THE IEEE
Intelligent Informatics
BULLETIN

IEEE COMPUTER SOCIETY

IEEE Computer Society
Technical Committee
on Intelligent Informatics

December 2021 Vol. 21 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 Multi-modal Data Mining and Advanced Machine Learning Research

A SHORT INTRODUCTION TO THE INTELLIGENT DATA ENGINEERING AND ANALYTICS LAB

Lin Li, Jingling Yuan, Qing Xie and Xian Zhong *

Intelligent Data Engineering and Analytics Lab, School of Computer and Artificial Intelligence, Wuhan University of Technology, Wuhan, China.*

E-mail: {Gcathylilin, yjl, felixxq, zhongx}@whut.edu.cn

Abstract

The Intelligent Data Engineering and Analytics (IDEA) Laboratory at Wuhan University of Technology (WUT) is a joint effort aiming to research and teach on data science and artificial intelligence. IDEA has intensive research collaboration with well-known industry partners. Leading by Professor Lin Li, the founding director, IDEA currently has nine academics and 50+ research students. In the past few years, the lab has secured over 10 million CNY competitive research grants from provincial and national research schemes funded by governments and corporations. IDEA conducted world-class research and had outcomes published in premier journals, such as VLDBJ, TOIS, TKDE, TSC, TOIT, IPM, and world-top conferences, such as AAAI, WWW, ICDE, ICDM, CIKM, ICMR, ICASSP. IDEA also received many national and international awards for their achievements in Information Management, Big Data, and Web Intelligence, including Top 1 of OpenLive QA Task at NTCIR-13, Best Student Paper Award in CCF Big Data 2020, and Best Student Paper Award in WISE 2020.

I. INTRODUCTION

The Laboratory of Intelligent Data Engineering and Analytics (IDEA) was founded by Professor Lin Li in 2011. The lab has intensively researched on topics of text mining, recommender system, Question & Answering, social computing, multi-modal machine learning, sequential prediction, Lightweight/Parallel Machine Learning. In the past few years, the lab has received over 10 millions RMB competitive research grants from provincial or national funding schemes offered by governments, corporations, and private sectors, such as National Natural Science Foundation of China (NSFC), National Social Science Fund of China, China Scholarship Council Project (CSC), Department of Science and Technology of Hubei Province, China, and Deloitte[1]. Supported by these funds, IDEA's research has been productive and sustainable, with outcomes published in top-tier data science and artificial intelligence conferences and journals, such as VLDBJ, TOIS, TKDE, TSC, TOIT, IPM, AAAI, WWW, ICDE, ICDM, CIKM, ICMR, and ICASSP.

The IDEA lab has received many awards at national and international levels. The work on Community Question & Answer (CQA) was awarded *Top 1 Online Test Results of OpenLive QA Task* at NTCIR-13[2]. The study on traffic flow forecasting received Best Student Paper Award in CCF Big Data 2020[3]. In the 21st International Conference on Web Information Systems Engineering (WISE 2020), IDEA's work on legal judgment prediction was also highly praised and awarded with Best Student Paper Award, too.

With a glance given in the Introduction, more details about IDEA will be provided below. Section II describes the research areas focused by IDEA. In Section III, the impact delivered by IDEA by its research to the real-world will be discussed. After that, some research activities conducted by IDEA will be highlighted in Section IV. Finally, in Section V IDEA's vision on future research will be presented.

II. RESEARCH AREAS

The IDEA lab is mainly focused on data science and artificial intelligence research and extensively involved in interdisciplinary research, such as computing social science, LegalAI, and intelligent transportation systems. The key research areas include:

- Recommender system
- Deep clustering
- Sequential prediction
- Cross-modal retrieval
- Multi-modal machine translation
- Question&Answering
- Social computing, multi-modal user profiling
- Law intelligence, legal judgment prediction
- Lightweight/Parallel machine learning

A. Recommender System

Recommender System, an active domain of information filtering, takes advantage of information from various sources to provide users with predictions and recommendations of products and services (e.g., movies, books, applications, websites, and travel destinations). The recommendation aims to facilitate user decision-making, improving user experience in Web services, achieve revenue increase for online businesses and merchants, and so on.

[1]https://www2.deloitte.com/cn/en.html
[2]http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings13/NTCIR/toc_ntcir.html
[3]http://bigdata2020.swu.edu.cn/beststudentpapers/

IDEA's strengths on recommender systems include:

- Geo-based and group recommendation
- Cross-domain, cold-start and long-tail recommendation
- Streaming recommendation
- Knowledge and content driven recommendation
- Interpretable, Interactive and dynamic recommendation

IDEA's research on recommender systems can be showcased on [1], [2], [3], [4], [5], [6], [7], [8], [9].

B. Cross-modal / Multimedia Retrieval

Besides the development of content-based multimedia retrieval, cross-modal retrieval - one of the most desired services in data powered machine learning application - focuses on the multi-modal datasets which contain more than one modality, such as Twitter tweets, Instagram messages, yelp dish recommendations and the information on other social platforms. Cross-modal retrieval aims to give users a multi-modal view on information acquisition, and promote advertising effectiveness to increase sales for online business and offline stores.There are many challenges, such as feature representation, semantic gap between image visual features and semantical meaning, and the lack of training samples. IDEA has made many significant achievements in this area, especially in:

- Image-text, Image-video cross-modal retrieval
- Implicit alignment cross-modal retrieval
- Dataset generation for specific areas (as shown in Figure 1)
- Effective automatic image annotation.
- Deep hashing for multi-label image retrieval

These research outcomes can be found on [10], [11], [12], [13], [14], [15], [16], [17], [18].

C. Sequential Prediction

Sequential Prediction focuses on modelling diverse kinds of interaction patterns across a series of elements in chronological order (e.g. purchase history, urban event, air quality index) to obtain hints about future elements. It is a basic technology for helping organizational and social entities in resource allocation and decision-making. Our strengths include:

- Spatial-temporal sequence prediction
- Cross-correlation based sequence prediction
- Knowledge distillation based sequential learning

See [19], [20], [21] for examples of our research in this area.

D. Multi-modal Machine Translation

Multi-modal machine translation (MMT) can use the image information corresponding to the source text and improve the translation quality. Since text and image belong to different data modality, bridging the modality gap between them is one of the challenges of MMT. Our studies include the following:

- Multi-modal Machine Translation with Attention
- Multi-perspective Multi-modal Machine Translation
- Machine Translation Enhancement with Multi-modal attention

See [22], [23], [24] for examples of our research in this area.

E. Question&Answering

QA is a classic natural language processing (NLP) task, which aims at building systems that automatically answer questions formulated in natural language, such as community question answering services(CQA), document-based question answering, question answering over knowledge base (KBQA), etc. Our strengths include:

- Cross-lingual Open QA.
- Translation based CQA
- Context enhanced KBQA.

See [25], [26], [27], [28] for examples of our research in this area.

F. Multi-modal User Profiling

Multi-modal user profiling means exploiting the technology of machine learning and the multi-modal data (e.g., text, image, code) generated by users to predict attributes of users, such as demographic attributes, hobby attributes, preference attributes, etc. The key work of multi-model user profiling is to label users with some highly refined features that can summarize user characteristics through analysis of various user information - in a sense, digitizing users. The potential applications of user profiling includes precision marketing, data statistics, decision support, etc. Our strengths include:

- Sentiment-based gender classification
- Multi-modal cooperative gender classification
- Emotion classification

See [29], [30], [31] for examples of our research in this area.

G. Legal Judgment Prediction

Legal Judgment Prediction (LJP) is one of the most critical tasks in LegalAI, especially in the Civil Law system. In the Civil Law system, the judgment results are decided according to the facts and the statutory articles. One will receive legal sanctions only after he or she has violated the prohibited acts prescribed by law. One of the LJP tasks mainly concerns how to predict the judgment results from both the fact description of a case and the contents of the statutory articles. Our strengths include:

- Legal framework-driven interpretable charge prediction.
- External knowledge enhanced multi-label charge prediction.
- Multi-type legal machine reading comprehension.

See [32], [33], [34] for examples of our research in this area.

H. Lightweight/Parallel Machine Learning

The lightweight machine learning models accelerate the models to achieve efficient inference by simplifying the structure, pruning or optimizing the construction unit. Parallel machine learning models use multiple processors to improve the efficiency and computational power of the model. Our strengths include:

- Object detection based on lightweight backbone.
- Semantic segmentation based on lightweight backbone.
- Video stitching based on GPU parallel acceleration.

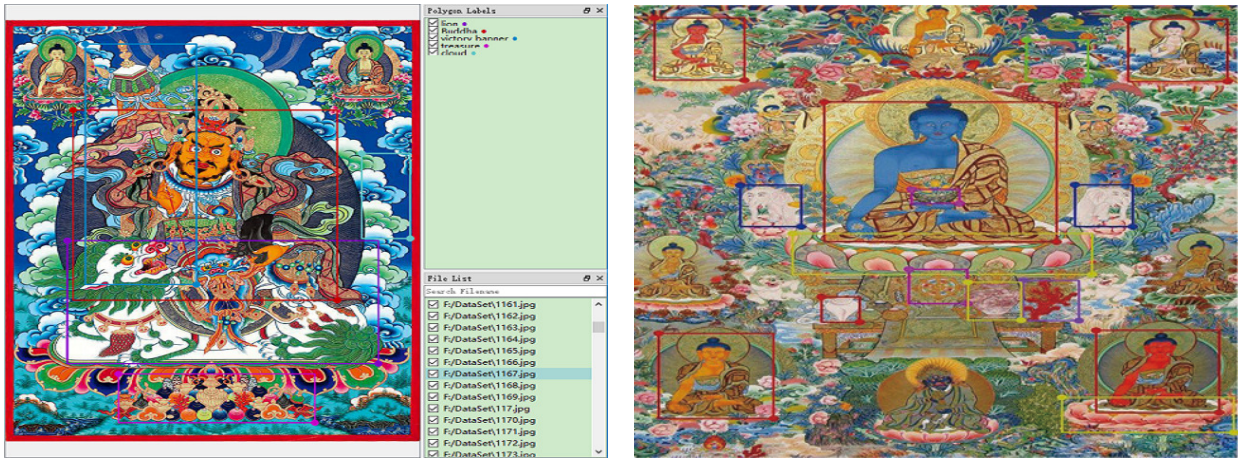


Fig. 1. The example of Dataset generation.

- Fast reduction and parallel processing of big data.

See [35], [36], [37], [38], [39] for examples of our research in this area.

I. Deep Clustering

Clustering aims to group data without label information, which is a crucial and challenging task in pattern analysis and machine learning. Deep clustering, a novel clustering method joints clustering and deep feature representation learning, has shown remarkable performance with real-world data (e.g., image, text documents, and graph-structured data). Our strengths include:

- Variational deep clustering
- Deep graph clustering
- Deep multi-view clustering

See [40], [41] for examples of our research in this area.

III. Research with Real-world Impact

This section summaries how the lab's research is applied when dealing with real world applications, improving business outcomes and benefiting society.

A. Rank Optimization of Personalized Web Information Retrieval

Personalized Web information retrieval is an effective way to improve the precision of traditional information retrieval, and then satisfy the information needs of Internet users. With the appearance of various types and large amount of information on the Web, users not only require retrieval precision, but also the retrieval efficiency and the privacy protection of their personal information. Funded by NSFC, this project focuses on how to improve the overall quality of Web information retrieval, including precision, efficiency and privacy protection. Our research includes query and time dependent re-ranking algorithms, the update mechanism of user profiles, the approaches to improving the efficiency of re-ranking algorithms, privacy protection based user profile modelling, privacy protection based re-ranking algorithms, and so on.

B. Small Business Credit Evaluation

With the development of digitization, networking, and socialization, the Internet has accumulated a large amount of information. The understanding of Internet information will help form a new type of credit evaluation method and promote financial innovation research in the Internet age. This project deeply analyzes and establishes a credit evaluation model for small business through the following activities:

1) Deep web data crawling and information extraction technology is studied to solve the problem of how traditional hyperlink-based web crawlers cannot crawl and index the information in the deep web;
2) Sentiment analysis algorithm for review data provides a viewpoint from Internet Word-of-mouth (positive, neutral and negative) to enhance the traditional credit evaluation methods. The main task is based on user reviews text crawled from some e-commerce platforms, and investigating how to transfer the rich sentiment analysis resources of traditional long texts to improve the sentiment classification quality of short texts.
3) Credit scoring algorithm based on enterprise association graph aims to extract the direct or indirect ownership or control relationship of enterprises in terms of capital, operation, purchase and sale, etc., and personal social network data from small business owners on the Internet.
4) A credit scoring model based on hidden factors is built. Because a small business with large amounts of data may be very sparse and missing features, hidden factor analysis is used to decompose sparse high-dimensional data into semantic low-dimensional hidden data, thus enhancing credit evaluation.

C. Multi-modal Machine Learning

With the rapid development of social networks and search engines, lots of interests has been witnessed in jointly dealing with multi-modal data such as text, image, audio and video. To cope with this scenario, information processing has to be transformed from the form of single modality to multi-modality. Therefore, challenges from the "media gap" (such

as how representations of different media types are inconsistent), are gaining increasing attention. Recently, deep neural networks(DNN), a major breakthrough in machine learning, has been employed to learn better multi-modal representations. This project works on multi-modal machine learning from representation, translation, fusion, alignment, and co-learning. Our recent studies in multi-modal representation are presented with new multi-modal algorithms and exciting multi-modal applications.

D. Continuous Querying and Query optimization on Streaming Data

As the volume of streaming data increases in current information era, how to perform efficient and optimized continuous query on streaming data has become one of the most significant problems for Data Stream Management System(DSMS). In the Big Data environment, the existing continuous query solution, including the approximate representation and similarity measure techniques, cannot meet the requirements of data variety and velocity, and current DSMSs fail to theoretically improve the techniques of continuous query on streaming data. Due to this situation, this project explores data characteristics in Big Data environment, and based on the theoretical analysis, aims to design efficient an framework of continuous query on streaming data, and perform query optimization for massive data stream applications. The key points of this project include: (1) customized approximate representation and similarity measure based on actual data stream; (2) the framework of efficient continuous query technique based on approximate representations; (3) multiple query optimization in massive data streams environments based on cost model and queries' structure. The outcomes of this project will theoretically propose the practical continuous query solutions as well as relevant techniques on streaming data, and provide the theoretical bases and key technique support for streaming data processing in Big Data environment.

E. Digital Content Management for Publication Industry

The publishing industry has accumulated a large amount of multimedia content. With the development of digitization, networking, and socialization, the unified and effective management and understanding of digital content will help promote the development of a new type of digital publication industry. This project collaborates with enterprises to industrialize scientific research results and provide software platform support for the digital publication industry. The main research content includes a digital content management platform and a digital content analysis platform.

1) The digital content management platform marks the knowledge fragments in the digital publication knowledge base, and introduces the semantic web technology to mark different types of associated descriptions and rich content tags.
2) The digital content analysis platform analyzes the information used to establish the user's interest model, and based on a deep understanding of digital content, sensitive word filtering technology, topic model analysis, semantic association and other technologies are studied to monitor online publishing, and discover content prohibited by laws and regulations.

F. AI+5G Service Robot and its Applications

This project explores the key technologies of AI+5G empowered service robot, and its applications for smart community platform: (1) Community service knowledge representation and reasoning methods will combine representation learning and transfer learning methods to study multi-level and multi-scale community service knowledge representation methods, cross-modal information representation methods for relationship analysis between entities/events. (2) The functional modulars are designed on key information extraction, user intention recognition, multi-semantic relationship extraction, automatic report generation, task-based multi-round question and answering, and so on. (3) Cloud-edge based smart community service platform is the integration of data privacy protection, multi-party secure computing, federated learning and other technologies.

G. Text mining for Contract Review

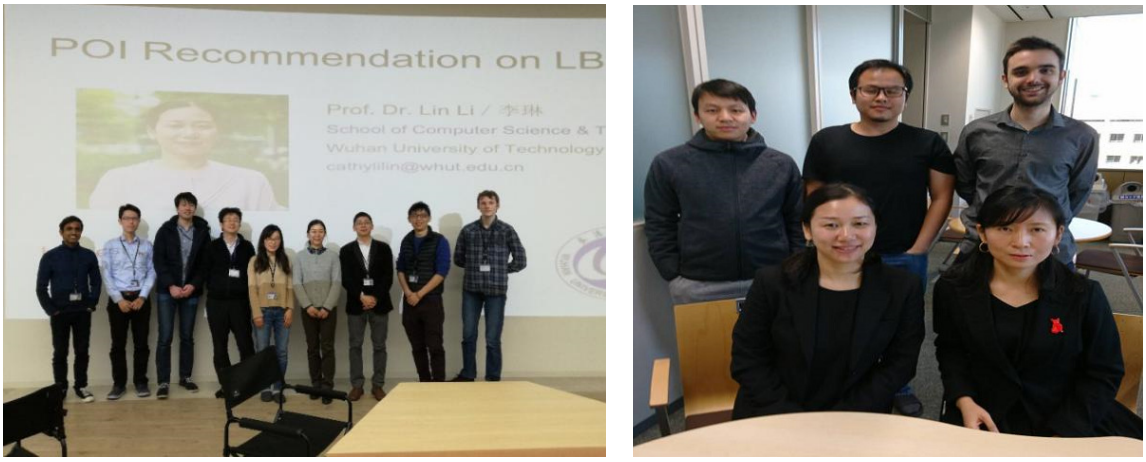
For a long time, the traditional contract management system has the problems of complex contract approval procedures and the lack of awareness of contract risk management. Therefore, the efficiency of business processing in the contract management process is low. At the social level, news outlets have reported that many people have signed "unequal" treaties due to lack of legal background. Enterprises and people have an urgent need for smart contract risk review and smart contract management. In order to solve this problem, this project explores natural language processing, deep transfer learning, information extraction, and OCR technology to automatically do contract risk review. Moreover, extracting summary generation and paper contract identification will help users analyze the contract process, avoid "contract traps", and improve the intelligent level of existing contract management.

