PROGRAM GUIDE

The 4th APWeb-WAIM International Joint Conference on Web and Big Data (APWeb-WAIM 2020)

September 18-20, 2020

Tianjin, China

Organizing Committee

Honorary Chairs

- Masaru Kitsuregawa, University of Tokyo, Japan
- Keqiu Li, Tianjin University, China

General Chairs

- Xiaofang Zhou, The University of Queensland, Australia
- Zhiyong Feng, Tianjin University, China

Program Committee Chairs

- Xin Wang, Tianjin University, China
- Rui Zhang, University of Melbourne, Australia
- Young-Koo Lee, Kyunghee University, Korea

Panel Chairs

- Bin Cui, Peking University, China
- Weining Qian, East China Normal University, China

Workshop Chairs

- Qun Chen, Northwestern Polytechnical University, China
- Jianxin Li, Deakin University, Australia

Tutorial Chairs

- Yunjun Gao, Zhejiang University, China
- Leong Hou U, University of Macau, Macau

Demo Chairs

- Xin Huang, Hong Kong Baptist University, Hong Kong
- Hongzhi Wang, Harbin Institue of Technology, China

Industry Chairs

- Feifei Li, University of Utah, USA & Alibaba
- Guoliang Li, Tsinghua University, China & Huawei, China

Publication Chairs

- Le Sun, Nanjing University of Information Science and Technology, China
- Yang-Sae Moon, Kangwon National University, Korea

Publicity Chairs

- Yi Cai, South China University of Technology, China
- Yoshiharu Ishikawa, Nagoya University, Japan
- Yueguo Chen, Renmin University of China, China

APWeb-WAIM Steering Committee Representative

- Yanchun Zhang, Victoria University, Australia

Schedule at a Glance

September 18, 2020, Friday

Time	Room 1	Room 2	Room 3	Room 4
8:30–9:00	Opening Ceremony			
9:00–10:20	Keynote 1 : The tragedy of the (Data) Commons **Prof. James Hendler**			
10:20–10:40	Coffee Break			
10:40–12:00	Keynote 2 : Towards efficient computation of network structural stability **Prof. Xuemin Lin**			
12:00–13:30				
13:30–15:40	Research Session 1 : Storage and Indexing	Research Session 2 : Data Mining 1	Research Session 3 : Data Management	Research Session 4 : Graph Data
15:40–15:50	Coffee Break			
15:50–18:00	Demo Session	Research Session 5: Data Mining 2	Research Session 6: Security, Privacy, and Trust	Research Session 7 : Neural Network Applications

September 19, 2020, Saturday

Time	Room 1	Room 2	Room 3	Room 4
9:00–10:20	Keynote 3 : Design and implementation of new database engine, OoODE **Prof. Masaru Kitsuregawa**			
10:20–10:40	Coffee Break			
10:40–12:00	Keynote 4 : Entity Linking and Data Privacy Protection for Spatiotemporal Data **Prof. Xiaofang Zhou**			
12:00–13:30				
13:30–15:40	Research Session 8 : Machine Learning	Research Session 9 : Knowledge Graph	Research Session 10 : Text Analysis	Research Session 11 : Information Extraction and Retrieval
15:40–15:50	Coffee Break			
15:50–18:00	Research Session 12: Machine Learning 2	Research Session 13: Recommender System	Research Session 14: Social Networks	Research Session 15: Spatial-Temporal Databases

September 20, 2020, Sunday

Time	Room 1	Room 2	Room 3	Room 4
9:00–12:00	Tutorial 1 : Neighborhood Query Processing and Surrounding Objects Retrieval in Spatial Databases	KGMA Workshop	SemiBDMA Workshop	
12:00–13:30				
14:00–17:00	Tutorial 2 : Distributed Graph Processing Systems	DeepLUDA Workshop		

Welcome Message from the General Chairs

On behalf of the Organizing Committee, it is our great pleasure to welcome you to The Fourth Asia Pacific Web (APWeb) and Web-Age Information Management (WAIM) Joint Conference on Web and Big Data (APWeb-WAIM 2020) and the beautiful city of Tianjin. Tianjin is a municipality direct under the Central Government, as well as an opening city. It's situated in the eastern part of the North China Plain, covering an area of 11,300 square km and with a population of six million.

APWeb and WAIM are two separate leading international conferences on research, development, and applications of Web technologies and database systems. Previous APWeb conferences were held in Beijing (1998), Hong Kong (1999), Xi'an (2000), Changsha (2001), Xi'an (2003), Hangzhou (2004), Shanghai (2005), Harbin (2006), Huangshan (2007), Shenyang (2008), Suzhou (2009), Busan (2010), Beijing (2011), Kunming (2012), Sydney (2013), Changsha (2014), Guangzhou (2015), and Suzhou (2016). Previous WAIM conferences were held in Shanghai (2000), Xi'an (2001), Beijing (2002), Chengdu (2003), Dalian (2004), Hangzhou (2005), Hong Kong (2006), Huangshan (2007), Zhangjiajie (2008), Suzhou (2009), Jiuzhaigou (2010), Wuhan (2011), Harbin (2012), Beidaihe (2013), Macau (2014), Qingdao (2015), and Nanchang (2016). Starting in 2017, the three conference committees agreed to launch a joint conference. The First APWeb-WAIM conference was held in Bejing (2017), the Second APWeb-WAIM conference was held in Macau (2018) and the Third APWeb-WAIM conference was held in Chengdu (2019). With the increased focus on big data, the new joint conference is expected to attract more professionals from different industrial and academic communities, not only from the Asia Pacific countries but also from other continents.

APWeb-WAIM 2020 will enable you to enjoy an outstanding program, exchange your ideas with leading researchers in various disciplines, and make new friends in the international science community. Some highlights include four keynote talks on the latest exciting topics of Web and big data, ranging from the fundamental topic of core database systems to the fast-growing artificial intelligence applications; a diverse range of tutorials and workshops; technical sessions with exciting talks and demonstrations, and social events.

