, pattern recognition, data mining, knowledge visualization, high performance computing, and World Wide Web. By promoting novel, high quality research findings, and innovative solutions to challenging Knowledge Graph problems, the conference continuously advanced the state-of-the-art in Knowledge Graph techniques.

ICKG-2020 hosted 84 papers. Themes and topics included: Foundations, algorithms, models, and theory of Knowledge Graph processing; Knowledge engineering with big data; Machine learning, data mining, and statistical methods for Knowledge Graph science and engineering; Acquisition, representation and evolution of fragmented knowledge; Fragmented knowledge modeling and online learning; Knowledge graphs and knowledge maps; Knowledge graph security, privacy and trust; Knowledge graphs and IoT data streams; Geospatial knowledge graphs; Ontologies and reasoning; Topology and fusion on fragmented knowledge; Visualization, personalization, and recommendation of Knowledge Graph navigation and interaction; Knowledge Graph systems and platforms, and their efficiency, scalability, and privacy; Applications and services of Knowledge Graph in all domains including web, medicine, education, healthcare, and business; Crowdsourcing, deep learning and edge computing for graph mining; and Rule and relationship discovery in knowledge graph computing.

Upcoming Events/Conferences

WI-IAT 2020
The 2020 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology
Melbourne, Australia (Virtual Conference)
December 14-17, 2020

Web Intelligence and Intelligent Agent Technology (WI-IAT) aims to achieve a multi-disciplinary balance between research advances in theories and methods usually associated with collective intelligence, data science, human-centric computing, knowledge management, network science, autonomous agents and multi-agent systems. It is committed to addressing research that both deepen the understanding of computational, logical, cognitive, physical, and social foundations of the future Web, and enable the development and application of intelligent technologies. WI-IAT '20 provides a premier forum and features high-quality, original research papers and real-world applications in all theoretical and technology areas that make up the field of Web Intelligence and Intelligent Agent Technology.

The 2020 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT '20) provides a premier international forum to bring together researchers and practitioners from diverse fields for presentation of original research results, as well as exchange and dissemination of innovative and practical development experiences on Web intelligence and intelligent agent technology research and applications. The theme for the WI-IAT 20 is "Web Intelligence = AI in the Connected World". The main topics and themes of this year's conference are Web of People, Web of Trust, Web of Things, Web of Data and Web of Agents, with a special track dedicated to Emerging Web in Health and Smart Living in the 5G era. There are approximately 79 topics spread across these 6 tracks.

WI-IAT '20 is a forum for research, application as well as Industry/Demo-Track paper submissions. Tutorial, Workshop and Special-Session proposals and papers are also welcome.

ICDM 2020
IEEE International Conference on Data Mining
Sorrento, Italy (Virtual Conference)
November 17-20, 2020
http://www.icdm2020.bigke.de

The IEEE International Conference on Data Mining series (ICDM) has established itself as the world's premier research conference in data mining. It provides an international forum for presentation of original research results, as well as exchange and dissemination of innovative, practical development experiences. The conference covers all aspects of data mining, including algorithms, software and systems, and applications. ICDM draws researchers and application developers from a wide range of data mining related areas such as statistics, machine learning, pattern recognition, databases and data warehousing, data visualization, knowledge-based systems, and high performance computing. By promoting novel, high quality research findings, and innovative solutions to challenging data mining problems, the conference seeks to continuously advance the state-of-the-art in data mining. Besides the technical program, the conference features workshops, tutorials, panels.

The topics of interest at this year's conference include: Foundations, algorithms, models and theory of data mining, including big data mining.

Deep learning and statistical methods for data mining; Mining from heterogeneous data sources, including text, semi-structured, spatio-temporal, streaming, graph, web, and multimedia data; Data mining systems and platforms, and their efficiency, scalability, security and privacy; Data mining for modelling, visualization, personalization, and recommendation; Data mining for cyber-physical systems and complex, time-evolving networks; and Applications of data mining in social sciences, physical sciences, engineering, life sciences, web, marketing, finance, precision medicine, health informatics, and other domains.

Accepted papers will be published in the conference proceedings by the IEEE Computer Society Press. Awards will be conferred at the conference to the authors of the best paper and the best student paper. A selected number of best papers will be invited for possible inclusion, in an expanded and revised form, in the Knowledge and Information Systems journal (http://kais.bigke.org/) published by Springer.