H. Surveillance Multi-modal Data Mining for Smart City

The centralized mode of cloud computing makes it difficult to efficiently process a large amount of surveillance multi-modal data generated by monitoring equipment. For basic applications such as target recognition, efficient retrieval, semantic analysis, etc., This project will research cloud-side collaboration lightweight target detection model, graph reasoning based cross-modal retrieval, and the generation of video text descriptions associated with multiple events, which will improve the level of smart city safety monitoring, traffic management, and infrastructure operations. This project is based on the surveillance multi-modal data, and in response to the current needs of cloud-side collaboration scenarios in smart cities. The main work will include a fast parallel processing architecture for surveillance multi-modal data, the lightweight target detection model for edge devices, the temporal and spatial correlation action recognition method, cross-modal efficient retrieval based on key frame extraction and graphic

(a) Certificate (b) Visiting The University of Tokyo

Fig. 2. Sakura Science Exchange Program



(a) NEC (b) NII

Fig. 3. Invited Talks

reasoning, multi-event correlation for the video content text description generation. This project will assist government management and decision-making, and improve the service level of smart cities.

IV. Research Activities

The School of Computer and Artificial Intelligence, Wuhan University of Technology, provides hardware and software experimental environment for IDEA lab with 8 Dell and Inspur high-end servers and workstations, and a PC cluster system consisting of 40 high-end PCs and 4 servers. With this support, IDEA lab woks on multi-modal data mining and advanced machine learning.

A. Sakura Science Exchange Program

Japan Science and Technology Agency (JST)invites young, talented people from other countries and regions to Japan through the Sakura Science Exchange Program in a collaboration of industry-academia-government, to introduce and offer experience in Japanese science and technology. Beginning in 2014, and for a period of 6 years, over 33,000 young people visited Japan on this program. Professor Lin Li has successfully achieved the course of Japan-Asia Youth Exchange program in Science and led the IDEA team to visit The University of Tokyo, 2018, as shown in Figure 2. During the visit, our team members communicated closely with Japanese students in terms of study and life.

B. Invited Talks

IDEA lab was invited by NEC and National Institute of Informatics(NII), Japan to deliver the talk titled “POI recommendation on LBSNs”, as shown in Figure 3. This talk presented an insight on the edge research of recommendation system, which drew great interest from both academia and industry.

C. Conference Attending

The members of the laboratory enhanced communication with their peers by participating in national and international conferences, and give oral reports at many conferences with

(a) APWeb-WAIM 2019

(b) BESC 2019

(c) CCF BigData 2019

(d) Oral presentation in ICDCS 2017

Fig. 4. Attending Conferences

topics including big data, social computing, and data mining, as shown in Figure 4.

1) The Asia Pacific Web (APWeb) and Web-Age Information Management (WAIM) Joint Conference on Web and Big Data (APWeb-WAIM 2019)
2) The 6th International Conference on Behavioral, Economic and Socio-Cultural Computing (BESC 2019)
3) The 7th CCF Conference, BigData 2019
4) The 37th IEEE International Conference on Distributed Computing Systems (ICDCS 2017)

D. AI+BigData Competition·and Study

Lab members actively participate in various AI and big data competitions and achieve good results. At the same time, they have in-depth exchanges and discussions with researchers from both academia and industry through related short-course learning, as shown in Figure 5 and Figure 6 .

1) 2017 KDD summer School
2) 2019 National Smart City CUP, China
3) NTCIR-13, Top 1, online test results of OpenLive QA task at NTCIR-13.
4) 2019 China Conference on Knowledge Graph and Semantic Computing (CCKS)

V. Looking into the Future

The overall goal of IDEA lab is to develop into a world-class research lab for data engineering and a study community for data analysis talents, leading the research of multi-modal data mining and machine learning. The applications of our studied technologies involve data analysis and mining in multiple fields such as finance, intelligent transportation, medical health, and smart cities. We also welcome research collaborations internationally.

(a) 2017 KDD Summer School (b) 2019 Smart City CUP

Fig. 5. AI+BigData Competition·and Study-1

(a) NTCIR-13 and CCKS 2019

Fig. 6. AI+BigData Competition·and Study-2

REFERENCES

[1] X. Chen, L. Li, G. Xu, Z. Yang, and M. Kitsuregawa, “Recommending related microblogs: A comparison between topic and wordnet based approaches,” in *AAAI*. AAAI Press, 2012.

[2] X. Li, G. Xu, E. Chen, and L. Li, “Learning user preferences across multiple aspects for merchant recommendation,” in *ICDM*. IEEE Computer Society, 2015, pp. 865–870.

[3] Y. Chen, X. Li, L. Li, G. Liu, and G. Xu, “Modeling user mobility via user psychological and geographical behaviors towards point of-interest recommendation,” in *DASFAA*, ser. Lecture Notes in Computer Science, vol. 9642. Springer, 2016, pp. 364–380.

[4] J. Chen, H. Li, Q. Xie, L. Li, and Y. Liu, “Streaming recommendation algorithm with user interest drift analysis,” in *APWeb/WAIM*, ser. Lecture Notes in Computer Science, vol. 11642. Springer, 2019, pp. 121–136.

[5] Q. Xie, F. Xiong, T. Han, Y. Liu, L. Li, and Z. Bao, “Interactive resource recommendation algorithm based on tag information,” *World Wide Web*, vol. 21, no. 6, pp. 1655–1673, 2018.

[6] Y. Wang, Q. Xie, L. Li, and Y. Liu, “An empirical study on effect of semantic measures in cross-domain recommender system in user cold-start scenario,” in *KSEM*, ser. Lecture Notes in Computer Science, vol. 12816. Springer, 2021, pp. 264–278.

[7] P. Wang, L. Li, R. Wang, G. Xu, and J. Zhang, “Socially-driven multi-interaction attentive group representation learning for group recommendation,” *Pattern Recognit. Lett.*, vol. 145, pp. 74–80, 2021.

[8] Q. Xie, Y. Zhu, F. Xiong, L. Li, Z. Bao, and Y. Liu, “Interactive resource recommendation with optimization by tag association and significance analysis,” *Neurocomputing*, vol. 391, pp. 210–219, 2020.

[9] S. Liu, Y. Liu, and Q. Xie, “Personalized resource recommendation based on regular tag and user operation,” in *APWeb*, ser. Lecture Notes in Computer Science, vol. 9932. Springer, 2016, pp. 111–123.

[10] Z. Xie, L. Liu, Y. Wu, L. Li, and L. Zhong, “Learning tfidf enhanced joint embedding for recipe-image cross-modal retrieval service,” *IEEE Trans. Serv. Comput.*, p. Accepted, 2021.

[11] Z. Xie, L. Li, X. Zhong, L. Zhong, and J. Xiang, “Image-to-video person re-identification with cross-modal embeddings,” *Pattern Recognit. Lett.*, vol. 133, pp. 70–76, 2020.

[12] Z. Zan, L. Li, J. Liu, and D. Zhou, "Sentence-based and noise-robust cross-modal retrieval on cooking recipes and food images," in *ICMR*. ACM, 2020, pp. 117–125.
[13] Z. Xie, L. Liu, L. Li, and L. Zhong, "Multi-subspace implicit alignment for cross-modal retrieval on cooking recipes and food images," in *CIKM*. ACM, 2021, p. Accepted.
[14] L. Li, M. Li, Z. Zan, Q. Xie, and J. Liu, "Multi-subspace implicit alignment for cross-modal retrieval on cooking recipes and food images," in *CIKM*. ACM, 2021, p. Accepted.
[15] H. Dai, Q. Xie, Y. Ma, Y. Liu, and S. Xiong, "Rgb-infrared person re-identification via image modality conversion," in *ICPR*. IEEE, 2020, pp. 592–598.
[16] Y. Ma, Y. Liu, Q. Xie, S. Xiong, L. Bai, and A. Hu, "A tibetan thangka data set and relative tasks," *Image Vis. Comput.*, vol. 108, p. 104125, 2021.
[17] Y. Ma, Y. Liu, Q. Xie, and L. Li, "Cnn-feature based automatic image annotation method," *Multim. Tools Appl.*, vol. 78, no. 3, pp. 3767–3780, 2019.
[18] Y. Ma, Q. Xie, Y. Liu, and S. Xiong, "A weighted knn-based automatic image annotation method," *Neural Comput. Appl.*, vol. 32, no. 11, pp. 6559–6570, 2020.
[19] K. Hu, L. Li, J. Liu, and D. Sun, "Duronet: A dual-robust enhanced spatial-temporal learning network for urban crime prediction," *ACM Trans. Internet Techn.*, vol. 21, no. 1, pp. 24:1–24:24, 2021.
[20] Y. Chu, L. Li, Q. Xie, and G. Xu, "C^2-guard: A cross-correlation gaining framework for urban air quality prediction," in *PAKDD*, ser. Lecture Notes in Computer Science, vol. 12712. Springer, 2021, pp. 779–790.
[21] K. Hu, L. Li, Q. Xie, J. Liu, and X. Tao, "What is next when sequential prediction meets implicitly hard interaction?" in *CIKM*. ACM, 2021, p. Accepted.
[22] Y. Han, L. Li, and J. Zhang, "A coordinated representation learning enhanced multimodal machine translation approach with multi-attention," in *ICMR*. ACM, 2020, pp. 571–577.
[23] L. Li, T. Tayir, K. Hu, and D. Zhou, "Multi-modal and multi-perspective machine translation by collecting diverse alignments," in *PRICAI*, 2021, p. Accepted.
[24] L. Li, T. Tayir, and K. Hu, "Multimodal machine translation enhancement by fusing multimodal-attention and fine-grained image features," in *MIPR*, 2021, p. Accepted.
[25] L. Li, M. Kong, D. Li, and D. Zhou, "A multi-granularity semantic space learning approach for cross-lingual open domain question answering," *World Wide Web*, vol. 24, no. 4, pp. 1065–1088, 2021.
[26] ——, "A cross-layer connection based approach for cross-lingual open question answering," in *NLPCC*, vol. 12430. Springer, 2020, pp. 470–481.
[27] M. Chen, L. Li, and Q. Xie, "Translation language model enhancement for community question retrieval using user adoption answer," in *APWeb/WAIM*, ser. Lecture Notes in Computer Science, vol. 10366. Springer, 2017, pp. 251–265.
[28] L. Li, M. Zhang, Z. Chao, and J. Xiang, "Using context information to enhance simple question answering," *World Wide Web*, vol. 24, no. 1, pp. 249–277, 2021.
[29] X. Zhong, J. Liu, L. Li, S. Chen, W. Lu, Y. Dong, B. Wu, and L. Zhong, "An emotion classification algorithm based on spt-capsnet," *Neural Comput. Appl.*, vol. 32, no. 7, pp. 1823–1837, 2020.
[30] L. Li, K. Hu, Y. Zheng, J. Liu, and K. A. Lee, "Coopnet: Multi-modal cooperative gender prediction in social media user profiling," in *ICASSP*. IEEE, 2021, pp. 4310–4314.
[31] Y. Zheng, L. Li, J. Zhang, Q. Xie, and L. Zhong, "Using sentiment representation learning to enhance gender classification for user profiling," in *APWeb/WAIM*, ser. Lecture Notes in Computer Science, vol. 11642. Springer, 2019, pp. 3–11.
[32] W. Duan, L. Li, and Y. Yu, "A relation learning hierarchical framework for multi-label charge prediction," in *PAKDD*, ser. Lecture Notes in Computer Science, vol. 12085. Springer, 2020, pp. 729–741.
[33] P. Nai, L. Li, and X. Tao, "A densely connected encoder stack approach for multi-type legal machine reading comprehension," in *WISE*, ser. Lecture Notes in Computer Science, vol. 12343. Springer, 2020, pp. 167–181.
[34] L. Li, L. Zhao, P. Nai, and X. Tao, "Charge prediction modeling with interpretation enhancement driven by double-layer criminal system," *World Wide Web*, pp. 1–20, 2021.
[35] J. Dong, J. Yuan, L. Li, and X. Zhong, "A lightweight high-resolution representation backbone for real-time keypoint-based object detection," in *ICME*. IEEE, 2020, pp. 1–6.
[36] J. Dong, J. Yuan, L. Li, X. Zhong, and W. Liu, "Optimizing queries over video via lightweight keypoint-based object detection," in *ICMR*. ACM, 2020, pp. 548–554.
[37] J. Yuan, M. Chen, T. Jiang, and T. Li, "Complete tolerance relation based parallel filling for incomplete energy big data," *Knowl. Based Syst.*, vol. 132, pp. 215–225, 2017.
[38] M. Chen, J. Yuan, L. Li, D. Liu, and Y. He, "Heuristic attribute reduction and resource-saving algorithm for energy data of data centers," *Knowl. Inf. Syst.*, vol. 61, no. 1, pp. 277–299, 2019.
[39] C. Du, J. Yuan, J. Dong, L. Li, M. Chen, and T. Li, "Gpu based parallel optimization for real time panoramic video stitching," *Pattern Recognition Letters*, vol. 133, pp. 62–69, 2020.
[40] R. Wang, L. Li, P. Wang, X. Tao, and P. Liu, "Feature-aware unsupervised learning with joint variational attention and automatic clustering," in *ICPR*. IEEE, 2020, pp. 923–930.
[41] R. Wang, L. Li, X. Tao, X. Dong, P. Wang, and P. Liu, "Trio-based collaborative multi-view graph clustering with multiple constraints," *Inf. Process. Manag.*, vol. 58, no. 3, p. 102466, 2021.

When AI Meets the Internet of Things

A BRIEF INTRODUCTION TO CISCO-LA TROBE CENTRE FOR AI AND IOT

Professor Wei Xiang

Cisco Research Chair of AI and IoT, Director of Cisco-La Trobe Centre for AI and IoT

La Trobe University, Melbourne, Australia

E-mail: w.xiang@latrob.edu.au

***Abstract*—The Cisco-La Trobe Centre for AI and Internet of Things (IoT) based at La Trobe University is Australia's only industry-sponsored research centre which specializes in combining the superpowers of AI and IoT technologies. The Cisco AIoT Centre was founded by Professor Wei Xiang, Cisco Research Chair of AI and IoT based at La Trobe University, and is technically sponsored by Cisco. The Centre currently has 15 affiliated academic staff and ~20 research students. Since its establishment in October 2020, the Centre has played instrumental roles in securing $12M external research funding from various government and industry sources. Through working closely with Cisco and its clients and ecosystem partners, the Centre has established both academic credentials and extensive industry linkages in priority areas of critical supply chain, sustainability, digital agriculture and digital health.**

***Index Terms*—Artificial Intelligence, Internet of Things**

I. INTRODUCTION

The Cisco-La Trobe Centre for AI and Internet of Things (IoT) was founded by Professor Wei Xiang in October 2020, who is Cisco Research Chair of AI and IoT at La Trobe University. The Cisco AIoT Centre is unique in the sense that it is Australia's only industry-sponsored research centre that specializes in exploiting the synergy of AI and IoT technologies. The Centre currently has 15 affiliated academic staff and ~20 research students. Since its establishment in late 2020, the Centre has played instrumental roles in securing $12M external research funding from various government and industry sources including the Australian Research Council (ARC). Through working closely with Cisco and its clients and ecosystem partners, the Centre has established both academic credentials and extensive industry linkages in priority areas of critical supply chain, sustainability, digital agriculture and digital health.

The Cisco AIoT Centre prides itself on the so-called "two-pillar" research strategy. That is, on one hand, the Centre members conduct cutting-edge and world-leading academic research. This includes publishing in some of the world's best research publications in its chosen fields, e.g., IEEE TPAMI, TNNLS, TIP, TMC, TSP, and JSAC, as well as winning national competitive research grants such as ARC and CRC grants. On the other hand, the Centre works closely with its industry partners such as Cisco and Optus to conduct impactful research projects aiming at industry-driven real-world problems.

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

II. RESEARCH AREAS

The Cisco-La Trobe Centre for AI and IoT focuses on combining AI and IoT technologies, as well as exploiting the synergy between the two powerful technologies. The Center also develops advanced AI and IoT solutions to a few selected vertical application domains including critical supply chains, digital health, digital agriculture, and environmental sustainability.

The key research areas of the Cisco AIoT Centre are detailed as follows.

A. Machine-to-machine and IoT Communications

Ubiquitous connectivity is crucial for IoT sensors. In a world with insatiable thirst for digital connectivity and the explosion of machine-to-machine communications, the need for reliable, high-speed and secure communications is crucial to building a connected society. Ubiquitous and energy-efficient M2M and IoT communications face great challenges in the sense that sensors are often deployed in the field and battery replacement is often not an option or incurs significant labor costs. As a result, power-efficient IoT communications are imperative.

Our strengths in this research area include:

1. Low-power wide-area IoT communications
2. LoRaWAN and NB-IoT sensor communications
3. 5G IoT communications
4. Wireless sensor networking
5. Massive IoT networks

See [10, 13, 15, 20] for examples of our research in this area.

B. Advanced 5G, 6G and Satellite Communications

5G communications networks are being rapidly deployed commercially all over the over, which are becoming critical enabling digital infrastructure for a plethora of applications such as Industry 4.0, smart cities, vehicular networking, etc. Meanwhile, the planning and early development of 6G mobile communications systems have already begun. 5G's ultra-broadband, ultra-reliable, and ultra low-latency communications cannot fully satisfy some emerging applications such as holographic video communications, haptic-based telemedicine, Industry 5.0, etc. Compared with 5G, 6G mobile communications not only excel in power consumption, latency, reliability, privacy and security, but also support non-terrestrial networks through integrated satellite-terrestrial communications with the objective of enabling ubiquitous and high-capacity global connectivity.

Our strengths in this research area include:

1. IoT over satellite communications
2. Integrating space and terrestrial networks for 6G
3. NOMA for 5G and 6G communications
4. Integrated sensing and communications for 5G/6G
5. Advanced error control coding for 5G/6G systems

See [14, 17, 19] for examples of our research in this area.

C. AI-driven Wireless Communications

Conventional wireless communications system designs have been guided by classic Shannon information theory. However, this design paradigm is nearly reaching its limits and can no longer satisfy the diverse service requirements of future wireless communications systems in bandwidth, throughput, latency, reliability, quality of experience (QoE), etc. Past and present wireless communications systems have been governed by mathematical models and mostly derived from classical communication theories. However, these traditional design techniques are unlikely to meet the QoS and QoE requirements imposed by next-generation wireless communications systems. On the other hand, the advent of AI and machine learning technique is heralding a paradigm shift from traditional communication theory oriented designs to AI-driven designs. It has been proven that AI techniques can achieve superior performances for wireless communications applications attributed to their exceptional learning and optimization capabilities for complex and dynamic communications scenarios.