We are grateful to the strong support of the Steering Committee of APWeb and WAIM, and we are honored to serve as General Chairs for such a unique joint conference. The conference would not have been possible without the dedication and the hard work of all members of the Organizing Committee. The Program Committee Chairs, Xin Wang (Tianjin University, China), Rui Zhang (University of Melbourne, Australia) and Young-Koo Lee (Kyunghee University, Korea) put tremendous effort into the creation of an exciting program. Many other individuals and organizations contributed to the success of this conference. We would like to acknowledge the efforts of Honorary Chairs (Masaru Kitsuregawa and Keqiu Li), Panel Chairs (Bin Cui and Weining Qian), Workshop Chairs (Qun Chen and Jianxin Li), Tutorial Chairs (Yunjun Gao and Leong Hou U), Demo Chairs (Xin Huang and Hongzhi Wang), Industry Chairs (Feifei Li and Guoliang Li), Publication Chairs (Le

Sun and Yang-Sae Moon) and Publicity Chairs (Yi Cai, Yoshiharu Ishikawa and Yueguo Chen).

In addition to members of the Organization Committee, many volunteers have contributed to the success of the conference. Volunteers helped in editing this conference booklet, and helped with local arrangements and on-site setups, and many other important tasks. While it is difficult to list all their names here, we would like to take this opportunity to sincerely thank them all.

Last but not least, we would like to extend our most sincere congratulations to all authors and speakers for a job well done. We look forward to welcoming you in person, and we hope that you will enjoy APWeb-WAIM 2020!

General Chairs

Xiaofang Zhou
The University of Queensland, Australia
Zhiyong Feng
Tianjin University, China

Welcome Message from the Program Committee Chairs

On behalf of the APWeb-WAIM 2020 Program Committee, we are delighted to welcome you to Tianjin! For more than 20 years in the past, Asia-Pacific Web (APWeb) and Web-Age Information Management (WAIM) have attracted professionals of different communities related to Web and big data who have common interests in interdisciplinary research to share and exchange ideas, experiences, and the underlying techniques and applications, including Web technologies, database systems, information management, software engineering, and big data.

The technical program APWeb-WAIM 2020 features four keynotes by Prof. James Hendler (Rensselaer Polytechnic Institute, USA), Prof. Xuemin Lin (The University of New South Wales, Australia), Prof. Masaru Kitsuregawa (The University of Tokyo, Japan), and Prof. Xiaofang Zhou (The University of Queensland, Australia), as well as two tutorials by Dr. Md. Saiful Islam (Griffith University, Australia), and Prof. Yanfeng Zhang (Northeastern University, China), Ph.D. Shufeng Gong (Northeastern University, China) and Ph.D. Qiange Wang (Northeastern University, China). We are grateful to these distinguished scientists for their invaluable contributions to the conference program.

Our gratitude goes to Program Committee members and external reviewers whose technical expertise and dedication were not only thorough and crucial for the technical assessment of the selection of papers, but also inspirational in making the whole process even more pleasurable. During the double-blind review process, each paper submitted to APWeb-WAIM 2020 received at least three high quality review reports. Based on the obtained reviews, our Senior Program Committee members provided recommendations for each paper so that the difficult task of making decisions for acceptance could be performed. Finally, out of 259 submissions in total, the conference accepted 68 regular (26%), 29 short research papers, and 8 demonstrations. The contributed papers address a wide range of topics, such as storage and indexing, data mining, data management, graph data, security, privacy, and trust, neural network applications, machine learning, knowledge graph, text analysis, information extraction and retrieval, recommender system, social networks, spatial-temporal databases. Amongst a number of highly rated manuscripts, several candidates for best papers have been shortlisted for awards, where the final selection will be decided during the conference. In particular, we would like to thank Springer for its cash sponsorship of the APWeb-WAIM 2020 Best Paper Award, which will be announced at the conference banquet.

In addition to the main conference program, we would also like to thank Zhuoming Xu (Hohai University, China), Saiful Islam (Griffith University, Australia), and Xin Wang (Tianjin University, China) for organizing The Third International Workshop on Knowledge Graph Management and Applications (KGMA 2020), Qun Chen (Northwestern Polytechnical University, China) and Jianxin Li (Deakin University, Australia) for organizing The Second International Workshop on Semi-structured Big Data Management and Applications (SemiBDMA 2020), and Tae-Sun Chung

(Ajou University, Korea) and Rize Jin (Tiangong University, China) for organizing The First International Workshop on Deep Learning in Large-scale Unstructured Data Analytics (DeepLUDA 2020), which are in conjunction with APWeb-WAIM 2020.

We thank the General Chairs Xiaofang Zhou and Zhiyong Feng for their patience and support, and Yanchun Zhang representing the Steering Committee of APWeb and WAIM for the guidance. Many thanks also to all the members of the Organizing Committee for their full support in preparation of the conference, especially with respect to Website, publications, registration and local arrangements, without which the conference would not be possible to be put together.

Finally, the high-quality program would not have been possible without the authors who chose APWeb-WAIM for disseminating their findings. We would like to thank our authors whose valuable and novel contributions are essential for both the continued success of APWeb-WAIM and the advancement of technology for humanity.

Program Committee Chairs

Xin Wang
Tianjin University, China
Rui Zhang
University of Melbourne, Australia
Young-Koo Lee
Kyunghee University, Korea

Keynotes

Keynote Speech I: The tragedy of the (Data) Commons

Time: 9:00-10:20, September 18, 2020, Friday

Abstract: The tragedy of the commons, first proposed by William Lloyd in 1833, is an economic problem in which every individual has an incentive to consume a resource at the expense of every other individual with no way to exclude anyone from consuming. It results in over consumption, under investment, and ultimately depletion of the resource. While the direct application of these principles to data seems like a bit of a reach, it becomes clear that data sharing risks much of the same problem - people wishing to protect their own data while having access to other people's. Motivation for sharing is thus weak, until incentives and policies are in place. However, now that these incentives and policies are coming into practice, the implementation will have high impact on the benefits. Thus, as we work to make the world a FAIRer place, we must consider how the way data, and especially metadata, is represented and shared.