ICHI 2020
The Eighth IEEE International Conference on Healthcare Informatics
Virtual Conference
Nov. 30-Dec. 3, 2020
https://ichi2020.de

The Eighth IEEE International Conference (ICHI 2020) will take place online in 2020. ICHI 2020 is a premier community forum concerned with the application of computer science, information science, data science, and informatics principles, as well as information technology, and communication science and technology to address problems and support research in healthcare, medicine, life science, public health, and everyday wellness.

ICHI 2020 serves as a venue for the discussion of innovative technologies and implementation science, highlighting end-to-end applications, systems, and technologies, even if available only in prototype form. The conference highlights the most novel technical contributions to stakeholder-centered technology innovation for benefiting human health and the related social and ethical implications. ICHI 2020 will feature keynotes, a multi-track technical program including papers, posters, panels, workshops, tutorials, an industrial track, and a doctoral consortium.

The tracks in this year's conference include Analytics, Human Factors and Systems. All accepted submissions will be published in IEEE Xplore and are indexed in other Abstracting and Indexing (A&I) databases. Selected papers will be invited to submit the extended version to the Journal of Healthcare Informatics Research and a fast track review will be conducted for the extended papers.

IEEE BigData 2020
The 2020 IEEE International Conference on Big Data (IEEE BigData 2020)
Atlanta, USA (Virtual Conference)
December 10-13, 2020
http://bigdataieee.org/BigData2020/

The IEEE International Conference on Big Data 2020 (IEEE BigData 2020) provides a leading forum for disseminating the latest research in Big Data. IEEE Big Data brings together leading researchers and developers from academia, research and the industry from all over the world to facilitate innovation, knowledge transfer and technical progress in addressing the 5 V's (Velocity, Volume, Variety, Value and Veracity) of Big Data. The purpose of the conference is to identify deep technical and scientific nature of big data problems, and share the future direction on the development of next-generation solutions for data-driven decision making. The conference will attract high-quality theory and applied research findings in big data science and foundations, big data infrastructure, big data management, big data search & mining, privacy/security, and big data applications.

The IEEE BigData 2020 received more than 610 full papers in the main conference and industry and government program. The conference hosts a number of workshops and poster presentations that are published along with accepted papers in the conference proceedings. The special sessions and tracks include themes such as Data Marketing, Intelligent Data Mining, HealthCare Data, Machine Learning on Big Data, Information Granulation and Explainable Artificial Intelligence in Safety Critical Issues. With a focus on Big Data, the conference also runs a number of sponsored tutorials run by experts in the fields of computing, computer science and data science.

AAMAS 2021
The 20th International Conference on Autonomous Agents and Multi-Agent Systems
London, UK
May 3-7, 2021
http://aamas2021.soton.ac.uk

AAMAS is the largest and most influential conference in the area of agents and multiagent systems, bringing together researchers and practitioners in all areas of agent technology and providing and internationally renowned high-profile forum for publishing and finding out about the latest developments in the field. AAMAS is the flagship conference of the non-profit International Foundation for Autonomous Agents and Multiagent Systems (IFAAMAS).

AAMAS'21 is inviting submissions of technical papers describing significant and original research on all aspects of the theory and practice of autonomous agents and multiagent systems. Papers are associated with area of interest, which include: Coordination, Organisations, Institutions, and Norms; Markets, Auctions, and Non-Cooperative Game Theory; Social Choice and Cooperative Game Theory; Knowledge Representation, Reasoning, and Planning; Learning and Adaptation; Modelling and Simulation of Societies; Humans and AI / Human-Agent Interaction; Engineering Multiagent Systems; Robotics; and Innovative Applications.

AAMAS-2021 will feature three special tracks, the Blue Sky Ideas Track, the JAAMAS Track, and the Demo Track, each with a separate Call for Papers. The focus of the Blue Sky Ideas Track is on visionary ideas, long-term challenges, new research opportunities, and controversial debate. The JAAMAS Track offers authors of papers recently published in the Journal of Autonomous Agents and Multiagent Systems (JAAMAS) that have not previously appeared as full papers in an archival conference the opportunity to present their work at AAMAS-2021. The Demo Track, finally, allows participants from both academia and industry to showcase their latest developments in agent-based and robotic systems.