Our strengths in this research area include:

1. Distributed and federated learning for communications and networking
2. Deep reinforcement learning for computation offloading and resource allocation
3. AI/ML approaches for SCMA and NOMA communications systems
4. AI/ML approaches for spatial modulation

See [3, 4, 7] for examples of our research in this area.

D. AI for IoT & Sensor Data Analytics

With the rise of the Internet of Things (IoT), industries are awash in massive amounts of big data from an increasing array of wired and wireless sensors. These sensors often form networks and continuously monitor and report on essential information like heart rates of a patient, fuel load of forest, etc. It should be taken note that original sensor data are 'useless', unless we turn them into information, insights, and knowledge using advanced data analytics techniques such as machine learning. Besides, raw sensor data straight out IoT sensors contain large-scale 'unclean' data, which need to undergo data cleaning processes before data analysis can be performed to useful information from cleaned IoT sensor data. Another important branch of data analytics is predictive analytics, which aims at making predictions about future outcomes based on historical data and analytics techniques such as statistical modeling and machine learning. With the aid of sophisticated predictive analytics tools, individuals and organizations alike can use past and present data to reliably forecast trends and behaviors in the future.

Our strengths in this research area include:

1. IoT sensor data imputation
2. Time series data analytics
3. Marine big data analytics
4. Predictive analysis for asset maintenance
5. Physical-informed neural networks

See [5, 8, 9, 12, 18] for examples of our research in this area.

E. Light Field Photography Based Computer Vision

Light field imaging and photography has emerged as a promising technology for capturing richer visual information from our world. As opposed to traditional photography, which captures a 2D projection of the light in the scene integrating the angular domain, a light field collects radiance from rays in all directions, demultiplexing the angular information lost in conventional photography. This higher-dimensional representation of visual data offers powerful capabilities for scene understanding, and substantially improves the performance of traditional computer vision problems such as depth sensing, post-capture refocusing, segmentation, etc. However, the high-dimensionality of light fields also brings up new challenges in light field data capturing, processing, compression, and display. Consequently, light field image and video processing has become increasingly popular in the field of computer vision and computer graphics.

Our strengths in this research area include:

1. Light field multi-view image depth estimation
2. Light field video streaming
3. Light field multi-view video coding
4. Light field based 3D telemedicine

See [1, 2] for examples of our research in this area.

F. IoT Security and Privacy

IoT devices and the data they collect can provide convenience, efficiency and insights into essentially every aspect of our world. Although the IoT brought huge benefits, privacy and security are among the significant challenges of the Internet of Things. Users need to trust IoT devices and related services are secure. Moreover, IoT safety must be considered to prevent the IoT system and its components from causing an unacceptable risk of injury or physical damage and at the same time considering social behaviour and ethical use of IoT technologies to enable effective security and safety. Unlike security and privacy issues in the conventional cyber space, the low capabilities of IoT devices in terms of their energy and computing capabilities, the unreliable nature of the wireless channel, and physical vulnerability are among the contributing factors to some unique security vulnerabilities in IoT. These factors make security and privacy for IoT more challenging than their conventional cyber space counterpart.

Our strengths in this research area include:

1. Physical-layer security for wireless IoT
2. Light weight IoT security algorithms
3. Secure machine-to-machine communications
4. AI and machine learning for IoT security

See [6] for examples of our research in this area.

G. Digital Twins for Industry 4.0

A digital twin is a digital representation of a physical object or system that spans its lifecycle, is updated from real-time data, and uses simulation, machine learning and reasoning to help decision-making. This amounts to creating a highly complex virtual model that is the exact counterpart (or twin) of a physical thing. Digital twins offer a real-time look at what's happening with physical assets, which can radically reduce development costs and alleviate maintenance burdens. Digital twins can be used to predict different outcomes based on variable data. With additional software and data analytics, digital twins can often optimize an IoT deployment for maximum efficiency, as well as help designers figure out where things should go or how they operate before they are physically deployed. Digital twins fuse many emerging technologies such as IoT, wireless communications, artificial intelligence, big data analytics and data visualization in the form of virtual and augmented reality. It is recognized as one of the pillars of Industry 4.0, and has found widespread applications in areas such as smart manufacturing, intelligent transport, and digital healthcare.

Our strengths in this research area include:

1. IoT for digital twins
2. Machine learning for digital twins
3. Digital twins for smart manufacturing

See [11, 16] for examples of our research in this area.

H. Explainable Artificial Intelligence (XAI)

Dramatic success in machine learning has led to a torrent of AI applications. Continued advances in AI techniques promise to produce autonomous systems that will perceive, learn, decide, and act on their own. However, the effectiveness of these systems is limited by the machine's current inability to explain their decisions and actions to human users. Conventional AI autonomous systems act like an opaque "black box", which produces outputs without yielding any decision rationale information. On the contrary, Explainable AI or XAI technology can makes a machine learning system behave like a "white box". This means that an XAI system not only produce more explainable models, while maintaining a high level of learning performance, but also enable human users to understand, appropriately trust, and effectively manage the XAI system. In essence, XAI is a set of processes and methods that allows human users to comprehend and trust the results and output created by machine learning algorithms. EAI helps characterize model accuracy, fairness, transparency and outcomes in AI-powered decision making. AI explainability also helps organizations adopt a responsible approach to AI development.

Our strengths in this research area include:

1. Explainable AI for medical image analysis
2. Explainable AI for time series data analytics
3. Model-dependent and model-agnostic explainability
4. Explainable Transformer networks

III. Research Collaborations with Industry and Real-World Impact

AI and IoT technologies have been seen dramatic success and uptake by nearly every vertical industry ranging from manufacturing to healthcare. One of the two pillars of the Cisco-La Trobe Centre for AI and IoT is to work with industry partners to deliver real-world impactful research.

In the following, we use two example use cases to showcase how the Cisco AIoT Centre works with its industry partners to tackle real-world problems and to deliver impactful research solutions. It is advised that some project details are deliberately concealed due to confidentiality commitments to our industry partners.

A. Digital Twin for Modelling the Molten Oxygen Electrolysis Reactor

This project funded by Janco Enterprise Ptd. Ltd. aims to build a digital twin for modelling the molten oxygen electrolysis reactor. The project objective is to maintain optimum Molten Oxide Electrolysis (MOE) reaction conditions in a stainless steel tube furnace. MOE is a method of extracting metal from their ore bodies using electricity and zero carbon in the reaction.

In the established digital twin, Physics-Informed Neural Networks (PINNs) are employed to approximate the solution of partial differential equations (PDEs) to produce a grid-based solution of Finite Element Analysis (FEA) in a fraction of time it would take matrix math approaches to arrive at a solution. This allows for approximation of systems in real-time with minimal feedback from sensors. Edge AI is used in conjunction with MOE to maintain optimum reaction conditions and to allow for optimum efficiency of the process.

B. Automated Water Channel Inspection Using AI

This project is in collaboration with Southern Rural Water (SRW) to investigate the possibility of utilising an AI system combined with Remotely Piloted Aerial Systems (RPAS) to complete inspections on its channel and drainage network in the Macalister Irrigation District. The Macalister Irrigation District is located approximately 200km south east of Melbourne, in the Gippsland region. In this area, SRW manages some 450+km of open Irrigation Channels, 500+km of open Drains and 50km of pipeline assets. These assets range in size from 2m in width to 10m wide. SRW has an obligation to inspect all its physical assets every 5 years on a rolling basis. Current methods for doing this involve maintenance teams and operators driving the channels and drains from top to bottom, often needing to hop between properties, to assess the condition of each segment and give it a rating. It is also very time consuming and at times dangerous, as operators may need to traverse rough terrain and private land to gain access to the assets.

SRW has engaged the Cisco-LTU AIoT Centre to create an application which analyses RPAS captured video footage, after being trained with SRW data, and detects problems in the channel system dynamically as new videos are fed into it. Future iterations of the application will potentially learn to predict likelihood of occurrences where conditions are met. The application will also produce a report, which captures the

GPS location detailed in the metadata on the video file, the extent and detail of the issue identified (and probability of accuracy) and captures a frame from the video that can be passed on to the maintenance team for human analysis and job planning.

IV. Research Facilities

The Cisco-La Trobe Centre for AI and IoT is located within the Digital Innovation Hub (DIH) at La Trobe University's Melbourne campus. The DIH is a $9M state-of-the-art innovation facility that is funded by the Victoria State Government. Cisco and Optus are two anchor tenants in the DIH, which will house the Cisco Innovation Central @ Melbourne (ICM) and an Optus 5G Lab.

The DIH features state-of-the-art IoT, networking and computer vision equipment donated by Cisco, alongside cutting-edge high performance computing and data storage infrastructure.

References

[1] K. Han, W. Xiang, E. Wang, and T. Huang, "A novel occlusion-aware vote cost for light field depth estimation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, early access via IEEEXplore.

[2] Y. Xu, K. Han, Y. Zhou, J. Wu, X. Xie, and W. Xiang, "Deep adaptive blending network for 3D magnetic resonance image denoising," *IEEE Journal of Biomedical and Health* Informatics, vol. 25, no. 9, pp. 3321-3331, Sep. 2021.

[3] P. Yang, xxx, and W. Xiang, "Transmit antenna selection for full-duplex spatial modulation based on machine learning," *IEEE Transactions on Vehicular Technology*, accepted for publication on 31 August 2021 (IF = 5.978).

[4] S. Gu, W. Lu, W. Xiang, N. Zhang, and Q. Zhang, "Repair delay analysis of mobile storage systems using erasure codes and relay cooperation," IEEE Transactions on Vehicular Technology, early access via IEEEXplore.

[5] J. Shi, Y. Zhao, W. Xiang, V. Monga, X. Liu, and R. Tao, "Deep scattering network with fractional wavelet transform," IEEE Transactions on Signal Processing, vol. 69, pp. 4740-4757, Aug. 2021 (IF = 4.931).

[6] W. Xu, B. Li, F. Zhao, and W. Xiang, "Artificial noise assisted secure transmission for uplink of massive MIMO systems," IEEE Transactions on Vehicular Technology, vol. 70, no. 7, pp. 6750-6762, Jul. 2021 (IF = 5.978).

[7] X. Yao, P. Yang, Z. Liu, M. Xiao, S. Li, and W. Xiang, "A novel hybrid code-domain index modulation," IEEE Communications Letters, accepted for publication on 13 July 2021 (IF = 3.436)

[8] M. Jahanbakht, W. Xiang*, and M. R. Azghadi, "Sea surface temperature forecasting with ensemble of stacked deep neural networks," IEEE Geoscience and Remote Sensing Letters, accepted for publication on 10 July 2021 (IF = 3.966)

[9] S. Baker, W. Xiang, and I. Atkinson, "A hybrid neural network for continuous and non-invasive estimation of blood pressure from raw electrocardiogram and photoplethysmogram waveforms," Computer Methods and Programs in Biomedicine, accepted for publication on 13 May (IF = 3.632).

[10] Y. Zhou, T. Cao, and W. Xiang, "Anypath routing protocol design via Q-learning for underwater sensor networks," IEEE Internet of Things Journal, vol. 8, no. 10, pp. 8173-8190, May 2021 (IF = 9.936).

[11] L. Lei, Y. Tan, G. Dahlenburg, W. Xiang, and K. Zheng, "Dynamic energy dispatch based on deep reinforcement learning in IoT-driven smart isolated microgrids," IEEE Internet of Things Journal, vol. 8, no. 10, pp. 7938-7953, May 2021 (IF = 9.936).

[12] N. Madhukumar, E. Wang, Y. Zhang, and W. Xiang*, "Consensus forecast of rainfall using hybrid climate learning model," IEEE Internet of Things Journal, vol. 8, no. 9, pp. 7270-7278, May 2021 (IF = 9.936).

[13] Y. Li, P. Yang, M. D. Renzo, Y. Xiao, M. Xiao, and W. Xiang, "Precoded optical spatial modulation for indoor visible light communications," IEEE Transactions on Communications, vol. 69, no. 4, pp. 2518-2531, Apr. 2021 (IF = 5.646).

[14] S. Gu, X. Sun, Z. Yang, T. Huang, W. Xiang, and K. Yu, "Energy-aware coded caching strategy design with resource optimization for satellite-UAV-vehicle integrated networks," IEEE Internet of Things Journal, accepted for publication on 12 Mar. 2021 (IF = 9.936).

[15] M. Jahanbakht, W. Xiang*, L. Hanzo, and M. R. Azghadi, "Internet of Underwater Things and big marine data analytics – A comprehensive survey," IEEE Communications Surveys & Tutorials, vol. 23, no. 2, pp. 904-956, Second quarter, 2021 (IF = 25.249).

[16] H. Wen, Y. Du, X. Chen, E. G. Lim, H. Wen, L. Jiang, and W. Xiang, "Deep learning-based multi-step solar forecasting for PV ramp-rate control using sky images," IEEE Transactions on Industrial Informatics, vol. 17, no. 3, pp. 1397-1406, Feb. 2021 (IF = 9.112).

[17] S. Gu, Q. Zhang, and W. Xiang, "Coded storage-and-computation: a new paradigm to enhancing intelligent services in space-air-ground integrated networks," IEEE Wireless Communications, vol. 27, no. 6, pp. 44-51, Dec. 2020 (IF = 11.391).

[18] S. Baker, W. Xiang, and I. Atkinson, "Continuous and automatic mortality risk prediction using vital signs in the intensive care unit: a hybrid neural network approach," Scientific Reports, vol. 10, Article Number 21282, Dec. 2020 (IF = 3.998).

[19] Y. Wang, S. Gu, L. Zhao, N. Zhang, W. Xiang, and Q. Zhang, "Repairable fountain coded storage systems for multi-tier mobile edge caching networks," IEEE Transactions on Network Science and Engineering, vol. 7, no. 4, pp. 2310-2322, Oct.-Dec. 2020 (5.213).

[20] P. Shi, Z. Wang, D. Li, and W. Xiang*, "Zigzag decodable online fountain codes with high intermediate symbol recovery rates," IEEE Transactions on Communications, vol. 68, no. 11, pp. 6629-6641, Nov. 2020 (IF = 5.646).

Text Categorisation on Semantic Analysis for Document Categorisation Using a World Knowledge Ontology

Xiaohui Tao*, Patrick Delaney and Yuefeng Li

E-mail: {Xiaohui.Tao, Patrick.Delaney}@usq.edu.au, y2.li@qut.edu.au

***Abstract*—An effective text categorisation approach can allow users easy access to useful and meaningful textual information. However, while many automatic categorisation techniques have been developed, there is still room for improvement in categorisation performance. In this work, we have proposed an innovative approach using a large world knowledge ontology built from the Library of Congress Subject Headings (LCSH) to categorise text documents. The semantic content of documents is represented by well-defined and well-specified subjects extracted from the ontology. The proposed approach has been successfully evaluated, using a large data set with linguist-generated categorisation results in empirical experiments.**

I. Introduction

An effective categorisation method can improve the efficiency of systems in accessing textual information. In particular, Web personalization systems benefit from categorising a user's local documents (e.g. browsing history, emails, tweets, and blogs) to concepts in a global knowledge base [37], [34]. A user profile is the simulation of the user's concept model, whereas a user's concept model is the user's local reflection of world knowledge with only the topics of interest to the user [35]. User local documents provide wealthy user background knowledge. Therefore, to acquire quality user profiles, user background knowledge needs to be discovered from user local documents and global world knowledge. Figure 1 illustrates a scenario of user profile acquisition, which is completed by discovering user background knowledge from the categorisation of user local documents to subjects in a world knowledge ontology. This method was successfully accomplished and evaluated by Tao et al. [35], using a world ontology constructed from the Library of Congress Subject Headings (LCSH) [1]. Other successfully accomplished models include Sieg et al. [32] using the Open Directory Project [2].

Effective document categorisation is mostly completed by human effort, with well-trained experts (e.g. linguists, librarians, and metadata experts) manually categorising documents in either traditional or digital forms and assigning descriptors (a list of subjects) to documents [9]. However, manual categorisation is expensive and time-consuming. Additionally, manual categorisation becomes problematic when dealing with large repositories. Many automatic methods have been developed to categorise documents based on semantic contents (e.g. [7], [1],

[1]http://id.loc.gov/authorities/
[2]http://www.dmoz.org/

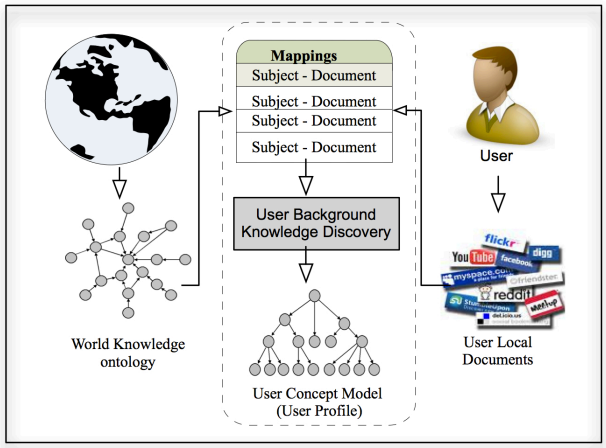


Fig. 1. Application of Document Categorisation exploiting World Knowledge

[4], [21], [26], [38], [39]). However, in pursuing efficient, automatic categorisation, two problems were revealed: (i) knowledge bases chosen for categorisation are usually inadequate for describing the real world, being either constructed in over-simplified structures or covering only a limited range of topics, resulting in inadequate subjects being assigned to documents for categorisation; and (ii) imperfect categorisation algorithms still have large room for improvement. Inadequate subjects have been used to categorise documents because imperfect algorithms were used. Kasper et al. [18] reviewed implicit, explicit and hybrid user acquisition frameworks, noting that there is a clear need to investigate the potential of learning algorithms. Refining user acquisition models against these identified problems will improve semantic content based document categorisation and add to research on user background knowledge discovery.

In this paper, we propose an automatic semantic categorisation approach using the LCSH ontology, a mature and well-defined world knowledge ontology previously evaluated by Tao et al. [35] successfully. Given a document, its semantic features are first discovered. The document's relevant subjects are then extracted from the world ontology based on the discovered features. Finally, these subjects are generalised in order to assign a list of competent subjects to the document for categorisation. An empirical experiment evaluated the proposed approach by comparing it against typical categorisation methods including *Rocchio* and kNN, based on the ground

truth of manual categorisation results. A large corpus was collected from the real world for the experiments. The evaluation result was promising and encouraging. The proposed approach makes the following contributions to research and practice:

- A method that categorises documents into multiple categories based on semantic analysis of content;
- An innovative algorithm that generalises subjects based on their semantic relationships;
- A novel approach using a large world knowledge ontology to guide semantic categorisation of documents.