James Hendler

Professor, Rensselaer Polytechnic Institute

Speaker Bio: James Hendler is the Director of the Institute for Data Exploration and Applications and the Tetherless World Professor of Computer, Web and Cognitive Sciences at RPI. He also is acting director of the RPI-IBM Artificial Intelligence Research Collaboration. Hendler has authored over 400 books, technical papers and articles in the areas of Semantic Web, artificial intelligence, agent-based computing and high-performance processing. Hendler is a Fellow of the AAAI, BCS, the IEEE, the AAAS and the ACM. He was the first computer scientist to serve on the Board of Reviewing editors for Science. In 2010, Hendler wwas selected as an "Internet Web Expert" by the US government. In 2013, he was appointed as the Open Data Advisor to New York State. In 2016, became a member of the National Academies Board on Research Data and Information and in 2018 became chair of the ACM's US technology policy committee and was elected a Fellow of the National Academy of Public Administration.

Keynote Speech II: Towards efficient computation of network structural stability

Time: 10:40-12:00, September 18, 2020, Friday

Abstract: With the emergence of large-scale networks, the study of network structural stability has recently received a great deal of attention in different areas such as social networks, the world wide web, and biology. The stability of a network indicates the ability of the network to maintain an acceptable level of service and/or to defend the attacks from the competitors. In this talk, we first introduce the stability models in different domains, their applications, and unique challenges that need to be addressed. We focus on three fundamental problems: (a) efficiently computing the stability of a given network, (b) motivating critical nodes and edges to enhance the network stability, and (c) defending critical nodes and edges against the attacks to network stability. Due to the fast evolvement of real-life networks, we also discuss the stability problems and their computation on dynamic graphs. We will explore the nature and lay the scientific foundation of these problems. Subsequently, we introduce novel computing paradigms and algorithms, indexing techniques, batch processing techniques, and distributed solutions. Finally, we discuss the future research directions in this important and growing research area.

Xuemin Lin

Professor, The University of New South Wales

Speaker Bio: Xuemin Lin is a UNSW distinguished Professor - Scientia Professor, and the head of database and knowledge research group in the school of computer science and engineering at UNSW. Xuemin is a distinguished visiting Professor at Tsinghua University and visiting Chair Professor at Fudan University. He is a fellow of IEEE. Xuemin's research interests lie in databases, data mining, algorithms, and complexities. Specifically, he is working in the areas of scalable processing and mining of large scale data, including graph, spatial-temporal, streaming, text and uncertain data.

Xuemin currently serves as the editor-in-Chief of IEEE Transactions on Knowledge and Data Engineering (Jan 2017-now). He was an associate editor of ACM Transactions Database Systems (2008-2014) and IEEE Transactions on Knowledge and Data Engineering (Feb 2013- Jan 2015), and an associate editor-in-Chief of IEEE Transactions on Knowledge and Data Engineering (2015-2016), respectively. He has been regularly serving as a PC member and area chairs/SPC in SIGMOD, VLDB, ICDE, ICDM, KDD, CIKM, and EDBT. He is a PC co-chair of ICDE2019 and VLDB2022.

Keynote Speech III: Design and implementation of new database engine, OoODE

Time: 9:00-10:20, September 19, 2020, Saturday

Abstract: I will talk about the database engine which we have developed for more than 10 years. Out of order execution will be explained. After the end of Moore's law, performance of single processor core will stop. This talk will cover how we can design new type database engine which can handle many many core machines.

Masaru Kitsuregawa

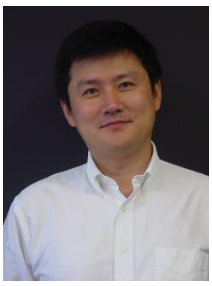
Professor, The University of Tokyo

Speaker Bio: Director General of National Institute of Informatics and Professor at Institute of Industrial Science, the University of Tokyo. Received Ph.D. degree from the University of Tokyo in 1983. Served in various positions such as President of Information Processing Society of Japan (2013-2015) and Chairman of Committee for Informatics, Science Council of Japan(2014-2016). He has wide research interests, especially in database engineering. He has received many awards including ACM SIGMOD E. F. Codd Innovations Award, IEICE Contribution Award, IPSJ Contribution Award, 21st Century Invention Award of National Commendation for Invention, Japan and C&C Prize, IEICE Contribution Award, IEEE Innovation in Societal Infrastructure Award and Japan Academy Award. In 2013, he awarded Medal with Purple Ribbon and in 2016, the Chevalier de la Legion D'Honneur. He is a fellow of ACM, IEEE, IEICE, IPSJ and honorary member of CCF.

Keynote Speech IV: Entity Linking and Data Privacy Protection for Spatio-Temporal Data

Time: 10:40-12:00, September 19, 2020, Saturday

Abstract: Spatial trajectory analytics involves a wide range of research topics including data management, query processing, data mining and recommendation systems. It can find many applications in intelligent transport systems, social media analysis, location-based systems, urban planning and smart city. New opportunities arise with massive and rapidly increasing volumes of high-quality spatio-temporal data from many sources such as GPS devices, mobile phones and social network applications. Integrating trajectory data is a fundamental step for making sense of spatio-temporal data. In this talk we will discuss our recent work on spatio-temporal entity linking and privacy protection for moving objects data.



Xiaofang Zhou

Professor, The University of Queensland

Speaker Bio: Professor Xiaofang Zhou is a Professor of Computer Science at The University of Queensland. His research focus is to find effective and efficient solutions for managing, integrating and analyzing very large amount of complex data for business, scientific and personal applications. He has been working in the area of spatial and multimedia databases, data quality, high performance database systems, data mining, streaming data analytics and recommendation systems. He is a Program Committee Chair for PVLDB 2020, SSTD 2017, CIKM 2016, ICDE 2013, and a General Chair of MDM 2018 and ACM Multimedia 2015. He has been an Associate Editor of The VLDB Journal, IEEE Transactions on Knowledge and Data Engineering, IEEE Transactions on Cloud Computing, World Wide Web Journal, Distributed and Parallel Databases, and IEEE Data Engineering Bulletin. He was the Chair of IEEE Technical Committee on Data Engineering (2015-2018), and a Fellow of IEEE.