AAAI 2021
The 35th AAAI Conference on Artificial Intelligence
Virtual Conference
February 2-9, 2021
https://aaai.org/Conferences/AAAI-21/

The Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI-21) will be held virtually February 2-9, 2021. The general chair will be Qiang Yang (Hong Kong University of Science and Technology, Hong Kong) and the program chairs will be Kevin Leyton-Brown (University of British Columbia, Canada) and Mausam (Indian Institute of Technology Delhi, India).

The purpose of the AAAI conference is to promote research in artificial intelligence (AI) and scientific exchange among AI researchers, practitioners, scientists, and engineers in affiliated disciplines. AAAI-21 will have a diverse technical track, student abstracts, poster sessions, invited speakers, tutorials, workshops, and exhibit and competition programs, all selected according to the highest reviewing standards. AAAI-21 welcomes submissions on mainstream AI topics as well as novel crosscutting work in related areas.

AAAI-21 welcomes submissions reporting research that advances artificial intelligence, broadly conceived. The conference scope includes machine learning (deep learning, statistical learning, etc), natural language processing, computer vision, data mining, multiagent systems, knowledge representation, human-in-the-loop AI, search, planning, reasoning, robotics and perception, and ethics. In addition to fundamental work focused on any one of these areas, work is encouraged that cuts across technical areas of AI, bridges between AI and a related research area (e.g., neuroscience; cognitive science) or develops AI techniques in the context of important application domains, such as healthcare, sustainability, transportation, and commerce. Most papers in AAAI-21 will be part of the “main track”. In addition to typical AI areas, there will be three focus areas within the main track that highlight timely topics. All main track papers will be reviewed according to the same criteria and via the same process. A second track of the conference will focus on AI for Social Impact. Papers in this track will be reviewed according to a different evaluation rubric than the main track. The same reviewing schedule will be followed for all papers.

SDM21
The 2021 SIAM International Conference on Data Mining
Virtual Conference
April 29 – May 1, 2021
https://www.siam.org/conferences/cm/conference/sdm21

Data mining is the computational process for discovering valuable knowledge from data – the core of modern Data Science. It has enormous applications in numerous fields, including science, engineering, healthcare, business, and medicine. Typical datasets in these fields are large, complex, and often noisy. Extracting knowledge from these datasets requires the use of sophisticated, high-performance, and principled analysis techniques and algorithms. These techniques in turn require implementations on high performance computational infrastructure that are carefully tuned for performance. Powerful visualization technologies along with effective user interfaces are also essential to make data mining tools appealing to researchers, analysts, data scientists and application developers from different disciplines, as well as usable by stakeholders.

SDM has established itself as a leading conference in the field of data mining and provides a venue for researchers who are addressing these problems to present their work in a peer-reviewed forum. SDM emphasizes principled methods with solid mathematical foundation, is known for its high-quality and high-impact technical papers, and offers a strong workshop and tutorial program (which are included in the conference registration). The proceedings of the conference are published in archival form, and are also made available on the SIAM website.

SDM21 has three main categories in Methods and Algorithms, Applications and Human Factors and Social Issues. Each category contains a multitude of related themes as topics for papers. Please visit the website for more information.

IJCAI 2021
The 30th International Joint Conference on Artificial Intelligence
Montreal, Canada
August 21-26, 2021
http://www.ijcai-21.org/

Submissions are invited for the 30th International Joint Conference on Artificial Intelligence (IJCAI-21), which is planned to be held in Montreal, Canada, from August 21st to August 26th, 2021. Starting from 1969, IJCAI has remained the premier conference bringing together the international AI community to communicate the advances and achievements of artificial intelligence research.

Submissions to IJCAI-21 should be significant, original, and previously unpublished results on all aspects of artificial intelligence. Papers on novel AI research problems and novel application domains are especially encouraged. The submission site for IJCAI-21 opens December 30, 2020.

IEEE Computer Society
1730 Massachusetts Ave, NW
Washington, D.C. 20036-1903

Non-profit Org.
U.S. Postage
PAID
Silver Spring, MD
Permit 1398