Semantic categorisation may also help capture users' demands and opinions in e-Commerce, as well as the rivals' intelligence, benefit from the improved efficient access of public text documents.

The paper is organised as follows: Section II discusses the related work; Section III formalises the definitions and the research problem in this work. After that, Section IV introduces the proposed semantic categorisation method. The experiment design is described in Section VI and the results are discussed in Section VII. Finally, Section VIII makes the conclusions.

II. Related Work

The semantic content of text documents has different representations, such as lexicons, categories, or patterns. A lexicon-based representation of documents is easily understood by users and computational systems. Text documents are represented by a set of descriptors chosen from controlled vocabularies defined in terminological ontologies, thesauruses, or dictionaries. However, when extracting lexical descriptors, some noisy descriptors are also extracted alongside meaningful, representative descriptors, due to the term ambiguity problem. The development of terminological ontologies, thesauruses, or dictionaries is also financially expensive and time-consuming, due to the large requirement of human effort. As a result, the lexicon-based representation of semantic content is inefficient.

Categorisations are widely used in methods to represent document contents, like those in [33], [29], [14], [35]. In this approach, the concepts revealed from text are represented by categories and organised in a tree or graphic structure. The relationships existing between concept nodes in the structure are explored in order to measure the competency of a concept describing or representing the content. However, usually simple relations (subsumption of one containing another or super- and sub-class) are used in categorisations rather than detailed, well-specified semantic relations (like is-a, part-of, and related-to). Thus, categorisation-based representation needs to improve for more detailed and precise levels of concept specification.

Pattern-based representation uses multiple phrases to represent document content [12], [10], [22], [24]. However, pattern-based categorisation suffers from issues caused by the length of patterns. Concepts are specific and discriminating only with substantially long patterns, but long patterns have low frequency. Consequently, the power of long patterns reduces because low frequency makes the patterns less applicable to problems [23]. In addition, because of the text-mining techniques used for pattern discovery, sometimes noisy patterns are extracted alongside useful patterns. Alternative weighting methods need to be investigated to overcome this problem in pattern-based content representation.

Many works utilise pattern-mining techniques to help build classification models, which is similar as the strategy employed in our work. Malik and Kender [28] proposed the "Democratic Classifier", a pattern-based classification algorithm using short patterns. However, the democratic classifier relies on the quality of training samples and cannot deal with the "no training set available" problem. Bekkerman and Matan [3] argued that most of the information on documents can be captured in phrases, and they proposed a text classification method that employs lazy learning from labelled phrases. The phrases in their work are in fact a special form of sequential patterns that are used in our work for feature extraction of documents.

Text classification is a common technique used to classify a stream of documents into categories by using the classifiers learned from the training samples [25]. This can be two types: *kernel-based* and *instance-based* [2]. Typical kernel-based classifier learning approaches include *Support Vector Machines* (SVMs) [17] and regression models [31]. Kernel-based approaches sometimes incorrectly classify negative samples into positive. Typical instance-based classification approaches include the kNN and its variants, which do not rely upon the statistical distribution of training samples. However, the instance-based approaches become unstable when classifying highly accurate positive samples from an unlabelled data set. Other reports, such as [30], have a different view and categorise text classification techniques into *document representations based classifiers* including SVMs and kNN and *word probabilities based classifiers* including Naive Bayesian, decision trees [17] and neural networks [44]. These classification techniques have different strengths and weaknesses, and should be chosen carefully depending on the problem space.

Unsupervised text classification aims to classify documents into classes that are absent of any labelled training documents. Many successful models have been proposed, such as [43]. However, on many occasions, the target classes may not have any labelled training documents available. One particular example is the "cold start" problem in recommender systems and social tagging [13]. Unsupervised classification can automatically learn an annotation model to make recommendations or label the tags when the products or tags are rare and have no useful associated information. Without associated training samples, Yang et al. [42] built a classification model for a target class by analysing the correlating auxiliary classes. The work in this paper is similar to that model, however, our model differs by exploiting a hierarchical world knowledge ontology for classification, instead of only auxiliary classes. Also exploiting a world knowledge base, Yan et al. [40] examined unsupervised relation extraction from Wikipedia articles and integrated linguistic analysis with web frequency information to improve unsupervised classification performance. By comparison, our work aims to exploit a world knowledge ontology to help unsupervised classification. Cai et al. [6] and Houle and Grira [16] proposed unsupervised approaches to evaluate and improve the quality of selecting features. Given

a set of data, their approach finds a subset containing the most informative, discriminative features. Though the work presented in this paper also relies on features selected from documents, the features are further investigated with their referring-to ontological concepts to improve the performance of classification.

Ontologies have been used to facilitate text classification by generating features using domain-specific and common-sense knowledge in large ontologies [11] and semantic relations in web personalization [34] and document retrieval [27]. Camous et al. [8] introduced a domain-independent method that uses the Medical Subject Headings (MeSH) ontology. The method observes the inter-concept relationships and represents documents by MeSH subjects, considering semantic relations. Another world ontology commonly used in text classification is Wikipedia [1]. For instance, Hu et al. [14] derived background knowledge from Wikipedia to represent documents and attempted to deal with the sparsity and high dimensionality problems in text classification. Compared to this prior research, our work uses the LCSH, a superior world knowledge ontology under continuous development for a hundred years by knowledge engineers.

Text classification models were originally designed to handle only single-label problems, where each document is classified into only one class. However, in many circumstances single-label text classification is inadequate, such as with social networks where multiple labels are needed [15], [20]. Similar to the work of Yang et al. [41], our method also targets multi-label text classification. However, rather than adopting active learning algorithms for multi-label classification, we exploit concepts and their structure in world knowledge ontologies [19].

III. Research Problem and Definitions

Let $\mathcal{D} = \{d_i \in \mathbb{D}, i = 1, \ldots, m\}$ be a set of text documents; $\mathcal{S} = \{s_1, \ldots, s_K\}$ be a large set of classes, where K is the number of classes. If there is an available training set $\mathcal{D}_t = \{d_j \in \mathbb{D}, j = m+1, \ldots, n\}$ with $y_j^k = \{0, 1\}, k = 1, \ldots, K$ provided for describing the likelihood of d_j belonging to class s_k, it is easy to learn a binary prediction function $p(y^k|d)$ and use it to classify $d_i \in \mathcal{D}$. However, our objective is to learn a prediction function $p(y^k|d)$ to classify d_i into $\{s_k\} \subset \mathcal{S}$ without $\mathcal{D}_t$ available. We refer to this problem as *unsupervised multi-label text classification*.

***Definition** 1:* Let $\Omega = \{d_1, d_2, d_3, \ldots, d_n\}$ be a finite and non-empty set of text documents. Given $d \in \Omega$, its semantic content can be categorised by using the mapping:

$$\eta : \Omega \to 2^{\mathcal{S}}, \quad \eta(d) = \{s \in \mathcal{S} | str(d, s) \geq min\_str\} \subseteq \mathbb{S}$$

and its reverse mapping:

$$\eta^{-1} : \mathcal{S} \to 2^{\Omega}, \quad \eta^{-1}(s) = \{d \in \Omega | str(d, s) \geq min\_str\} \square$$

Note that $str(d, s)$ is the strength describing the competency of s to categorise d, and min_str is the threshold defining the desirable competency level.

To illustrate the problem, a sample document is shown in Fig. 2. This screenshot was taken from the online catalogue of the University of Melbourne Library[3]. The catalogue information is about a book with the title and summarised content:

> *Economic espionage and industrial spying. Dimensions of economic espionage and the criminalization of trade secret theft – Transition to an information society - increasing interconnections and interdependence – International dimensions of business and commerce – Competitiveness and legal collection versus espionage and economic crime – Tensions between security and openness – The new rule for keeping secrets - the Economic Espionage Act – Multinational conspiracy or natural evolution of market economy.*

and a list of librarian manually-assigned subjects:

> *Business intelligence; Trade secrets; Computer crimes; Intellectual property; Commercial crimes.*

The title, summarised content, and subjects in Fig. 2 depict the ultimate goal we pursue: given a text document (e.g., the title and summarised content in Fig. 2), categorise it to an indexed set of subjects extracted from the world ontology (e.g., the listed subjects in Fig. 2). Ideally, the extracted subjects should be the same as these linguist manually-assigned subjects, because they represent human intellectual work in semantic categorisation. However, at this stage, attaining the same result as human work is unrealistic. Therefore, finding similar assignment of subjects with human work is the aim of our work. The sample case in Fig. 2 will be used through the rest of the paper to assist the explanation.

The world knowledge ontology is constructed from the Library of Congress Subject Headings (LCSH), a knowledge system developed for organising information in large library collections. It has been under continuous development for over a hundred years to describe and classify human knowledge. Because of the dedicated endeavours of knowledge engineers from generation to generation, the LCSH has become a de facto standard for concept cataloguing and indexing, superior to other knowledge bases. Tao et al. [35] previously compared the LCSH with the Library of Congress Classification, the Dewey Decimal Classification, and Yahoo! categorisation, and reported that the LCSH has broader topic coverage, more meaningful structure, and more accurate semantic relations. The LCSH has been widely used as a means for many knowledge engineering and management works [9]. In this work, the class set $\mathcal{S} = \{s_1, \ldots, s_K\}$ is encoded from the LCSH subject headings.

***Definition** 2: (SUBJECT)* Let $\mathcal{S}$ be the set of subjects, an element $s \in \mathcal{S}$ is a 4-tuple $s := \langle label, neighbour, ancestor, descendant \rangle$, where

- label is a set of sequential terms describing s; $lable(s) = \{t_1, t_2, \ldots, t_n\}$;
- $neighbour$ refers to the set of subjects in the LCSH that directly link to s, $neighbour(s) \subset \mathcal{S}$;

[3]http://cat.lib.unimelb.edu.au/. Note that the screenshot has been altered for display - the alteration was completed without pruning away any meaningful content.

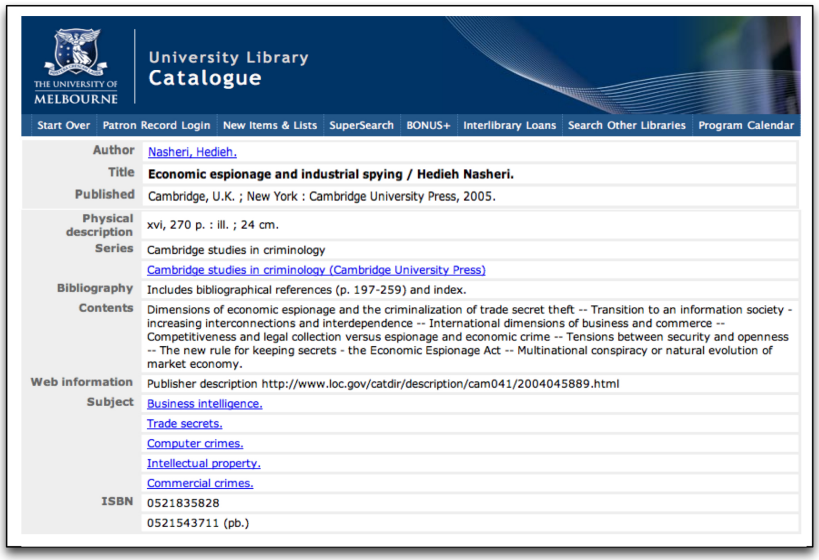


Fig. 2. A Sample Document with Subjects Manually Assigned by Librarians.

- $ancestor$ refers to the set of subjects directly and indirectly link to s and locating at more abstractive level than s in the LCSH, $ancestor(s) \subset \mathcal{S}$;
- $descendant$ refers to the set of subjects directly and indirectly link to s and locating at more specific level than s in the LCSH, $descendant(s) \subset \mathcal{S}$. □

The semantic relationships of subjects are encoded from the references defined in the LCSH for subject headings, including *Broader Term*, *Used for*, and *Related to*. The $ancestor(s)$ in Definition 2 returns the *Broader Term* subjects of s; the $descendant(s)$ is the reversed function of $ancestor(s)$, with additional subjects *Used for* s; the $neighbour(s)$ returns the subjects *Related to* s.

With Definition 2, the world knowledge ontology is defined:

***Definition** 3: (ONTOLOGY)* Let $\mathcal{O}$ be a world ontology. $\mathcal{O}$ contains a set of subjects linked by their semantic relations in a hierarchical structure. $\mathcal{O}$ is a 3-tuple $\mathcal{O} := \langle \mathcal{S}, \mathcal{R}, \mathcal{H}_{\mathcal{R}}^{\mathcal{S}} \rangle$, where

- $\mathcal{S}$ is the set of subjects defined in Definition 2;
- $\mathcal{R}$ is the set of relations linking any pair of subjects;
- $\mathcal{H}_{\mathcal{R}}^{\mathcal{S}}$ is the hierarchical structure of $\mathcal{O}$ constructed by $\mathcal{S} \times \mathcal{R}$. □

IV. Theoretical Framework

A lexicon-based representation is based on the statistic of occurring terms. Such a representation is easy to understand by users and systems. However, along with meaningful, representative features, some noisy terms are also extracted, caused by sense ambiguity of terms. To deal with this problem, pattern-based representation is studied, which uses frequent sequential patterns (phrases) to represent document contents [24]. The pattern-based representation is superior to lexicon-based, as the context of terms co-occurred in phrases is considered. However, the pattern-based presentation suffers from a limitation caused by the length of patterns. Though a long pattern is wealthy with information and so more discriminative, it usually has low frequency and as a result, becomes inapplicable. To overcome the problem, we represent the content of documents by a set of weighted closed frequent sequential patterns discovered by pattern mining techniques.

***Definition** 4: (FEATURES)* Given a document $d = \{t_1, t_2, \ldots, t_n\}$ as a sequential set of repeatable terms, the feature set, denoted as $\mathcal{F}(d)$, is a set of weighted phrase patterns, $\{\langle p, w(p) \rangle\}$, extracted from d that satisfies the following constraints:

- $\forall p \in \mathcal{F}(d), p \subseteq d$.
- $\forall p_1, p_2 \in \mathcal{F}(d)(p_1 \neq p_2), p_1 \not\subset p_2 \wedge p_2 \not\subset p_1$.
- $\forall p \in \mathcal{F}(d), w(p) \geqslant \vartheta$, a threshold. □

The initial classification of d to $s_k \in \mathcal{S}$ is done through accessing a term-subject matrix created by the subjects and their labels. Adopting the features discovered previously, we use a feature-subject mapping approach to initially assign subject classes to the document.

***Definition** 5: (TERM-SUBJECT MATRIX)* Let $\mathcal{T}$ be the term space of $\mathcal{S}, \mathcal{T} = \{t \in \bigcup_{s \in \mathcal{S}} label(s)\}, \langle \mathcal{S}, \mathcal{T} \rangle$ is the matrix coordinated by $\mathcal{T}$ and $\mathcal{S}$, where a mapping exists:

$$\mu : \mathcal{T} \rightarrow 2^{\mathcal{S}}, \quad \mu(t) = \{s \in \mathcal{S} | t \in label(s)\}$$

and its reverse mapping also exists:

$$\mu^{-1} : \mathcal{S} \rightarrow 2^{\mathcal{T}}, \quad \mu^{-1}(s) = \{t \in \mathcal{T} | s \in \mu(t)\} \qquad \square$$

Adopting Definition 4 and 7, we can initially classify $d_i \in \mathcal{D}$ into a set of subjects using the following prediction:

$$\widehat{y}_i^k = I(s_k \in h \circ g \circ f(d_i)), i = 1, \ldots, m \qquad (1)$$

where $I(z)$ is an indicator function that outputs 1 if z is true and zero, otherwise; $f(d) = \{p | \langle p, w(p) \rangle \in F(d)\}$; $g(\rho) = \{t \in \cup_{p \in \rho} p\}$; $h(\tau) = \{s \in \cup_{t \in \tau} \mu(t)\}$.

The initial classification process easily generates noisy subjects because of direct feature-subject mapping. Against the problem, we introduce a method to generalise the initial subjects to optimise the classification. We observed that in initial classification some subjects extracted from the ontology are overlapping in their semantic space. Thus, we can optimise the classification result by keeping only the dominating subjects and pruning away those being dominated. This can be

done by investigating the semantic relations existing between subjects. Let s_1 and s_2 be two subjects and $s_1 \in ancestor(s_2)$ ($s_2 \in descendant(s_1)$). s_1 refers to an broader semantic space than s_2 and thus, is more general. Vice versa, s_2 is more specific and focused than s_1. Hence, if some subjects are covered by a common ancestor, they can be replaced by the common ancestor without information loss. The common ancestor is unnecessary to be chosen from the initial classification result, as choosing an external common ancestor also satisfies the above rule. After generalising the initial classification result, we have a smaller set of subject classes, with no information lost but some focus. (The handling of focus problem is presented in next section.)

***Definition** 6: (GENERALISED CLASSIFICATION)* Given a document d and its initial classification result, a subject set denoted by $S^I(d)$, the generalised classification result, denoted as $S^G(d)$, is the set of subjects satisfying:

1) $\forall s \in S^I(d), \exists s' \in S^G(d), s \neq s', s \in descendants(s')$.
2) $\forall s_1, s_2 \in S^G(d)(s_1 \neq s_2), s_1 \notin descendants(s_2) \wedge s_2 \notin descendants(s_1)$.

V. Framework in Practice

To design a semantic content-based document categorisation approach, two critical difficulties must be addressed: choosing a competent knowledge base, and proposing a categorising algorithm with less imperfection. This work was designed to address these two difficulties. A world knowledge ontology constructed from the LCSH is utilised to work as the knowledge base for the semantic content based categorisation. Documents are categorised to the subjects in the LCSH ontology through three steps: discovering features from the documents; extracting subjects from the LSCH ontology based on the features; generalising the subjects to finalise categorisation. The conceptual framework for the design is illustrated in Fig. 3, which consists of three modules, each one designed for one step.