Tutorials

Tutorial I: Neighborhood Query Processing and Surrounding Objects Retrieval in Spatial Databases

Time: 9:00-12:00 September 20, 2020, Sunday

Abstract: A nearest neighbourhood query (NHQ) retrieves the closest group of collocated objects from a spatial database for a given query location. On the other hand, a reverse nearest neighborhood query (RNHQ) returns all groups of collocated objects that find the given query nearer than any other alternatives. Both NHQ and RNHQ queries might have many practical applications including on demand facility placement and smart urban planning. This tutorial also introduces another query, called direction-based spatial skyline query (DSQ), for retrieving surrounding objects from a spatial database for a given user location. The retrieved objects are not dominated by other data objects in the same direction w.r.t. the query. A DSQ query is not only rotationally invariant, but also fair and stable. Like NHQ and RNHQ queries, retrieval of surrounding objects has many applications such nearby point-of-interests retrieval surrounding a user and digital gaming. This tutorial presents the challenges, algorithms, data indexing and data pruning techniques for processing NHQ, RNHQ and DSQ queries in spatial databases. Encouraging experimental results and future research directions in NHQ, RNHQ, DSQ queries and their variants are discussed.

Md. Saiful Islam
Doctor, Griffith University

Speaker Bio: Dr. Md. Saiful Islam is a Lecturer in the School of Information and Communication Technology, Griffith University, Australia. Before joining Griffith, he was a Research Fellow in the School of Engineering and Mathematical Sciences, La Trobe University, Australia from May 2016 to January 2017 and Research Associate in the Department of Computer Science and Software Engineering at Swinburne University of Technology from November 2013 to April 2016. He has finished his Ph.D. in Computer Science and Software Engineering from Swinburne University of Technology, Australia in February 2014. He has received his BSc (Hons) and MS degree in Computer Science and Engineering from University of Dhaka, Bangladesh, in 2005 and 2007, respectively. He has received the Faculty of Information and Communication Technologies Dean's Award Research Excellence (2nd Prize), Swinburne University of Technology in 2013. He has received the best paper awards in ACM SSDBM 2017 (as first author), DASFAA 2019 (as co-author) and ADMA 2019 (as co-author). He has published more than 50 research papers in prestigious computer science journals such as The VLDB Journal, IEEE TKDE, Elsevier Information Systems, Journal of Systems and Software, Future Generation Computer Systems, Journal of Network and Computer Applications, Springer World Wide Web and MDPI Sensors, and conferences such as IEEE ICDE, ACM CIKM, ACM SSDBM, IJCNN, IEEE CEC, DASFAA and WISE. He is a regular

reviewer for the top journals such as IEEE TKDE, The VLDB Journal, IEEE TFS, IEEE TPDS, IEEE TII, Elsevier Knowledge-based Systems and Future Generation Computer Systems. He was a program committee member and associate reviewer of many top computer science conferences such as IEEE ICDE, SIGMOD, SSDBM, DASFAA and ADMA, and senior program committee member for APWeb-WAIM 2020. He is a co-guest editor for the special issue “IoT and Artificial Intelligence Approaches to Defeat COVID-19 Outbreak” for MDPI Sensors. His current research interests are in the areas of database usability, spatial and graph data management, artificial intelligence, deep learning, health informatics, human-in-the-loop and big data analytics.

Tutorial II: Distributed Graph Processing Systems

Time: 14:00-17:00 September 20, 2020, Sunday

Abstract: During the past 10 years, there has been a surging interest in developing distributed graph processing systems. This tutorial provides a comprehensive review of existing distributed graph processing systems. We firstly review the programming models for distributed graph processing and then summarize the common optimization techniques for improving graph execution performance, including graph partitioning methods, communication mechanisms, parallel processing models, hardware-specific optimizations, and incremental graph processing. We also present an emerging hot topic, distributed Graph Neural Networks (GNN) frameworks, and review recent progress on this topic.

Yanfeng Zhang

Professor, Northeastern University

Speaker Bio: Yanfeng Zhang is a Professor at the College of Computer Science and Engineering, Northeastern University, Shenyang, China. He received the Ph.D. degree in computer science from Northeastern University in 2012 and spent 3 years at UMass Amherst as a visiting student during his Ph.D. study. His primary research areas are large-scale graph processing, asynchronous distributed computation, and big data management. He has published many technical papers in the above areas. His paper on "prioritized iteration" in ACM SoCC 2011 was honored with “Paper of Distinction”.

Shufeng Gong

Philosophic Doctor, Northeastern University

Speaker Bio: Shufeng Gong is currently a Ph.D. Candidate in Northeastern University, China. He received the B.S. degree in computer science from Harbin Normal University and the M.S. degree in computer science from Northeastern University of China. His current research interests include distributed graph computation and large-scale data mining.

Qiange Wang

Philosophic Doctor, Northeastern University

Speaker Bio: Qiange Wang is currently a Ph.D. Candidate in Northeastern University, China. He received the B.S. degree in computer science from Northeastern University of China. His current research interests include distributed graph computation and big data management.

Conference Sessions

Research Session 1: Storage and Indexing

Time: 13:30-15:40, September 18, 2020, Friday
Chair: Zhenying He, Fudan University

Index-Based Scheduling for Parallel State Machine Replication

Guodong Zhao[1], Gang Wu[1,2], Yidong Song[1], Baiyou Qian[1], Donghong Han[1]
[1]Northeastern University, [2]Naning University

Abstract. State Machine Replication is a fundamental approach to designing web services with fault tolerance. However, its requirement for the deterministic execution of transactions often results in single-threaded replicas, which cannot fully exploit the multicore capabilities of today's processors. Therefore, parallel SMR has become a hot topic of recent research. The basic idea behind it is that independent transactions can be executed in parallel, while dependent transactions must be executed in their relative order to ensure consistency among replicas. The dependency detection of existing parallel SMR methods is mainly based on pairwise transaction comparison or batch comparison. These methods cannot simultaneously guarantee both effective detection and concurrent execution. Moreover, the scheduling process cannot execute concurrently, which introduces extra scheduling overhead as well. In order to further reduce scheduling overhead and ensure the parallel execution of transactions, we propose an efficient scheduler based on a specific index structure. The index is composed of a Bloom Filter and the associated transaction queues, which provides an efficient dependency detection and preserve necessary dependency information respectively. Based on the index structure, we further devise an elaborated concurrent scheduling process. The experimental results show that the proposed scheduler is more efficient, scalable and robust than the comparison methods.