Feature Discovery Module. Pattern Taxonomy Method has been employed in this module to discover features from the given document, based on the theory of closed frequent sequential patterns. As the outcome of this module, a set of patterns with weights greater than a minimum value is selected to represent the features of the document;

Knowledge Extraction Module. A term-subject matrix has been established in this module to extract appropriate subjects from the LCSH world ontology, based on the features extracted in previous step. The matrix has two attributes: joint set of terms from the label of all subjects; the set of all subjects in the world ontology. Given a set of patterns (features), a mapping set of subjects is extracted, in which each element is assigned with a strength value representing its level of competency to categorise the document;

Knowledge Generalisation Module. The subjects extracted in previous step are investigated in this module for their semantic relations with other subjects in the neighbourhood and their location in the structure of the world ontology. The subjects referring to common semantic space are merged and replaced by their common ancestor subject. Finally, a refined indexed list of subjects are generalised to represent the semantic content and to categorise the document.

The LCSH world ontology and proposed semantic categorisation approach is be explained in the following sections.

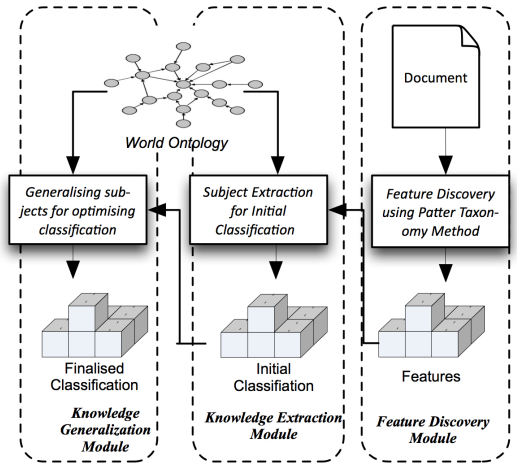


Fig. 3. Conceptual Framework

A. The LCSH World Ontology

Textual information has some properties that make semantic categorisation difficult. The structure and format of text documents are usually complex and the topics are heterogeneous, meaning the content may change constantly [36]. An efficient text document categorisation method must deal with these properties. As shown in many previous works like [46], [35] and [32], an effective strategy is using world knowledge ontologies. Ontologies are formal descriptions and specifications of conceptualisation. By nature, ontologies are a powerful technique for clarifying and then solving complex, heterogeneous problems. World knowledge is commonsense knowledge possessed by people and acquired through their experiences and education [45]. To categorise text documents with constant changes, world knowledge provides constant support because it updates alongside the progress of civilisation. The ontology (or any knowledge base) chosen to guide efficient, automatic text categorisation should be competent to deal with these properties.

The world knowledge ontology in this work is constructed based on the LCSH, similar to the work of [35]. The LCSH was developed for organising and retrieving information from a large volume of library collections. As discussed by Chan [9], the LCSH has many superiorities for handling the problems in text categorisation:

- The LCSH system is an ideal world knowledge base covering an exhaustive range of topics (Competent to deal with the complexity and heterogeneity problems);
- The LCSH represents the natural growth and distribution of human intellectual work. For over a hundred years, the knowledge contained in the LCSH has undergone continuous revision and enrichment (Competent to deal with the constant change problem);

TABLE I
COMPARISON OF DIFFERENT WORLD TAXONOMIES

	LCSH	LCC	DDC	Yahoo!
# of topics	394,070	4,214	18,462	100,000
Structure	Directed Acyclic Graph	Tree	Tree	Directed Acyclic Graph
Depth	37	7	23	10
Semantic Relations	Broader, Used-for, Related-to	Super- and Sub-class	Super- and Sub-class	Super- and Sub-class

- The LCSH has the most comprehensive non-specialised controlled vocabulary in English (Providing competent subjects to categorise documents.)

Though the majority of libraries utilising the LCSH are located in the United States, almost all libraries around the world have their systems convertible to the LCSH. The LCSH system is also superior to other world knowledge taxonomies. Table I presents a comparison of the LCSH with the Library of Congress Classification (LCC), the Dewey Decimal Classification (DDC), and the Yahoo! categorisation (YC). The LCSH has the largest number of topics, and the most specific semantic relations and structure. LCSH descriptors are classified by professionals, and the classification quality is guaranteed by well-defined and continuously-refined cataloging rules - in many respects, the LCSH has become a de facto standard for subject cataloging and indexing [9]. A world ontology constructed from the LCSH has also been proven promising by Tao et al. [35], for the problem of user background knowledge discovery from user local text documents. In summary, the LCSH is an ideal, competent world knowledge ontology for semantic categorisation of text documents.

The concepts in the world ontology are called *subjects* that are encoded from subject headings in the LCSH authorities. The semantic relationships of subjects are encoded from the references defined in the LCSH authorities for subject headings, such as *Broader Term*, *Used for*, and *Related to*. The $ancestor(s)$ function in Definition 2 returns the *Broader Term* subjects of s (they are semantically broader and thus, more general than s); the $descendant(s)$ returns the subjects that are *Used for* s and the subjects for which s is their *Broader Term* (s is semantically broader and thus, more specific than these subjects); the $neighbour(s)$ returns the *Related to* subjects of s.

B. Feature Discovery from Text

Given a document $d = \{t_1, t_2, \ldots, t_m\}$, let $\mathcal{F}(d) = \{\langle p, w(p)\rangle\}$, a set of competent patterns with weights, be the feature set of d. $\mathcal{F}(d)$ is to be discovered using the closed frequent sequential pattern mining technique.

We first introduce the concept of *sequential patterns*. A sequential pattern $p = \langle t_1, \ldots, t_r\rangle$ is an ordered list of terms. Given two sequential patterns p_1 and p_2, if p_1 is a subsequence of p_2, we say p_1 is a sub-pattern of p_2, and p_2 a super-pattern of p_1.

A pattern's *frequent* level depends on its occurrence frequency in the document. Let $P(d)$ be the set of all n-gram ($0 < n <= |d|$) patterns that can be extracted from d; $termset(p)$ be a function that returns the set of terms in a pattern p and $termset(p) \subseteq d$. $coverset(p)$ is the covering set of p for d, and includes all patterns $p' \in P(d)$ satisfying $termset(p) \subseteq termset(p')$; $coverset(p) = \{p' | p' \in P(d), termset(p) \subseteq termset(p')\} \subset P(d)$. The absolute support $sup_a(p)$ is the number of occurrences of p in $P(d)$; $sup_a(p) = |coverset(p)|$. The relative support $sup_r(p)$ is the fraction of the patterns that contain $termset(p)$; $sup_r(p) = \frac{|coverset(p)|}{|P(d)|}$. p is then called *frequent pattern* if its sup_a (or sup_r) $\geq min\_sup$, a minimum support.

We then define the concept of *closed* patterns. Given a set of patterns $P' \subseteq P(d)$, we can also define its *termset* by:

$$termset(P') = \{t | \forall p \in P' \Rightarrow t \in p\} \tag{2}$$

The closure of a pattern p is defined as:

$$Cls(p) = termset(coverset(p)) \tag{3}$$

A pattern p is then called *closed* if and only if $termset(p) = Cls(p)$.

The definition of *closed frequent sequential patterns* relies on a property of closed patterns. Given a closed pattern p, for all patterns $p_1 \supset p$, we have

$$sup_a(p_1) < sup_a(p) \tag{4}$$

A frequent sequential pattern p is called *closed* if there exists no super-pattern p_1 of p such that $sup_a(p_1) = sup_a(p)$.

Based on these definitions, given a d, its feature set $\mathcal{F}(d) = \{\langle p, w(p)\rangle\}$ is discovered, where $w(p)$ is the frequency of p in d. Table II shows the features (closed frequent sequential patterns) discovered from the sample document illustrated in Fig. 2. Note that $min\_sup = 2$, and the feature discovery is based on the text following pre-processing.

TABLE II
FEATURE DISCOVERED FROM THE SAMPLE DOCUMENT

Features	Frequency
dimens	2
espionag	4
econom espionag	3
secret	2
econom	4

C. Subject Extraction from World Ontology

Let $\mathbb{T}$ be the term space of $\mathbb{S}$ in $\mathcal{O}$ and $\mathbb{T} = \bigcup_{s\in\mathbb{S}} label(s)$. A matrix coordinated by $\mathbb{T}$ and $\mathbb{S}$ can be obtained:

***Definition** 7:* Let $\langle \mathbb{S}, \mathbb{T}\rangle$ be the matrix coordinated by $\mathbb{T}$ and $\mathbb{S}$, where a mapping exists:

$$\mu : \mathbb{T} \to 2^{\mathbb{S}}, \quad \mu(t) = \{s \in \mathbb{S} | t \in label(s)\} \subseteq \mathbb{S}$$

and its reverse mapping also exists:

$$\mu^{-1} : \mathbb{S} \to 2^{\mathbb{T}}, \quad \mu^{-1}(s) = \{t \in \mathbb{T} | s \in \eta(t)\} \subseteq \mathbb{T}.\square$$

By $\mu : \mathbb{T} \to 2^{\mathbb{S}}$, a term $t \in \mathbb{T}$ maps to a set of subjects $\mathcal{S}_t \subseteq \mathbb{S}$. Thus, given the feature set $\mathcal{F}(d) = \{\langle p, w(p)\rangle\}$, a set of subjects can be extracted from $\mathbb{S}$:

$$\mathcal{S}_d = \bigcup_{t \in termset(\mathcal{F}(d))} \mu(t) \tag{5}$$

TABLE III
SUBJECTS EXTRACTED FOR THE SAMPLE DOCUMENT

Subject	Strength
Espionage	16.83
Espionage, economic	13.01
Space surveillance	13.01
Dimensions	9.24
Espionage, industry	9.24
Business espionage	8.98
Espionage literature	8.98
Espionage story	8.98
... ...	...

where $\mathcal{S}_d \subseteq \mathbb{S}$; $\mu(t) = \emptyset$ if $t \notin \mathbb{T}$.

By $\mu^{-1} : \mathbb{S} \rightarrow 2^{\mathbb{T}}$, a subject $s \in \mathbb{S}$ maps to a set of terms $\{t\} \subseteq \mathbb{T}$. Hence, with Eq. (5), a set of terms can be extracted from $\mu^{-1}(s)$ to expand d:

$$termset(d) = \bigcup_{s \in \mathcal{S}_d} \mu^{-1}(s) \tag{6}$$

Note that $termset(d) \neq d$. There exist some terms $\{t | t \in termset(d), t \notin d\}$ that are suggested by $\mathcal{S}_d$; there also exist some terms $\{t | t \notin termset(d), t \in d\}$ not in the term space $\mathbb{T}$ and thus, mapping to an empty subject set.

Because $\mathcal{S}_d$ is extracted using $\mathcal{F}(d) = \{\langle p, w(p)\rangle\}$, considering the weights of feature patterns, we can evaluate the terms $t \in termset(d)$:

$$w(t) = \sum_{p \in \{p | t \in termset(p), p \in \mathcal{F}(d)\}} w(p) \tag{7}$$

Considering the distribution of the terms spreading in other subject labels, the normalized form of term evaluation is defined as:

$$nw(t) = w(t) \times log(\frac{|\mathcal{S}_d|}{sf(t, \mathcal{S}_d)}) \tag{8}$$

where $sf(t, \mathcal{S}_d) = |\{s | t \in \mu^{-1}(s), s \in \mathcal{S}_d\}|$.

Subjects in $\mathcal{S}_d$ can finally be evaluated for their competence of summarizing d, using $nw(t)$ for all $t \in \mu^{-1}(s)$:

$$str(d, s) = \sum_{t \in \mu^{-1}(s)} nw(t) \tag{9}$$

By using the normalized form of terms, the subjects are competent for not only describing d but also distinguishing d from other documents in the document space Ω.

To prune away noisy subjects, a threshold, min_str, is applied to subject extraction. The subjects with $str(d, s) \geq min\_str$ are kept, whereas those with $str(d, s) < min\_str$ are dropped. During the experiments, different values were tested for min_str. The results revealed that setting min_str as the top 5th $str(d, s)$ value, a variable but a static value, gave the system the best performance. Table III shows the valid subjects extracted from the world ontology for the sample document in Fig. 2, using the features shown in Table II. Note that only the top subjects are displayed, because a total of 80 subjects survived the pruning process.

D. Generalising Subjects for categorisation

The subject set extracted from the ontology (as described in Section V-C) suffers from problems, such as the set being easily oversized and many subjects overlapping in their referring-to semantic space. As a result, the system complexity becomes high and its performance becomes difficult to handle when using the subject set. The extracted subject set must be generalized for semantic categorisation.

An example of how subjects extracted from the ontology overlap in their semantic space is displayed in Table III. Through common sense, we know that *Espionage* dominates *Espionage, economic*, *Espionage, industrial*, and *Business espionage*; that *Espionage literature* dominates *Espionage story*. This overlapping is caused by the same feature terms occurring in different subject labels. The overlapping space needs to be clarified and the noisy subjects need to be removed.

The algorithm of generalizing subjects is proposed based on the observation of the semantic overlapping of subjects. The algorithm is accomplished via investigating the relationships existing between these subjects. From Definitions 2 and 3, we know subjects in the world ontology are linked by semantic relations. Within the taxonomical structure, let s_1 and s_2 be two subjects and $s_1 \in ancestor(s_2)$ ($s_2 \in descendant(s_1)$). s_1 refers to a larger semantical extent than s_2, and thus, is more general than s_2. On the other hand, s_2 is more specific than s_1, thus focuses more on its referring-to topic. Such semantic relations can be revealed from an example. Let s_1 be *Automobile* and s_2 *Sedan*. *Automobile* contains *Car*, *Truck*, etc; *Car* contains *Sedan*, *Hatchback*, etc. *Automobile* covers broader extent than *Sedan*; vice versa, *Sedan* is more focused than *Automobile*. Therefore, if one subject is a descendant of another, the descendant can be removed because its referring-to semantical extent has already been covered by the other. By doing so, we have no information loss but limited focus (e.g., replacing *Sedan* by *Car*). With the same rule, if *Sedan* and *Hatchback* are both in the set, they may be replaced by their common ancestor *Car* without information loss, even if *Car* is not in the extracted set. Based on these, if some extracted subjects are under the same umbrella of an ancestor, their referring-to semantic extent is covered by that referred-to by their ancestor. Therefore, by losing no information but only limiting focus, we can replace these subjects with their ancestor, whether this common ancestor is in the extracted subject set or not.

The issue becomes how much focus we can afford to lose. A common ancestor chosen to replace its descendant subjects cannot be too far from the replaced descendants in the taxonomic structure, or the main focus will be lost. One extreme example is that we should never use *Thing* to replace any subject. *Thing* as the root dominates all subjects in the ontology. An ancestor subject being too far from its descendants reduces meaning. Therefore, we use only the lowest common ancestor (LCA) to replace the descendant subjects. The LCA is defined as the common ancestor of a set of subjects with the shortest distance to these subjects in the taxonomic structure of ontology. The LCA dominates descendant subjects and covers their semantic extent with only

limited loss of focus.

input : $\mathcal{S}_i = \{s_1, s_2, \ldots, s_j\}$ (subject set extracted i), $\mathcal{O}$;
output: $\mathcal{S'}_i = \{s_1, s_2, \ldots, s_k\}$ (subject set generalized to map i).

$\mathcal{S'}_i = \emptyset, \mathcal{S}_{temp} = \emptyset, \mathcal{S}_{redundant} = \emptyset$;
foreach $s \in \mathcal{S}_i$ **do**
  Extract $S(s)$ from $\mathcal{O}$ where $S(s) = \{s' | s' \in ancestor(s), \delta(s \mapsto s') \leq 3\}$;**foreach** $s_n \in \mathcal{S}_i$ where $s_n \neq s$ **do**
    Extract $S(s_n)$ from $\mathcal{O}$ like Step 3;
    if $S(s) \cap S(s_n) \neq \emptyset$ **then** $\{\widehat{s} = \mathcal{LCA}(S(s) \cup S(s_n)), str(i, \widehat{s}) = str(i, s) + str(i, s_n)$;
      $\mathcal{S}_{temp} = \mathcal{S}_{temp} \cup \{\widehat{s}\}$;
      $\mathcal{S}_{redundant} = \mathcal{S}_{redundant} \cup \{s, s_n\}$;
    }
  end
  if $\mathcal{S}_{temp} \neq \emptyset$ **then** $\{\mathcal{S'}_i = \mathcal{S'}_i \cup \mathcal{S}_{temp}$;
    $\mathcal{S}_i = \mathcal{S}_i - \mathcal{S}_{redundant}$; $\mathcal{S}_{temp} = \emptyset$; $\mathcal{S}_{redundant} = \emptyset\}$;
  else $\mathcal{S'}_i = \mathcal{S'}_i \cup \{s\}$;
end
return $\mathcal{S'}_i$.

Algorithm 1: Generalizing Subjects

Algorithm 1 explains the process of semantic categorisation of a document via generalising the subjects initially extracted from the ontology. $\delta(s_1 \mapsto s_2)$ is a function measuring the distance between two subjects, which is completed by counting the number of edges travelled from s_1 to s_2 in the taxonomic structure of ontology. $\mathcal{LCA}(S(s_1) \cup S(s_2))$ is a function returning $\widehat{s}$, the LCA of s_1 and s_2 in a joint subject set, $S(s_1) \cup S(s_2)$.

Table IV presents the categorisation results generalized from the subjects displayed in Table III, with the min_str set as the top 5th $str(i, s)$ value again. Similar subjects like *Espionage*, *Espionage, economic*, *Espionage, industrial*, and *Business espionage*, have been merged and replaced by their LCA *Espionage* and *Business Intelligence*; *Espionage literature* and *Espionage story* replaced by *Spy story*. Consequently, the 80 subjects initially extracted from the world ontology (as described in Section V-C previously) are generalized to a much shorter list with only five subjects. This semantic categorisation result is meaningful, and in terms of semantics very close to the subjects listed with the sample document in Fig. 2, which were manually assigned by linguists.

VI. Experimental Evaluation

A. Experiment Design

Ideally, to categorise a document, the subjects automatically generated by the proposed approach should be exactly the same as those specified by specialist librarians. Though such a goal is unrealistic, the ideal scenario inspirited the design of our evaluation experiments. The proposed method was evaluated, based on the ground truth of manual assignment of subjects from linguists and compared against typical baseline classification methods.