GHSH: Dynamic Hyperspace Hashing on GPU

Zhuo Ren, Yu Gu, Chuanwen Li, Fangfang Li, Ge Yu
Northeastern University

Abstract. Hyperspace hashing which is often applied to NoSQL data-bases builds indexes by mapping objects with multiple attributes to a multidimensional space. It can accelerate processing queries of some secondary attributes in addition to just primary keys. In recent years, the rich computing resources of GPU provide opportunities for implementing high-performance HyperSpace Hash. In this study, we construct a fully concurrent dynamic hyperspace hash table for GPU. By using atomic operations instead of locking, we make our approach highly parallel and lock-free. We propose a special concurrency control strategy that ensures wait-free read operations. Our data structure is designed considering GPU specific hardware characteristics. We also propose a warp-level pre-combinations data sharing strategy to obtain high parallel acceleration. Experiments on an Nvidia RTX2080Ti GPU suggest that GHSH performs about 20-100X faster than its counterpart on CPU. Specifically, GHSH performs updates with up to 396 M updates/s and processes

search queries with up to 995 M queries/s. Compared to other GPU hashes that cannot conduct queries on non-key attributes, GHSH demonstrates comparable building and retrieval performance.

An Index Method for the Shortest Path Query on Vertex Subset for the Large Graphs
Zian Pan, Yajun Yang, Qinghua Hu
Tianjin University, China
Abstract. Shortest path query is an important problem in graphs and has been well-studied. In this paper, we study a special kind of shortest path query on a vertex subset. Most of the existing works propose various index techniques to facilitate shortest path query. However, these indexes are constructed for the entire graphs, and they cannot be used for the shortest path query on a vertex subset. In this paper, we propose a novel index named pb-tree to organize various vertex subsets in a binary tree shape such that the descendant nodes on the same level of pb-tree consist of a partition of their common ancestors. We further introduce how to calculate the shortest path by pb-tree. The experimental results on three real-life datasets validate the efficiency of our method.

EPUR: An Efficient Parallel Update System over Large-Scale RDF Data
Xiang Kang, Pingpeng Yuan, Hai Jin
Huazhong University of Science and Technology
Abstract. RDF is a standard model for data interchange on the web and is widely adopted for graph data management. With the explosive growth of RDF data, how to process RDF data incrementally and maximize the parallelism of RDF systems has become a challenging problem. The existing RDF data management researches mainly focus on parallel query, and rarely pay attention to the optimization of data storage and update. Also, the conventional parallel models for parallel query optimizations are not suitable for data update. Therefore, we propose a new design of an efficient parallel update system which is novel in three aspects. Firstly, the proposed design presents a new storage structure of RDF data and two kinds of indexes, which facilitates parallel processing. Secondly, the new design provides a general parallel task execution framework to maximize the parallelism of the system. Last but not least, parallel update operations are developed to handle incremental RDF data. Based on the innovations above, we implement an efficient parallel update system (EPUR). Extensive experiments show that EPUR outperforms RDF-3X, Virtuoso, PostgreSQL and achieves good scalability on the number of threads.

Research Session 2: Data Mining 1

Time: 13:30-15:40, September 18, 2020, Friday
Chair: Hui Li, Xiamen University

MLND: A weight-adapting Method for Multi-label Classification based on Neighbor Label Distribution
Lei Yang, Zhan Shi, Dan Feng, Wenxin Yang, Jiaofeng Fang, Shuo Chen, Fang Wang

Huazhong University of Science and Technology
Abstract. In multi-label classification, each training sample is associated with a set of labels and the task is to predict the correct set of labels for the unseen instance. Learning from the multi-label samples is very challenging due to the tremendous number of possible label sets. Therefore, the key to successful multi-label learning is exploiting the label correlations effectively to facilitate the learning process. In this paper, we analyze the limitations of existing methods that add label correlations and propose MLND, a new method which extracts the label correlations from neighbors. Specifically, we take neighbor's label distribution as new features of a instance and obtain the label's confidence according to the new features. Nevertheless, the neighbor information is unreliable when the intersection of nearest neighbor samples is small, so we use information entropy to measure the uncertainty of the neighbor information and combine the original instance features with the new features to perform multi-label classification. Experiments on three different real-world multi-label datasets validate the effectiveness of our method against other state-of-the-art methods.

Multi-task Attributed Graphical Lasso

Yao Zhang[1], Yun Xiong[1], Xiangnan Kong[2], Xinyue Liu[2], Yangyong Zhu[1]
[1]Fudan University, [2]Worcester Polytechnic Institute
Abstract. Sparse inverse covariance estimation, i.e., Graphical Lasso, can estimate the connections among a set of random variables basing on their observations. Recent research on Graphical Lasso has been extended to multi-task settings, where multiple graphs sharing the same set of variables are estimated collectively to reduce variances. However, different tasks usually involve different variables. For example, when we want to estimate gene networks w.r.t different diseases simultaneously, the related gene sets vary. In this paper, we study the problem of multitask Graphical Lasso where the tasks may involve different variable sets. To share information across tasks, we consider the attributes of variables and assume that the structures of graphs are not only determined by observations, but influenced by attributes. We formulate the problem of learning multiple graphs jointly with observations and attributes, i.e., Multi-task Attributed Graphical Lasso (MAGL), and propose an effective algorithm to solve it. We rely on the LogDet divergence to explore latent relations between attributes of the variables and linkage structures among the variables. Multiple precision matrices and a projection matrix are optimized such that the `1-penalized negative log-likelihood and the divergence are minimized.

Characterizing Robotic and Organic Query in SPARQL Search Sessions

Xinyue Zhang[1], Meng Wang[1], Bingchen Zhao[2], Ruyang Liu[1], Jingyuan Zhang[1], Han Yang[3]
[1]Southeast University, [2]Tongji University, [3]Peking University
Abstract. SPARQL, as one of the most powerful query languages over knowledge graphs, has gained significant popularity in recent years. A large amount of SPARQL query logs have become available and provided new research opportunities to discover user interests, understand query intentions, and model search behaviors. However, a significant portion of the queries to SPARQL endpoints on the Web are robotic queries that are generated by automated scripts. Detecting and

separating these robotic queries from those organic ones issued by human users is crucial to deep usage analysis of knowledge graphs. In light of this, in this paper, we propose a novel method to identify SPARQL queries based on session-level query features. Specifically, we define and partition SPARQL queries into different sessions. Then, we design an algorithm to detect loop patterns, which is an important characteristic of robotic queries, in a given query session. Finally, we employ a pipeline method that leverages loop pattern features and query request frequency to distinguish the robotic and organic SPARQL queries. Differing from other machine learning based methods, the proposed method can identify the query types accurately without labelled data. We conduct extensive experiments on six real-world SPARQL query log datasets. The results demonstrate that our approach can distinguish robotic and organic queries effectively and only need 7.63 × 10-4 seconds on average to process a query.

NSTI-IC: An Independent Cascade Model based on Neighbor Structures and Topic-aware Interests

Chuhan Zhang, Yueshuang Yin, Yong Liu
HeiLongJiang University

Abstract. With the rapid development of social networks, discovering the propagation mechanism of information has become one of the key issues in social network analysis, which has attracted great attention. The existing propagation models only take into account individual influence between users and their neighbors, ignoring that different topologies formed by neighbors will have different influence on the target user. In this paper, we combine the influence of neighbor structure on different topics with the distribution of user interest on different topics, propose an propagation model based on structure influence and topic-aware interest, called NSTI-IC. We use an expectation maximization algorithm and a gradient descent algorithm to learn parameters of NSTI-IC. The experimental results on real datasets show that NSTI-IC model is superior to classical IC and structInf-IC models in terms of MSE and accuracy.

Temporal Knowledge Graph Incremental Construction Model for Recommendation

Chunjing Xiao, Leilei Sun, Wanlin Ji
Civil Aviation University of China

Abstract. Knowledge graph(KG) has been proven to be effective to improve the performance of recommendation because of exploiting structural and semantic paths information in a static knowledge base. However, the KG is an incremental construction process with interactions occurring in succession. Although some works have been proposed to explore the evolution of knowledge graph, which updates the entity representations by considering the previous interactions of related entities. However, we believe that the semantic path information between the involved entities and the occurring interaction itself also can refine their representations. To this end, we propose a temporal knowledge graph incremental construction model, which updates the entity representations by considering interaction itself and high-order semantic paths information. Specifically, different length semantic paths between user and item are automatically extracted when

an interaction occurs. Then we respectively employ recurrent neural network and standard multi-layer perceptron(MLP) to capture different length path semantic information and interaction itself information for updating the entity representations. Finally, we use MLP to predict the probability that a user likes an item after seamlessly integrating these variations into a unified representation. We conduct experiments on real-world datasets to demonstrate the superiority of our proposed model over all state-of-the-art baselines.

meanNet: A Multi-layer Label Mean based Semi-supervised Neural Network Approach for Credit Prediction

Guowei Wang[1], Lin Li[1], Jianwei Zhang[2]
[1]Wuhan University of Technology, [2]Iwate University

Abstract. Currently, semi-supervised deep learning usually combines supervised and unsupervised way to train its model, which intends to make good use of the information of unlabeled data. When applying semi-supervised learning in credit prediction, the distribution of credit data has its own characteristics. It is observed that there are multiple data-dense divisions even for one class because credit prediction needs to be considered from multiple perspectives. We argue that utilizing this information can improve the performance of semi-supervised learning. In this paper, we propose a novel multi-layer label mean based semi-supervised deep learning for credit prediction which is called meanNet. Our multi-layer structure approach takes into consideration class center points in different layers. We estimate the class center points of each class and the goal of multi-layer label mean is to maximize the distance of class center points at each layer. In addition, we add the cost-sensitive loss function to meanNet for the inconsistent misclassification cost between classes of credit datasets. Experiments are conducted on two public financial datasets and the results show that our approach can improve the credit prediction performance compared with popular baselines.

Author Contributed Representation for Scholarly Network

Binglei Wang, Tong Xu, Hao Wang, Yanmin Chen, Le Zhang, Lintao Fang, GuiQuan Liu, Enhong Chen
University of Science and Technology of China

Abstract. Scholarly network analysis is a fundamental topic in academia domain, which is beneficial for estimating the contribution of researchers and the quality of academic outputs. Recently, a popular fashion takes advantage of network embedding techniques, which aims to learn the scholarly information into vectorial representations for the task. Though great progress has been made, existing studies only consider the text information of papers for scholarly network representation, while ignoring the effects of many intrinsic and informative features, especially the different influences and contribution of authors and cooperations. In order to alleviate this problem, in this paper, we propose a novel Author Contributed Representation for Scholarly Network (ACR-SN) framework to learn the unique representation for scholarly networks, which characterizes the different authors' contribution. Specifically, we first adopt a graph convolutional network (GCN) to capture the structure information in the citation network. Then, we calculate the correlations between

authors and each paper, and aggregate each embedding of authors according to their contribution by using the attention mechanism. Extensive experiments on two real world datasets demonstrate the effectiveness of ACR-SN and reveal that authors' contribution to the paper varies with the corresponding authorities and interested fields.

Research Session 3: Data Management

Time: 13:30-15:40, September 18, 2020, Friday
Chair: Yajun Yang, Tianjin University

Quantitative Contention Generation for Performance Evaluation on OLTP Databases
Chunxi Zhang[1], Rong Zhang[1], Weining Qian[1], Ke Shu[2], Aoying Zhou[1]
[1]East China Normal University, [2]PingCAP Ltd.
Abstract. Although we have achieved significant progress in improving the scalability of transactional database systems (OLTP), the presence of contention operations in workloads is still the fundamental limitation in improving throughput. The reason is that the overhead of managing conflict transactions with concurrency control mechanism is proportional to the amount of contentions. As a consequence, contention workload generation is urgent to evaluate performance of modern OLTP database systems. Though we have kinds of standard benchmarks which provide some ways in simulating resource contention, e.g. skew distribution control of transactions, they can not control the generation of contention quantitatively; even worse, the simulation effectiveness of these methods is affected by the scale of data. So in this paper we design a scalable quantitative contention generation method with fine contention granularity control, which is expected to generate resource contention specified by contention ratio and contention intensity.