TABLE IV
Generalized Subjects for the Sample Document

Subject	Strength
Espionage	269.78
Business Intelligence	203.83
Space surveillance	17.96
Spy story	16.27
Dimensions	9.24

TABLE V
Statistics of the Testing Set

Description	Stat.
Number of documents crawled	227,219
Number of documents used in experiments	31,902
Shortest length of documents in experiments	30
Longest length of documents in experiments	952
Average length of documents in experiments	85

The experiments were performed using a large testing set crawled from the catalogue of the University of Melbourne library[4]. The subject headings assigned to the catalogue items were manually specified by LCSH authorities through specialist librarians trained to specify subjects for a document without bias [9]. A sample catalogue item was presented in Section III. The title and content of catalogue items were used to form the content.

The text of each item in the catalogue was parsed first to remove unused information in this work, such as author name and Dewey Decimal Codes (Fig. 2 is an example document at this stage). The title and body of documents were equally removed during this process. General pre-processing techniques such as stopword removal and Porter stemming were applied to the preparation of the testing set for the experiment. Table V shows the statistics of the testing set (The length of documents refers to the number of terms in the documents after stopword removal). In the experiments, we used only documents having at least 30 terms. Documents shorter than that did not provide substantial frequent patterns, as revealed in preliminary experiments. By using the catalogue items in a library as the corpus, we could easily obtain a large testing set as well as a perfect ground truth for evaluation.

The subjects manually assigned to the documents by linguists provided the ideal ground truth in the experiments to measure the effectiveness of the proposed approach, against the automatically generated subjects. The objective evaluation methodology also assured the solidity and reliability of the experimental evaluation for our proposed method.

B. Baseline Models

Given that the LCSH ontology contains 394,070 subjects in our implementation, the semantic categorisation problem could also be understood as a $\mathcal{X}$-class classification problem where $\mathcal{X} = |\mathbb{S}| = 394,070$. Hence, we chose two typical multi-class classification approaches, *Rocchio* and kNN, for the baseline models in the experiments.

Rocchio is a simple and efficient classification method using centroid to define the class boundaries. The centroid of a subject s is computed as the vector average:

$$\vec{\mu}(s) = \frac{1}{|D_s|} \sum_{d \in D_s} \vec{v}(d)$$

[4]http://www.library.unimelb.edu.au/

In the experiments, the training set D_s contained only a single document $d = label(s)$. The $\vec{v}(d)$ was evaluated by using the frequency of terms in $label(s)$. The distance between a document and a subject class was measured by cosine similarity. The document was then classified into the subject classes with the top cosine value (Considering that $\mathcal{X} = |\mathbb{S}| = 394070$ is a huge number, using only the top value has already generated a considerably large set of subjects).

Unlike *Rocchio*, k Nearest Neighbour (kNN) determines the decision boundary locally and classifies documents into the major class of its k closest neighbours. When inputting a document d from the testing set, we extracted the closest neighbours $NN(d)$ that had the highest cosine similarity value with d. Because the testing documents were usually short, a large number of documents had the same cosine values. Thus, we set $k = 1$ to limit the number of considerable neighbours and ensure the highest possible accuracy. The distance of a s and a d is then evaluated by aggregating the cosine value of each $d' \in NN(d)$ to s. Again, d was classified into the subjects with only the top cosine value.

C. Performance Measuring Methods

The performance of the experimental models were measured by standard methods, precision and recall [5]. For the semantic categorisation problem, precision measured the ability of a method to categorise a document with highly-focused subjects, and recall with high-coverage of possible subjects.

As discussed previously, considering that $\mathcal{X} = |\mathbb{S}| = 394070$, pursuing the exact same subjects as those manually assigned by linguists is an unrealistic task. Thus, in respect to the testing set and the ground truth featured by the LCSH, performance was evaluated by:

$$precision = \frac{|\mathcal{FT}(S_{tgt}) \cap \mathcal{FT}(S_{grt})|}{|\mathcal{FT}(S_{tgt})|}$$

$$recall = \frac{|\mathcal{FT}(S_{tgt}) \cap \mathcal{FT}(S_{grt})|}{|\mathcal{FT}(S_{grt})|}$$

where $\mathcal{FT}(S) = \bigcup_{s \in S} \mu^{-1}(s)$ (see Definition 7); tgt referred to the target experimental model; grt referred to ground truth subjects.

In the experiments we also employed *micro*-F_1 Measure:

$$F_1 = \frac{2 \times precision \times recall}{precision + recall}$$

Precision and recall were evenly weighted in F_1 Measure. Each document's categorisation result was evaluated first and then all results were averaged for the final F_1 value. As with precision and recall, greater F_1 values indicated better performance.

VII. Results and Discussions

A. Experimental Results

Calling the proposed semantic categorisation approach the *OntoSum* model, the experiments compare the effectiveness of the performance of *OntoSum* against the baselines *Rocchio* and kNN models. Their effectiveness performances are depicted in Fig. 4, 5, and 6 for precision, recall, and F-Measure results, respectively. The value axis indicates the effectiveness rate between 0 and 1; the category axis indicates the number of documents whose categorisation results meet the indicating effectiveness rate. The number of documents is counted for those with only valid values (> 0).

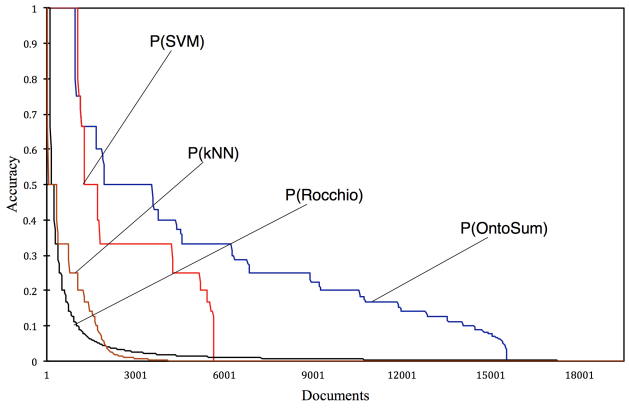


Fig. 4. Experimental Precision Results

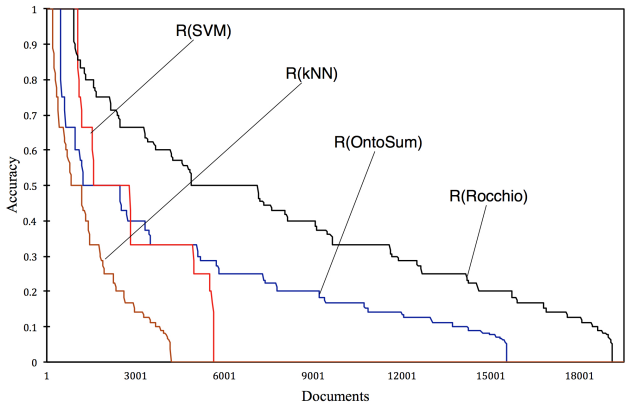


Fig. 5. Experimental Recall Results

The overall average performance is presented in Table VI. The F_1 measure equally considers both precision and recall when measuring performance. Thus the F_1 results are an overall effectiveness performance. The average F_1 results shown in Table VI reveal that the *OntoSum* model has achieved much better overall performance (0.125115) than the baseline models (0.019980 and 0.016305). This is also depicted in Fig. 6, where the $F(OntoSum)$ line is located at much higher bound level compared with the $F(Rocchio)$ and $F(kNN)$ lines.

Precision measures the accuracy of categorisation. For this, the *OntoSum* model also outperformed the baseline models. The average precision results in Table VI show this, with *OntoSum* 0.157992 vs. *Rocchio* 0.020259 and kNN 0.02077. Additionally, in Fig. 4, $P(OntoSum)$ is much higher than the

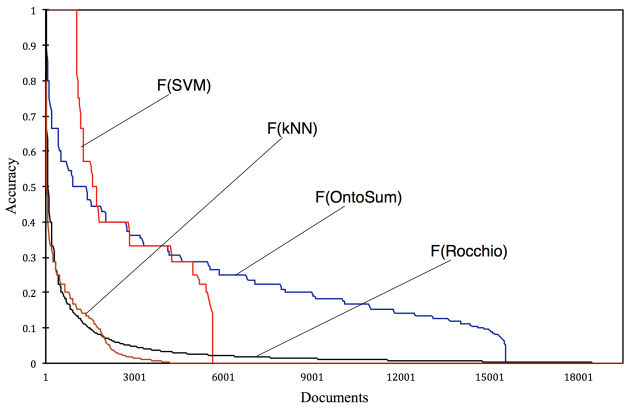


Fig. 6. Experimental F-Measure Results

TABLE VI
EFFECTIVENESS PERFORMANCE ON AVERAGE

	Precision	Recall	F-Measure
OntoMap	0.157992	0.134965	0.125115
SVM	0.0834775	0.093606	0.087678
Rocchio	0.020259	0.290226	0.019980
kNN	0.02077	0.053931	0.016305

other two.

Recall measures the semantic coverage of categorisation. The recall performance in the experiments shows a slightly different result compared to F_1 Measure and precision performance. The *Rocchio* model achieved the best recall performance (0.290226 on average), outperforming both the *OntoSum* (0.134965) and kNN model (0.053931). This is also illustrated in Fig. 5, in which $R(OntoSum)$ is in the middle of $R(Rocchio)$ and $R(kNN)$.

B. Discussions

There was a gap between the recall performance of the *OntoSum* and the baselines. After investigation, we found that the categorisation result of the *Rocchio* model was usually a large set of subjects (935 on average for each document), whereas the *OntoSum* model was 10 and the kNN 106. Due to the nature of recall, more features would be covered if the subject size became larger. As a result, the *Rocchio* categorisation with the largest size achieved the best recall performance. The subject sets generated by the kNN model had a larger size than those of the *OntoSum*. However, when taking neighbours into account, a large deal of noisy data was also brought into the neighbourhood - the average number of neighbours was 336. This was caused by the very large subject set in ontology and short documents. Thus, the categorisation became inaccurate, although only the subjects with the top similarity values were chosen to categorise a document. That is why the *OntoSum* sat in the middle of the *Rocchio* and kNN.

A different number of levels were tested in the sensitivity study for choosing the right number of levels to find the lowest common ancestor when generalising subjects for final categorisation (The relevant discussion is in Section V-D). Table VII displays the results for finding such a level. In the same experimental environment, when tracing three levels to find a LCA, the *OntoSum* model's performance - including F_1 Measure, precision, and recall - was better than that by five levels. In addition, tracing only three levels gave us lower complexity. Therefore, we chose three levels to find LCAs.

TABLE VII
SENSITIVITY STUDY RESULTS FOR TRACING A RIGHT NUMBER OF LEVELS IN ONTOLOGY TO FIND THE LOWEST COMMON ANCESTORS (LCAS)

	Precision	Recall	F-Measure
Level = 3	0.157992	0.134965	0.125115
Level = 5	0.154302	0.111632	0.111373

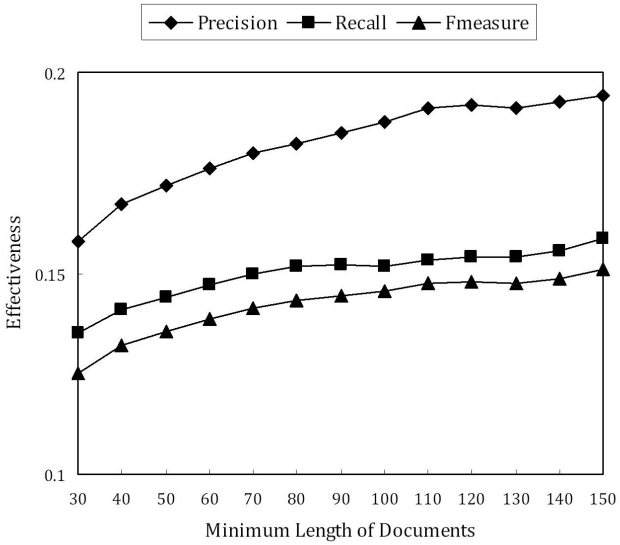


Fig. 7. Effectiveness of categorising Documents with Different Length

We also found that the performance of the *OntoSum* model slightly improved when the documents were relatively long. Figure 7 depicts the performance made by the *OntoSum* model on the documents with different minimum lengths. When the length of documents increased, the effectiveness sightly increased as well. Such an improvement is believed to be the result of the contribution of closed frequent sequential patterns discovered from documents (see Section V-B for details). When the *OntoSum* had the best performance with only documents longer than 150 terms, the average number of closed frequent sequential patterns was 27; when only with documents with length$>=$ 90, the average number of patterns was 17; when considering all documents (length$>=$ 30), the average number of discovered patterns dropped to 11. These results reveal that more useful and meaningful patterns would help the semantic categorisation in our approach. Given that more patterns would lead to more subjects extracted from the ontology, these facts also suggest that the generalising algorithm in the proposed approach successfully handled the extracted subjects well without sacrificing much information.

VIII. Conclusions

Semantic categorisation of text documents has become more important than ever, given that information in electronic form grown explosively. Many categorisation techniques have bottlenecks, such as being too expensive because of the large involvement of human effort, or are ineffective due to inadequate knowledge bases. The contribution of the work presented here addresses these bottlenecks, by introducing a semantic categorisation approach using a large world knowledge ontology built from the LCSH. A subject generalization algorithm has also been proposed in the work aiming to improve the performance of semantic categorisation. The approach was successfully evaluated through comparing typical text classification methods across a large testing set, measured by the categorisation manually made by linguists. This work contributes to text classification by demonstrating the value of an adequate and competent world knowledge ontology.

References

[1] B. Altinel, M.C. Ganiz, Semantic text classification: A survey of past and recent advances, Information Processing and Management, 54 (6) (2018) 1129-1153. doi:https://doi.org/10.1016/j.ipm.2018.08.001.

[2] B. Altınel, M.C. Ganiz, A new hybrid semi-supervised algorithm for text classification with class-based semantics, Knowledge-Based Systems, 108, 2016, 50–64.

[3] R. Bekkerman, M. Gavish, High-precision phrase-based document classification on a modern scale, in: Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, KDD '11, ACM, New York, NY, USA, 2011, pp. 231–239. doi:http://doi.acm.org/10.1145/2020408.2020449.
URL http://doi.acm.org/10.1145/2020408.2020449

[4] A. Bossard, Using document structure for automatic summarization, in: SIGIR '09: Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval, ACM, New York, NY, USA, 2009, pp. 849–849. doi:http://doi.acm.org/10.1145/1571941.1572170.

[5] C. Buckley, E. M. Voorhees, Evaluating evaluation measure stability, in: Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval, 2000, pp. 33–40.

[6] D. Cai, C. Zhang, X. He, Unsupervised feature selection for multi-cluster data, in: Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, KDD '10, ACM, New York, NY, USA, 2010, pp. 333–342. doi:http://doi.acm.org/10.1145/1835804.1835848.
URL http://doi.acm.org/10.1145/1835804.1835848

[7] F. Camastra, A. Ciaramella, A. Maratea, L.H. Son, A. Staiano, Semantic maps for knowledge management of web and social information, in: Computational Intelligence for Semantic Knowledge Management 2020 (pp. 39-51). Springer, Cham.

[8] F. Camous, S. Blott, A. F. Smeaton, Ontology-based medline document classification, in: Proceedings of the 1st international conference on Bioinformatics research and development, BIRD'07, Springer-Verlag, Berlin, Heidelberg, 2007, pp. 439–452.

[9] L. M. Chan, Library of Congress Subject Headings: Principle and Application, Libraries Unlimited, 2005.

[10] Z. Dou, R. Song, J.-R. Wen, A large-scale evaluation and analysis of personalized search strategies, in: WWW '07: Proceedings of the 16th international conference on World Wide Web, ACM Press, New York, NY, USA, 2007, pp. 581–590. doi:http://doi.acm.org/10.1145/1242572.1242651.

[11] E. Gabrilovich, S. Markovitch, Feature generation for text categorization using world knowledge, in: Proceedings of The Nineteenth International Joint Conference for Artificial Intelligence, Edinburgh, Scotland, 2005, pp. 1048–1053.
URL http://dl.acm.org/citation.cfm?id=1762370.1762413

[12] Y. Gao, Y. Xu, Y. Li, Pattern-based topics for document modelling in information filtering, IEEE Trans. Knowl. Data Eng. 27, 6 (2015), 1629—1642

[13] J. Gope, S.K. Jain, A survey on solving cold start problem in recommender systems, in: 2017 International Conference on Computing, Communication and Automation (ICCCA), IEEE, 2017 May 5, pp. 133–138.

[14] X. Hu, X. Zhang, C. Lu, E. K. Park, X. Zhou, Exploiting wikipedia as external knowledge for document clustering, in: KDD '09: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, New York, NY, USA, 2009, pp. 389–396. doi:http://doi.acm.org/10.1145/1557019.1557066.

[15] S. Huang, W. Peng, J. Li, D. Lee, Sentiment and topic analysis on social media: a multi-task multi-label classification approach, in: Proceedings of the 5th annual ACM web science conference, 2013, pp. 172–181.

[16] M. E. Houle, N. Grira, A correlation-based model for unsupervised feature selection, in: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management, CIKM '07, ACM, New York, NY, USA, 2007, pp. 897–900. doi:http://doi.acm.org/10.1145/1321440.1321570.
URL http://doi.acm.org/10.1145/1321440.1321570

[17] T. Joachims, Text categorization with Support Vector Machines: learning with many relevant features, in: Proceedings of the 10th European conference on machine learning, no. 1398, Springer Verlag, Heidelberg, DE, Chemnitz, DE, 1998, pp. 137–142.
URL citeseer.ist.psu.edu/joachims97text.html

[18] G. Kasper, D. de Siqueira Braga, D.M. Lima Martins, B. Hellingrath, User profile acquisition: A comprehensive framework to support personal information agents, in: Proceedings of the 2017 IEEE Latin American Conference on Computational Intelligence (LA-CCI), 2017, pp. 1–6. IEEE.

[19] Z. Kastrati, A.S. Imran, S.Y. Yayilgan, The impact of deep learning on document classification using semantically rich representations, Information Processing and Management 56(5), 2019, 1618–1632.

[20] I. Katakis, G. Tsoumakas, I. Vlahavas, Multilabel text classification for automated tag suggestion, in: In: Proceedings of the ECML/PKDD-08 Workshop on Discovery Challenge, 2008.