Evaluating Fault Tolerance of Distributed Stream Processing Systems
Xiaotong Wang[1], Cheng Jiang[1], Junhua Fang[2], Ke Shu[3], Rong Zhang[1], Weining Qian[1], Aoying Zhou[1]
[1]East China Normal University, [2]Soochow University, [3]PingCAP Ltd.
Abstract. Since failures in large-scale clusters can lead to severe performance degradation and break system availability, fault tolerance is critical for distributed stream processing systems (DSPSs). Plenty of fault tolerance approaches have been proposed over the last decade. However, there is no systematic work to evaluate and compare them in detail. Previous work either evaluates global performance during failure-free run-time, or merely measures throughout loss when failure happens. In this paper, it is the first work proposing an evaluation framework customized for quantitatively comparing runtime overhead and recovery efficiency of fault tolerance mechanisms in DSPSs. We define three typical configurable workloads, which are widely-adopted in previous DSPS evaluations. We construct five workload suites based on three workloads to investigate the effects of different factors on fault tolerance performance. We carry out extensive experiments on two well-known open-sourced DSPSs. The results demonstrate performance gap of two systems, which is useful for choice and evolution of fault tolerance approaches.

Pipelined Query Processing using Non-Volatile Memory SSDs

Xinyu Liu, Yu Pan, Wenxiu Fang, Rebecca J. Stones, Gang Wang, Yusen Li, Xiaoguang Liu
Nankai University

Abstract. NVM Optane SSDs are faster than traditional flash-based SSDs and more economical than DRAM main memory, so we explore query processing with the inverted index on NVM aiming at reducing costs, but this leads to NVM-to-DRAM I/O which negatively affects the search engine's responsiveness. To alleviate this problem, we propose a pipelining scheme to overlap CPU computation with NVM-to-DRAMI/O. We further propose some optimizations: variable coalesced block size, data prefetching, and block skipping. The experiments on the Gov2 and ClueWeb document corpuses indicate a reduction in CPU waiting time caused by NVM-to-DRAM I/O byaround 85% for Maxscore, Wand, and BlockMaxW and queries vs. not using pipelining, while maintaining comparable query throughput (loss within 6%) vs. an in-memory inverted index (DRAM-based scheme). For RankAnd queries, we occupy 3% of the inverted index in memory for caching to achieve similar query efficiency (within 6%) vs. the DRAM-based scheme.

Tool Data Modeling Method Based On An Object Deputy Model

Qianwen Luo, Chen Chen, Song Wang, Rongrong Li, Yuwei Peng
Wuhan University

Abstract. With the development of intelligent manufacturing industry, the management of tool data in machine tool processing is becoming more and more important. Due to the richness of machine tool data and the complexity of relationships in them, it's hard for a traditional relational database to manage the tool data. Therefore, a new method should be proposed to manage these data in a better way. In this work, we propose a tool data modeling method based on the object deputy model, which utilizes the characteristics of the class and the objects to express the meaning of the tool data and the various semantic constraints on them. Unlike the traditional relational model, objects are connected with a two-way pointer in the object deputy model where an object can have one or more deputy objects that inherit the properties and methods of the source object, and the deputy objects can have their own properties and methods. Besides, the two-way pointer between the source class and its deputy class makes the cross-class query easier in two aspects: One is to make complex queries expressed in intuitive statements, and the other is to improve query efficiency. We implemented and evaluated our model on an object deputy database. Experiments show that our method is better than the traditional relational ones.

Parallel Variable-Length Motif Discovery in Time Series using Subsequences Correlation

Chuitian Rong[1], Lili Chen[1], Chunbin Lin[2], Chao Yuan[1]
[1]Tiangong University, [2]Amazon AWS

Abstract. The repeated patterns in a long time series are called as time series motifs. As the motifs can reveal much useful information, time series motif discovery has been received extensive attentions in recent years. Time series motif discovery is an important operation for time series analysis in many fields, such as financial data analysis, medical and health monitoring. Although

many algorithms have been proposed for motifs discovery, most of existing works are running on single node and focusing on finding fixed-length motifs. They cannot process very long time series efficiently. However, the length of motifs cannot be predicted previously, and the Euclidean distance has many drawbacks as the similarity measure. In this work, we propose a parallel algorithm based on subsequences correlation called as PMDSC (Parallel Motif Discovery based on Subsequences Correlation), which can be applied to find time series motifs with variable lengths. We have conducted extensive experiments on public data sets,the results demonstrate that our method can efficiently find variable-length motifs in long time series.

Multi-grained Cross-modal Similarity Query with Interpretability

Mingdong Zhu[1], Derong Shen[2], Lixin Xu[1], Gang Ren[1]
[1]Henan Institute of Technology, [2]Northeastern University

Abstract. Cross-modal similarity query has become a highlighted research topic for managing multimodal datasets such as images and texts. Existing researches generally focus on query accuracy by designing complex deep neural network models, and hardly consider query efficiency and interpretability simultaneously, which are vital properties of cross-modal semantic query processing system on large-scale datasets. In this work, we investigate multi-grained common semantic embedding representations of images and texts, and integrate interpretable query index into the deep neural network by developing a novel Multi-grained Cross-modal Query with Interpretability (MCQI) frame-work. The main contributions are as follows: (1) By integrating coarse-grained and fine-grained semantic learning models, a multi-grained cross-modal query processing architecture is proposed to ensure the adaptability and generality of query processing. (2) In order to capture the latent semantic relation between images and texts, the framework combines LSTM and attention mode, which enhances query accuracy for the cross-modal query and constructs the foundation for interpretable query processing. (3) Index structure and corresponding nearest neighbor query algorithm are proposed to boost the efficiency of interpretable queries. Comparing with state-of-the-art methods on widely-used cross-modal datasets, the experimental results show the effectiveness of our MCQI approach.

Unsupervised Cross-Modal Retrieval by Coupled Dual Generative Adversarial Networks

Jingzi Gu[1,2], Peng Fu[1], Jinchao Zhang[1], Lulu Wang[1], Bo Li[1,2], Weiping Wang[1,2]
[1]Institute of Information Engineering, Chinese Academy of Sciences,
[2]School of Cyber Security, University of Chinese Academy of Sciences

Abstract. Textual-visual cross-modal retrieval has become a hot research topic in both computer vision and natural language processing communities. However, existing deep cross-modal hashing methods either rely on amounts of labeled information or have no ability to learn an accuracy correlation between different modalities. In this paper, we address the unsupervised cross-modal retrieval problem using a novel framework called coupled dual generative adversarial networks(CDGAN). This framework consists of two cycle networks: a text-to-image-to-text(t2t) network and an image-to-text-to-image(i2i) network. The t2t network is used to learn the relation

among an original text, the generated image and the generated text using the similarity of original and generated image-text, and the i2i network is used to learn the relation among an original image, the generated text and the generated image. Therefore, two groups of mixed hash codes of image-text are learned in this framework. Furthermore, our proposed CDGAN seamlessly couples these two cycle networks with generative adversarial mechanism so that the hash codes can be optimized simultaneously. Extensive experiments show that our framework can well match images and sentences with complex content, and it can achieve the state-of-the-art cross-modal retrieval results on two popular benchmark datasets.