[21] S. Kulkarni, A. Singh, G. Ramakrishnan, S. Chakrabarti, Collective annotation of wikipedia entities in web text, in: KDD '09: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, New York, NY, USA, 2009, pp. 457–466. doi:http://doi.acm.org/10.1145/1557019.1557073.

[22] Y. Li, N. Zhong, Mining Ontology for Automatically Acquiring Web User Information Needs, IEEE Transactions on Knowledge and Data Engineering 18(4) (2006) 554–568.

[23] Y. Li, A. Algarni, S.-T. Wu, Y. Xu, Mining negative relevance feedback for information filtering, in: Proceedings of the IEEE/WIC/ACM international conference on Web Intelligence, 2009, pp. 606–613.

[24] Y. Li, A. Algarni, N. Zhong, Mining positive and negative patterns for relevance feature discovery, in: Proceedings of 16th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2010, pp. 753–762.

[25] B. Liu, Y. Dai, X. Li, W. Lee, P. Yu, Building text classifiers using positive and unlabeled examples, in: Proceedings of the Third IEEE International Conference on Data Mining, ICDM2003, 2003, pp. 179–186.

[26] P. Luo, F. Lin, Y. Xiong, Y. Zhao, Z. Shi, Towards combining web classification and web information extraction: a case study, in: KDD '09: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, New York, NY, USA, 2009, pp. 1235–1244. doi:http://doi.acm.org/10.1145/1557019.1557152.

[27] B. Maleszka, A method for ontology-based user profile adaptation in personalized document retrieval systems, in: 2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC), IEEE, 2016, pp. 003187-003192.

[28] H. H. Malik, J. R. Kender, Classifying high-dimensional text and web data using very short patterns, in: Proceedings of the 2008 Eighth IEEE International Conference on Data Mining, IEEE Computer Society, Washington, DC, USA, 2008, pp. 923–928. doi:10.1109/ICDM.2008.139.
URL http://dl.acm.org/citation.cfm?id=1510528.1511336

[29] G. Qiu, K. Liu, J. Bu, C. Chen, Z. Kang, Quantify query ambiguity using odp metadata, in: SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, ACM Press, New York, NY, USA, 2007, pp. 697–698. doi:http://doi.acm.org/10.1145/1277741.1277864.

[30] D. Ravindran, S. Gauch, Exploiting hierarchical relationships in conceptual search, in: Proceedings of the 13th ACM international conference on Information and Knowledge Management, ACM Press, New York, USA, 2004, pp. 238–239. doi:http://doi.acm.org/10.1145/1031171.1031221.

[31] F. Sebastiani, Machine learning in automated text categorization, ACM Computing Surveys (CSUR) 34 (1) (2002) 1–47. doi:http://doi.acm.org/10.1145/505282.505283.

[32] A. Sieg, B. Mobasher, R. Burke, Learning ontology-based user profiles: A semantic approach to personalized web search, The IEEE Intelligent Informatics Bulletin 8(1) (2007) 7–18.

[33] A. Sieg, B. Mobasher, R. Burke, Web search personalization with ontological user profiles, in: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management, ACM, New York, NY, USA, 2007, pp. 525–534. doi:http://doi.acm.org/10.1145/1321440.1321515.

[34] A. Singh, A. Sharma, N. Dey, Semantics and agents oriented web personalization: state of the art, International Journal of Service Science, Management, Engineering, and Technology (IJSSMET), 6,(2), (2015), 35–49.

[35] X. Tao, Y. Li, N. Zhong, A personalized ontology model for web information gathering, IEEE Transactions on Knowledge and Data Engineering, IEEE computer Society Digital Library. IEEE Computer Society 23 (4) (2011) 496–511. doi:http://doi.ieeecomputersociety.org/10.1109/TKDE.2010.145.

[36] J. Teevan, C. Alvarado, M. S. Ackerman, D. R. Karger, The perfect search engine is not enough: a study of orienteering behavior in directed search, in: Proceedings of the SIGCHI conference on Human factors in computing systems, 2004, pp. 415–422. doi:http://doi.acm.org/10.1145/985692.985745.

[37] J. Teevan, S. T. Dumais, E. Horvitz, Personalizing search via automated analysis of interests and activities, in: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval, 2005, pp. 449–456. doi:http://doi.acm.org/10.1145/1076034.1076111.

[38] X. Wan, J. Yang, Multi-document summarization using cluster-based link analysis, in: SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval, ACM, New York, NY, USA, 2008, pp. 299–306. doi:http://doi.acm.org/10.1145/1390334.1390386.

[39] S. Wang, G. Englebienne, S. Schlobach, Learning concept mappings from instance similarity, in: A. Sheth, S. Staab, M. Dean, M. Paolucci, D. Maynard, T. Finin, K. Thirunarayan (Eds.), The Semantic Web - ISWC 2008, Vol. 5318 of Lecture Notes in Computer Science, Springer Berlin / Heidelberg, 2008, pp. 339–355.

[40] Y. Yan, N. Okazaki, Y. Matsuo, Z. Yang, M. Ishizuka, Unsupervised relation extraction by mining wikipedia texts using information from the web, in: Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2 - Volume 2, ACL '09, Association for Computational Linguistics, Stroudsburg, PA, USA, 2009, pp. 1021–1029.
URL http://dl.acm.org/citation.cfm?id=1690219.1690289

[41] B. Yang, J.-T. Sun, T. Wang, Z. Chen, Effective multi-label active learning for text classification, in: KDD '09: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, New York, NY, USA, 2009, pp. 917–926. doi:http://doi.acm.org/10.1145/1557019.1557119.

[42] T. Yang, R. Jin, A. K. Jain, Y. Zhou, W. Tong, Unsupervised transfer classification: application to text categorization, in: Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, KDD '10, ACM, New York, NY, USA, 2010, pp. 1159–1168. doi:http://doi.acm.org/10.1145/1835804.1835950.
URL http://doi.acm.org/10.1145/1835804.1835950

[43] Z. Yang, D. Yang, C. Dyer, X. He, A. Smola, E. Hovy, Hierarchical attention networks for document classification, in: Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: human language technologies, San Diego, California, USA, 2016, Jun pp. 1480–1489.

[44] L. Yu, S. Wang, K. K. Lai, An integrated data preparation scheme for neural network data analysis, IEEE Transactions on Knowledge and Data Engineering 18 (2) (2006) 217–230. doi:http://dx.doi.org/10.1109/TKDE.2006.22.

[45] L. Zadeh, Web intelligence and world knowledge - the concept of Web IQ (WIQ), in: Processing NAFIPS '04, IEEE Annual Meeting of the Fuzzy Information, 2004., Vol. 1, 2004, pp. 1–3.

[46] D. Zha, C. Li. Multi-label dataless text classification with topic modeling, Knowledge and Information Systems, 61 (2019) 137-–160. https://doi.org/10.1007/s10115-018-1280-0

Prediction and Categorization of Heart Arhythymia

Nishitha Doris Rebecca and S.N. Prasad

School of ECE, Reva University, Bangalore

E-mail: ndorisrebecca@gmail.com, prasadsn@reva.edu.in

***Abstract*— Heart arrhythmia is a state of the heart in which the heartbeat is unbalanced, either too fast, too slow or unstable. Electrocardiography (ECG) is used for the recognition of Heart arrhythmia. It registers the electrical activities of the heart of a patient for a period through electrodes attached to the skin. Due to the ECG signals that reflect the physiological conditions of the heart, medical specialists tend to utilize ECG signals to detect and analyze heart arrhythmia. The most important skill of medical doctors is being able to identify the dangerous types of heart arrhythmia from ECG signals. In spite of this, interpretation of the ECG waveforms performed by a professional medical doctor manually is proven to be monotonous and time consuming. As a result, the development of automatic systems for identifying abnormal conditions from diurnal recorded ECG data is of primary importance. Suitable and timely medical treatment measures can be effectively applied when such irregular heart conditions are identified instantly using health monitoring equipment and tools utilizing machine learning algorithms. Therefore, an important investigation in this regard would be machine learning approaches.**

***Index Terms*— Cardiac autonomic nervous system, Cardiac arrhythmias, Atrial fibrillation, Ventricular tachyarrhythmia, Denervation, Nerve stimulation, Neuromodulator**

I. INTRODUCTION

Effective treatment and management now exists for many arrhythmias. Devices and high-level catheters, along with computerised-plotting systems that permit for ablation treatment and therapy, have produced some notable and incredible clinical electrophysiology into one of the most rapidly multiplying cardiology subspecialties.

Pacemakers are the acknowledged standard model of supervision for those with bradycardia, and if facilities are available, patients with Wolff-Parkinson-White syndrome or similar arrhythmias should be looked up for ablation.

However, knowledge of the underlying biology has not kept up with technical improvements, and queries about clinical management remain. Foremost, although we know some of the common factors that incline and prompt arrhythmias, the evaluation precision is not always sufficient to justify prophylaxis or intervention. Secondly, if we want to suppress arrhythmia not responsive to ablation, we have few options. In the past thirty years, the range of available drugs has scarcely expanded, and available drug treatments have pro-arrhythmic risk, other toxic effects, low tolerability, and variable efficacy. However, recent developments suggest that most of the arrhythmia biology is tractable. Improvements in the relevant genetics and genomics, and the availability of data and new model systems, are reassuring. The promising picture is one of many molecular perturbations that come together and interact in individuals to generate arrhythmia-prone hearts, expressed through the phenotypic variability familiar to clinicians.

II. BACKGROUND

A. *Machine Learning*

Artificial Intelligence (AI) is a rapidly advancing technology, that can learn, reason, plan, perceive or process natural language. From evaluations made by many scientists, the term “machine learning” is interchangeably used along with the term “artificial intelligence”, given that the possibility of learning is the main characteristic of intelligent agents. The most significant purpose of machine learning is the formation and building of computer program code that can learn, test and improvise and adapt accordingly, using past experience and data.

B. *Supervised Learning*

This assessment is determined on the study of a system based on distinct methods of supervised learning. In supervised learning, the system must “learn”, including using target function. This target function is an abstract of a model which describes the data. In order to conclude the best target result, the learning system, given a training set, must take appropriate hypothesis for the function and be represented by h.

In supervised learning, there are two kinds of learning tasks: classification and regression. Classification models envisage distinct classes using its trained data, such as e.g. blood groups, while regression models forecast numerical values. Some of the most common procedures are Support Vector Machines (SVM), Decision Trees (DT), Genetic Algorithms (GA), Artificial Neural Networks (ANN) and Instance Based Learning (IBL), such as k- Nearest Neighbors (k-NN).

III. METHODOLOGY

Analysis of the heart state or normal ECG waves is not considered an easy task. As a point of fact, the ECG signal is nonstationary and thus, symptoms of a disease, if any, may not occur regularly. Thus, medical specialists need to document the records and closely observe the heartbeat for a long time to categorize the rhythm into regular or irregular type. For ECG signal analysis, the size of the generated data can be huge, which requires a lot of time and effort, therefore

there is a need for an automatic classification system.

A. Dataset

For the current study, publicly available Physio Net, MIT–BIH arrhythmia database sampled at 360 Hz is used. Then, the heartbeats from the complete dataset collected are categorized into five arrhythmia classes as suggested by the ANSI/AAMI EC57:1998 standard. The MIT-BIH database comprises of 48 registered records. Each recorded note has a period of 30 minutes with sampling frequency of 360 Hz. Table 1 shows the heartbeat distribution.

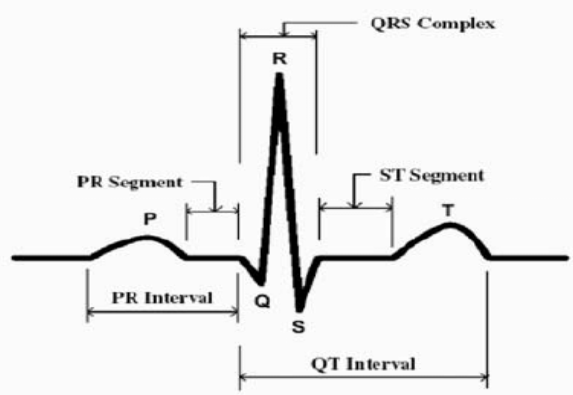
Table 1. Distribution of Heartbeats

Heartbeat	ECG Recording Containing Respective Type
N	100, 101, 105, 112, 115, 000, 000
LBBB	109, 111, 207, 214
RBBB	124, 212, 231, 232
PVC	105, 109, 116, 119, 214, 000, 000

For suitable feature selection, we intend to use Machine Learning Algorithms: K-Nearest Neighbors, Logistic Regression, Naïve Bayes and SVM.

Each standard ECG signal has components that are composed of P-wave, QRS complex, followed by T wave as shown in Figure 1. On inspecting the shape, the correlation between these waves and the duration of each waves is used to analyze the diagnosis and category of the arrhythmia.

Figure 1. Components of ECG Signal.



B. Pre-processing the ECG Signal for the Intended System

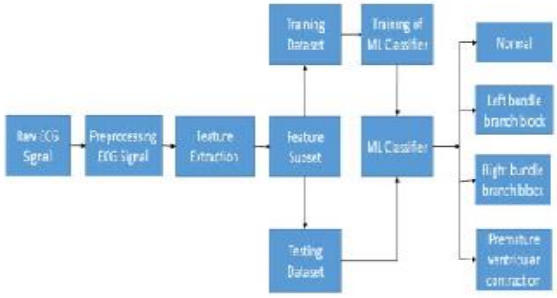
Input to the system will be detailed and rare ECG signals. This detailed signal contains noise. Pre-processing of the ECG signal detaches this noise. Three different DE noising techniques are used: median filter, moving average filter and notch filter. Following this, features are removed from the filter ECG signal. In total 9 characteristics are removed for each beat using discrete wavelet transform, namely R point location, area under QRS complex, duration of QR, RS, RR points, R peak, R normal, area under autocorrelation and SVD of ECG. Various techniques such as FFT, CWT and DWT etc., will be used for the extraction of different features from the expired ECG signal. The resulting feature dataset will then be split into training and testing datasets. Training dataset will be fed to the different Machine Learning classifier. In the suggested system, an SVM classifier and an ANN classifier will be used.

Different hypotheses and weights will be taken into consideration to raise the accuracy of ranking. At the end, the best combination of the pre-processing and classification techniques resulting from the proposed system will be used to more accurately identify the type of heart arrhythmia.

C. System Evaluation Approach

The proposed system architecture is shown in Figure 2.

Figure 2. Proposed System Architecture



For conducting evaluation, we used three standard metrics: sensitivity, specificity, and accuracy. These metrics are used to quantify the performance of the system.

Sensitivity is a measure of the capacity of the positive samples and is denoted by:

$$(Sn) = (TP/ TP + FN) * 100$$

Where TP represents the real positive and FN represents the false negative.

Specificity is measures of the capacity of test the negative samples, defined by:

$$(Sp) = (TN / TN + FP) * 100$$

Where TN represents the true negative and FP represents the false positive.

The precision accuracy is described as the ability of the test to correctly identify a classified type with and without positives. It reflects both sensitivity and specificity.

$$\text{Accuracy (Ac)} = (TP + TN) / (TP + TN + FP + FN) * 100$$

IV. Conclusion

The proposed machine learning system can be used in hospitals or medical diagnostic centres, where a large dataset is available. They can can assist medical specialists in developing more precise analytics decisions and to cut down the number of causalities due to heart diseases in the future. This classification technique is based on the algorithms K-Nearest Neighbors, Logistic Regression, Naïve Bayes and SVM, with an assumption of independence among predictors.

BIBLIOGRAPHY

[1] H.I. Bulbul, N. Usta, and M. Yildiz. "Classification of ECG arrhythmia with machine learning techniques." In 2017 *16th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pp. 546-549. IEEE, 2017.

[2] J. Park, S. Lee, and K. Kang. "Arrhythmia detection using amplitude difference features based on random forest." In 2015 *37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 5191-5194. IEEE, 2015.

[3] T. Paul, A. Chakraborty, and S. Kundu. "Hybrid shallow and deep learned feature mixture model for arrhythmia classification." In 2018 *Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT)*, pp. 1-4. IEEE, 2018.

[4] S. Savalia and V. Emamian. "Cardiac arrhythmia classification by multi-layer perceptron and convolution neural networks." *Bioengineering* 5, no. 2 (2018): 35.

[5] K. Padmavathi and K. Sri Ramakrishna. "Classification of ECG signal during atrial fibrillation using autoregressive modeling." *Procedia Computer Science* 46 (2015): 53-59.

[6] A. Sharma and K. Bhardwaj. "Identification of normal and abnormal ECG using neural network." *International Journal of Information Research and Review* 2 (2015): 695-700.

[7] S. Subbiah, R. Patro, and P. Subbuthai. "Feature extraction and classification for ecg signal processing based on artificial neural network and machine learning approach." In *International conference on inter disciplinary research in engineering and technology*, pp. 50-57. 2015.

[8] G. Huang, G.B. Huang, S. Song, and K. You. "Trends in extreme learning machines: A review." *Neural Networks* 61 (2015): 32-48.

[9] R.G. Afkhami, G. Azarnia, and M.A. Tinati. "Cardiac arrhythmia classification using statistical and mixture modeling features of ECG signals." *Pattern Recognition Letters* 70 (2016): 45-51.

Selected Ph.D. Thesis Abstracts

This Ph.D thesis abstracts section presents theses defended in 2020 and 2021. These submissions cover a range of research topics and themes under intelligent informatics, such as proximity mining, decentralized community information systems, deep anomaly detection,and automated reasoning and cybersecurity.

DATASET PROXIMITY MINING FOR SUPPORTING SCHEMA MATCHING AND DATA LAKE GOVERNANCE

Ayman Alserafi
alserafi@essi.upc.edu
Universitat Politcnica de Catalunya (UPC), Spain, and
Universit Libre de Bruxelles (ULB), Belgium
https://hdl.handle.net/2117/345323

With the huge growth in the amount of data generated by information systems, it is common practice today to store datasets in their raw formats (i.e., without any data preprocessing or transformations) in large-scale data repositories called Data Lakes (DLs). Such repositories store datasets from heterogeneous subject-areas (covering many business topics) and with many different schemata. Therefore, it is a challenge for data scientists using the DL for data analysis to find relevant datasets for their analysis tasks without any support or data governance. The goal is to be able to extract metadata and information about datasets stored in the DL to support the data scientist in finding relevant sources. This shapes the main goal of this thesis, where we explore different techniques of data profiling, holistic schema matching and analysis recommendation to support the data scientist.

We propose a novel framework based on supervised machine learning to automatically extract metadata describing datasets, including computation of their similarities and data overlaps using holistic schema matching techniques. We use the extracted relationships between datasets in automatically categorizing them to support the data scientist in finding relevant datasets with intersection between their data. This is done via a novel metadata-driven technique called proximity mining which consumes the extracted metadata via automated data mining algorithms in order to detect related datasets and to propose relevant categories for them. We focus on flat (tabular) datasets organised as rows of data instances and columns of attributes describing the instances.

Our proposed framework uses the following four main techniques: (1) Instance-based schema matching for detecting relevant data items between heterogeneous datasets, (2) Dataset level metadata extraction and proximity mining for detecting related datasets, (3) Attribute level metadata extraction and prox- imity mining for detecting related datasets, and finally, (4) Automatic dataset categorization via supervised k-Nearest Neighbour (kNN) techniques. We implement our proposed algorithms via a prototype that shows the feasibility of this framework. We apply the prototype in an experiment on a real-world DL scenario to prove the feasibility, effectiveness and efficiency of our approach, whereby we were able to achieve high recall rates and efficiency gains while improving the computational space and time consumption by two orders of magnitude via our proposed early-pruning and pre-filtering techniques in comparison to classical instance-based schema matching techniques. This proves the effectiveness of our proposed automatic methods in the early-pruning and pre-filtering tasks for holistic schema matching and the automatic dataset categorisation, while also demonstrating improvements over human-based data analysis for the same tasks.

SCAFFOLDING DECENTRALIZED COMMUNITY INFORMATION SYSTEMS FOR LIFELONG LEARNING COMMUNITIES

Peter deLange
lange@dbis.rwth-aachen.de
RWTH Aachen University, Germany

Initially, the Web was developed as a decentralized system of information repositories that facilitate organizational knowledge transfer by allowing anyone to create and access content. However, Web publishing required both technical expertise and hardware infrastructure. With the rise of the Web 2.0, social networking sites and content management systems enabled all users to create Web content. But it simultaneously put the users at the mercy of the platform operators. Services could be shut down, erasing content and disrupting communities.

Decentralized community information systems radically change this dynamic by establishing participants as equal peers, which form a self-governing community. This way, a community regains control over their data, while being able to scale the infrastructure according to their needs.

In this dissertation, we followed a design science approach that provides support for communities to create and host their own, decentralized community information systems. On the one hand, we produced several artifacts to provide possible answers to the question of what properties such an infrastructure needs to fulfill. With the blockchain- based decentralized service registry, we propose a solution for making community knowledge accessible in a secure and verifiable way. On the other hand, we transfer the metaphor of educational scaffolding to the domain of service development. It is based on the idea, that a scaffold serves as a temporary supporting structure during a building's construction phase. As the construction site develops and the building gets completed, the scaffold gradually gets removed up to the point, that it is

not needed anymore. With the community application editor, communities are provided with such a scaffolding environment for requirements elicitation, wireframing, modeling and coding their decentralized community applications. Once deployed on the infrastructure, those applications and development efforts remain available, even after the contributing members might have left, serving as the community's long term memory.

We demonstrated and evaluated our artifacts on a large European scale, with three longitudinal studies conducted within several communities from different areas of technology enhanced learning, such as the European voluntary service, vocational and educational training providers and in higher education mentoring scenarios. All in all, this shift from data being stored in centralized repositories to a decentralized infrastructure, hosted by community members, opens up possibilities for a more democratic and egalitarian management of community knowledge.

DEEP ANOMALY DETECTION IN DISTRIBUTED SOFTWARE SYSTEMS

Sasho Nedelkoski
nedelkoski@tu-berlin.de
Distributed and Operating Systems, der Technischen Universität Berlin, Berlin, Germany

ARTIFICIAL Intelligence for IT Operations (AIOps) combines big data and machine learning to replace a broad range of IT Operations tasks. The task of anomaly detection has a prominent position in ensuring the required reliability and safe operation in distributed software systems. However, the frequent software and hardware updates, system heterogeneity, and massive amount of data create a challenging environment. The detection of anomalies in these systems predominantly relies on metric, log, and trace data. Each of them provides a different view of the internal states of the systems. By induction, improving the detection in every data source increases the overall anomaly detection performance in the system.

This thesis provides the following contributions. (1) We present a method based on variational inference and recurrent neural network to address the detection of anomalies in system metric data that possibly exhibit multiple modes of normal operation. (2) We propose a novel log parsing through language modelling that enables learning of log representations for downstream anomaly detection. We identify the learning of log representations as a major challenge toward a robust anomaly detection. Therefore, we additionally design a method that learns log representations by distinguishing between normal data from the system of interest and easily accessible anomaly samples obtained through the internet. (3) We describe a self-supervised anomaly detection task that utilizes the entire trace information to robustly detect anomalies that propagate through system components. (4) In a rule-based approach, we combine the presented methods for a multi-view anomaly detection.

The methods presented in this thesis were implemented in prototypes and evaluated on various datasets including production data from a cloud provider. They provided (1) an F1 score of 0.85 on metric data, (2) parsing accuracy of 99% and F1 score improvement of 0.25 in log anomaly detection, (3) increase in F1 score of 7% in trace anomaly detection over the state of the art, and (4) broadened spectrum of detected anomalies. The results were peer-reviewed and published at renowned international conferences.

HYPOTHESIS GENERATION VIA AUTOMATED REASONING WITH APPLICATIONS TO CYBERSECURITY

Jose N. Paredes
jose.paredes@cs.uns.edu.ar
Universidad Nacional del Sur (UNS), Bahia Blanca, Argentina

IN recent years, a wide variety of malicious behaviors have taken root in social platforms such as fake news, hate speech, malware diffusion, among others. This kind of behavior leads to a set of related problems, which has produced unforeseen consequences in many arenas; motivated by this situation, this thesis focuses on the study of automated hypothesis generation systems to address such issues. As a first contribution, two basic approaches are considered for the detection of a specific type of malicious behavior that we call adversarial deduplication. In the first approach, the generation of hypotheses is based on the use of well-defined logical rules, though the essence of its operation is supported by results obtained from applying previously-deployed machine learning techniques (that is, a part of the symbols necessary for the logical machinery are yielded by ML tools). In the second approach, ML techniques are used with a more central role; specifically, we use classifiers to tackle the problem, and hypothesis generation is carried out by simpler rules that are fired when the output of the classifiers exceeds a certain threshold.

Given that the initial proposal focuses on a specific problem (and therefore suffers some of the same limitations as ad-hoc approaches in the literature), our ultimate aim is to develop more robust and general systems. In particular, it is crucial to be able to handle situations not only single problems, but rather consider their multiplicity and take advantage of the relationships that may exist between them. In order to make progress in this direction, the main contribution is the presentation of the NETDER (Network Diffusion and Existential Rules) architecture to reason about malicious behavior on social platforms, which, in principle, seeks to serve as a guide for the implementation of software tools in such domains. NETDER includes four main modules: Data Ingestion (handles issues such as data cleaning, inconsistency, data analytics, among others, as well as other higher-level issues such as trust and uncertainty management); Ontological Reasoning (manages the knowledge base, both for background knowledge as well as for the network, and provides inference services); Network Diffusion (handles the evolution of the network in the form of diffusion processes, and checks conditions for the Ontological Reasoning Module); and Query Answering (handles the coordination of the two previous modules in order to answer the specific queries that users issue to the system).

After presenting the architecture, we study of the computational cost of query answering in its different instantiations, given that this process fundamentally drives the generation of hypotheses. Our analysis yields an interesting set of results that range from polynomial time tractability to the possibility that termination is not guaranteed, depending on the features that are made available to the model. Additionally, we develop a use case to illustrate how the approach can be applied in a cybersecurity domain to reason about products that are at risk of attack based on Darknet forum posts.

The final contribution is an experimental evaluation of the NETDER architecture. Given the difficulty of obtaining adequate datasets with ground truth (which is necessary to carry out performance evaluations), it was necessary to develop a general testbed designed with the purpose of generating social networks with complete traces of posting activities, potentially involving all kinds of malicious content such as fake news, malicious actors, botnets, links to malware, hate speech, etc. Our results constitute an important step towards achieving the ultimate goal, which is to develop automated robust hypothesis generation systems that can be used to address malicious behavior in social platforms.

RELATED CONFERENCES, CALL FOR PAPERS/PARTICIPANTS

TCII Sponsored Conferences

WI-IAT 2021
The 2021 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology
Melbourne, Australia (Hybrid conference with online and offline modes)
December 14-17, 2021
https://www.wi-iat.com/wi-iat2021/index.html

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'21 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. WI-IAT'21 welcomes research and application, as well as Industry/Demo-Track paper submissions.

This year celebrates the 20th anniversary of the WI-IAT. The 2021 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT'21) 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. Academics, professionals and industry people are encouraged to exchange their ideas, findings and strategies in utilizing the power of human brains and man-made networks to create a better world. More specifically, the fields of how intelligence is impacting the Web of People, the Web of Data, the Web of Things, the Web of Trust, the Web of Agents, and a special track: emerging Web in health and smart living in the 5G Era. Therefore, the theme of WI-IAT'21 will be "Web Intelligence = AI in the Connected World". There are approximately 79 topics spread across these 6 tracks.

Special events will be arranged for the celebration of WI-IAT's 20th anniversary. Every regular registration has included one ticket access to the WI-IAT 20th Anniversary Celebrations on-site at Berth Dockland, 45 New Quay Promenade. Docklands VIC, 3008.

ICDM 2021
The 21st IEEE International Conference on Data Mining
Auckland, New Zealand (Virtual Conference)
December 7-10, 2021
https://icdm2021.auckland.ac.nz/

The IEEE International Conference on Data Mining (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 and practical development experiences. The conference covers all aspects of data mining, including algorithms, software, systems, and applications. ICDM draws researchers, application developers, and practitioners from a wide range of data mining related areas such as big data, deep learning, pattern recognition, statistical and machine learning，databases, 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 advance the state-of-the-art in data mining.

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. There is also an encouragement of submissions for emerging topics of high importance, such as ethical data analytics, automated data analytics, data-driven reasoning, interpretable modeling, modeling with evolving environment, cyber-physical systems, multi-modality data mining, and heterogeneous data integration and mining.

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 2022
The 10th IEEE International Conference on Healthcare Informatics
Lisbon, Portugal
April 14-15, 2022
https://waset.org/healthcare-informatics-conference-in-april-2022-in-lisbon

The International Conference on Healthcare Informatics (ICHI 2022) aims to bring together leading academic scientists, researchers and research scholars to exchange and share their experiences and research results on all aspects of Healthcare Informatics. It also provides a premier interdisciplinary platform for researchers, practitioners and educators to present and discuss the most recent innovations, trends, and concerns as well as practical challenges encountered and solutions adopted in the fields of Healthcare Informatics.

The international research conference program is

designed for original research contributions and presentations in all research fields, with overarching categories including: Medical and Health Sciences Research; Human and Social Sciences Research, and Engineering and Physical Sciences Research. Each category has a number of underlying topics that are timely and emerging areas of interest in the domain. Prospective authors are kindly encouraged to contribute to and help shape the conference through submissions of their research abstracts, papers and e-posters. Also, high quality research contributions describing original and unpublished results of conceptual, constructive, empirical, experimental, or theoretical work in all areas of Healthcare Informatics are cordially invited for presentation at the conference.

ICHI 2022 has teamed up with the Special Journal Issue on Healthcare Informatics. A number of selected high-impact full text papers will also be considered for the special journal issues. All submitted papers will have the opportunity to be considered for this Special Journal Issue. The paper selection will be carried out during the peer review process as well as at the conference presentation stage. Submitted papers must not be under consideration by any other journal or publication. The final decision for paper selection will be made based on peer review reports by the Guest Editors and the Editor-in-Chief jointly. Selected full-text papers will be published online free of charge.

IEEE BigData 2021
The 2021 IEEE International Conference on Big Data (IEEE BigData 2021)
Virtual Conference
December 15-18, 2021
https://bigdataieee.org/BigData2021/

The 2021 IEEE International Conference on Big Data (IEEE BigData 2021) will continue the success of the previous IEEE Big Data conferences. It will provide a leading forum for disseminating the latest results in Big Data Research, Development, and Applications. The conference called for original research papers (and significant work-in-progress papers) in any aspect of Big Data, with emphasis on 5Vs (Volume, Velocity, Variety, Value and Veracity), including the Big Data challenges in scientific and engineering, social, sensor/IoT/IoE, and multimedia (audio, video, image, etc) big data systems and applications.

The IEEE BigData 2021 conference will feature regular and short papers across the main track and Industry and Government Program sessions. The main paper sessions will have 32 topics based around big data themes, while the Industry and Government Program sessions has 7 topics ranging from Machine Learning to IoT. This year's conference also has the BigData Cup Challenges, with focus on a Reinforcement Learning based Recommender System Challenge.

IEEE ICKG 2022
The 12th IEEE International Conference on Knowledge Graph (ICKG-2022)
Amsterdam, Netherlands
November 4-5, 2022
https://waset.org/knowledge-graphs-conference-in-november-2022-in-amsterdam

International Conference on Knowledge Graphs aims to bring together leading academic scientists, researchers and research scholars to exchange and share their experiences and research results on all aspects of Knowledge Graphs. It also provides a premier interdisciplinary platform for researchers, practitioners and educators to present and discuss the most recent innovations, trends, and concerns as well as practical challenges encountered and solutions adopted in the fields of Knowledge Graphs.

Tracks presented at the conference included: Machine Learning and Knowledge Graphs; Reasoning with Knowledge Graphs; Knowledge Graph Analytics and Applications; Knowledge Graphs and NLP; Knowledge graphs for Explainable AI; Multimodal Knowledge Graphs, Social Network and Representation Learning; Knowledge Graphs for Cultural Heritage; Knowledge Graphs for Geospatial Information Systems; Domain Knowledge Graphs, and Knowledge Graphs for Education.

AAMAS 2022
The 21st International Conference on Autonomous Agents and Multi-Agent Systems
Auckland, New Zealand
May 9-13, 2022
https://aamas2022-conference.auckland.ac.nz/

AAMAS (International Conference on Autonomous Agents and Multiagent Systems) is the largest and most influential conference in the area of agents and multiagent systems. The aim of the conference is to bring together researchers and practitioners in all areas of agent technology and to provide a single, high-profile, internationally renowned forum for research in the theory and practice of autonomous agents and multiagent systems. AAMAS is the flagship conference of the non-profit International Foundation for Autonomous Agents and Multiagent Systems (IFAAMAS).

AAMAS-22 sought 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 areas of interest, including: 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-2022 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-2022. The Demo Track, finally, allows participants from both academia and industry to showcase their latest developments in agent-based and robotic systems.

AAAI 2022
The 36th AAAI Conference on Artificial Intelligence
Vancouver, BC, Canada
February 22-March 1, 2022
https://aaai.org/Conferences/AAAI-22/

The Thirty-Sixth AAAI Conference on Artificial Intelligence (AAAI-21) will be held in Vancouver, BC, Canada, February 22 - March 1, 2021. The general chair will be Katia Sycara (Carnegie Mellon University, USA) and the

program cochairs will be Vasant Honavar (Pennsylvania State University, USA) and Matthijs Spaan (Delft University of Technology, Netherlands).

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-22 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-22 welcomes submissions on mainstream AI topics as well as novel crosscutting work in related areas. AAAI-22 will be co-located with the Thirty-Fourth Innovative Applications of Artificial Intelligence Conference (IAAI-22) and the Twelfth AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-22).

AAAI-22 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 we expressly encourage work that cuts across technical areas of AI, (e.g., machine learning and computer vision; computer vision and natural language processing; or machine learning and planning), 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.

As in past years, AAAI-22 will include a Track on AI for Social Impact (AISI). Submissions to this track will be reviewed according to a rubric that emphasizes the fit between the techniques used and a problem of social importance, rather than simply rewarding technical novelty. In particular, reviewers will assess novelty of the AI problem formulation studied; the paper's engagement with previous literature on the application problem (whether in the AI literature or elsewhere); both novelty of and justification for the proposed solution; quality of evaluation; facilitation of follow-up work; and overall scope and promise for social impact. Further details are available at https://aaai.org/Conferences/AAAI-22/aiforsocialimpactcall/

SDM22
The 2022 SIAM International Conference on Data Mining
Hybrid Conference
April 28-30, 2022
https://www.siam.org/conferences/cm/conference/sdm22

Data mining is the computational process for discovering valuable knowledge from data – the core of Data Science. It has enormous application 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, which are based on sound theoretical and statistical foundations. 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.

The SDM conference provides a venue for researchers who are addressing these problems to present their work in a peer-reviewed forum. It also provides an ideal setting for graduate students to network and get feedback for their work (as part of the doctoral forum) and everyone new to the field to learn about cutting-edge research by hearing outstanding invited speakers and attending presentations and tutorials (included with conference registration). A set of focused workshops is also held on the last day of the conference.

SDM22 has three main themes in Methods and Algorithms, Applications and Human Factors and Social Issues, each with a number of related topics. Each category contains a multitude of related themes as topics for papers. The proceedings of the conference are published in archival form and are made available on the SIAM web site, to be posted online April 2022.

IJCAI-ECAI 2022
The 31st International Joint Conference on Artificial Intelligence
Vienna, Austria
July 23-29, 2022
http://www.ijcai-21.org/

IJCAI-ECAI 2022 is the 31st International Joint Conference on Artificial Intelligence and the 23rd European Conference on Artificial Intelligence, the premier international gathering of researchers in AI. Starting from 1969, IJCAI has brought together the international AI community to communicate the advances and achievements of artificial intelligence research.

Following the tradition of previous IJCAIs, IJCAI-ECAI 2022 will feature the following tracks: workshops and tutorials; a doctoral consortium; the main technical program; an early career spotlight track; special tracks on "AI for Good" and on "AI, the Arts and Creativity"; survey, sister best paper and journal tracks; a demo, a video and a robot exhibition track; a diversity and inclusion program, as well as competitions and challenges.

Submissions to IJCAI-ECAI 22 should report on significant, original, and previously unpublished results on any aspect of artificial intelligence. Papers on novel AI research problems, on AI techniques for novel application domains, and papers that cross discipline boundaries within AI are especially encouraged. The submission site for IJCAI-ECAI 22 opens December 30, 2021.

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

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