Research Session 4: Graph Data

Time: 13:30-15:40, September 18, 2020, Friday
Chair: Meng Wang, Southeast University

LOCATE: Locally Anomalous Behavior Change Detection in Behavior Information Sequence
Dingshan Cui[1], Lei Duan[1], Xinao Wang[1], Jyrki Nummenmaa[2], Ruiqi Qin[1], Shan Xiao[1]
[1]Sichuan University, [2]Tampere University
Abstract. With the availability of diverse data reflecting people's behavior, behavior analysis has been studied extensively. Detecting anomalies can improve the monitoring and understanding of the objects' (e.g., people's) behavior. This work considers the situation where objects behave significantly differently from their previous (past) similar objects.We call this locally anomalous behavior change. Locally anomalous behavior change detection is relevant to various practical applications, e.g.,detecting elderly people with abnormal behavior. In this paper, making use of objects, behavior and their associated attributes as well as the relations between them, we propose a behavior information sequence (BIS) constructed from behavior data, and design a novel graph information propagation autoencoder framework called LOCATE (locally anomalous behavior change detection), to detect the anomalies involving the locally anomalous behavior change in the BIS. Two real-world datasets were used to assess the performance of LOCATE. Experimental results demonstrated that LOCATE is effective in detecting locally anomalous behavior change.

Partition-Oriented Subgraph Matching on GPU
Jing Chen, Yu Gu, Qiange Wang, Chuanwen Li, Ge Yu
Northeasten University
Abstract. Subgraph isomorphismis a well known NP-hard problem that finds all the matched subgraphs of a query graph in a large data graph. The state-of-the-art GPU-based solution is the vertex-oriented joining strategy, which is proposed by GSI. It effectively solves the problem of parallel write conflicts by taking vertices as processing units. However, this strategy might result in load-imbalance and redundant memory transactions when dealing with dense query graph. In this paper, we design a new storage structure Level-CSR and a new partition-oriented joining strategy. To avoid the influence of vertices with large degrees, we divide the dense vertices in traditional CSR

into several GPU-friendly tasks and store them in Level-CSR. Then, an efficient execution strategy is designed based on the partitioned tasks. The partition strategy can improve the load imbalance caused by the irregularity of real-world graphs, and further reduce the redundant global memory access caused by the redundant neighbor set accessing. Besides, to further improve the performance, we propose a well-directed filtering strategy by exploiting a property of real-world graphs. The experiments show that compared with the state-of-the-art GPU based solutions, our approach can effectively reduce the number of unrelated candidates, minimize memory transactions, and achieve load balance between processors.

A Unified Framework for Processing Exact and Approximate Top-k Set Similarity Join

Cihai Sun[1,2], Hongya Wang[1], Yingyuan Xiao[3], Zhenyu Liu[4]
[1]Donghua University, [2]Shanghai University of International Business and Economics,
[3]Tianjin University of Technology
[4]Shanghai Key Laboratory of Computer Software Testing & Evaluation

Abstract. An interesting observation was made that only a few (far shorter than the prefix) low-frequency tokens are enough to help finding similarity pairs for processing top-k set joins. This phenomenon is ubiquitous in all real datasets we have experimented with, covering do-mains as varied as text, social network, protein sequence data. Possible explanations are discussed. Based on this observation, we propose an algorithm called AEtop-k for processing both approximate and exact top-k similarity join in a unified framework. Comprehensive experiments demonstrate that, compared with the state-of-the-art algorithm on a large collection of real-life datasets, the approximate version of our algorithm can achieve up to 10000x speedup with little sacrifice on accuracy and the exact version runs up to 5x faster than the existing algorithm.

GSimRank: A General Similarity Measure on Heterogenous Information Networks

Chuanyan Zhang, Xiaoguang Hong, Zhaohui Peng
Shandong University

Abstract. Measuring similarity of objects in information network is a primitive problem and has attracted many studies for widely applications, such as recommendation and information retrieval. With the advent of large-scale heterogeneous information network that consist of multi-type relationships, it is important to research similarity measure in such networks. However, most existing similarity measures are defined for homogeneous network and cannot be directly applied to HINs since different semantic meanings behind edges should be considered. This paper proposes GSimRank that is the extended form of the famous SimRank to compute similarity on HINs. Rather than summing all meeting paths for two nodes in SimRank, GSimRank selects linked nodes of the same semantic category as the next step in the pairwise random walk, which ensure the two meeting paths share the same semantic. Further, in order to weight the semantic edges, we propose a domain-independent edge weight evaluation method based on entropy theory. Finally, we proof that GSimRank is still based on the expected meeting distance model and provide experiments on two real world datasets showing the performance of GSimRank.

Natural Answer Generation via Graph Transformer

Xiangyu Li, Sen Hu, Lei Zou
Peking University

Abstract. Natural Answer Generation (NAG), which generates natural answer sentences for the given question, has received much attention in recent years. Compared with traditional QA systems, NAG could offer specific entities fluently and naturally, which is more user-friendly in the real world. However, existing NAG systems usually utilize simple retrieval and embedding mechanism, which is hard to tackle complex questions. They suffer issues containing knowledge insufficiency, entity ambiguity, and especially poor expressiveness during generation. To address these challenges, we propose an improved knowledge extractor to retrieve supporting graphs from the knowledge base, and an extending graph transformer to encode the supporting graph, which considers global and variable information as well as the communication path between entities. In this paper, we propose a framework called G-NAG, including a knowledge extractor, an incorporating encoder, and an LSTM generator. Experimental results on two complex QA datasets demonstrate the efficiency of G-NAG compared with state-of-the-art NAG systems and transformer baselines.

LSimRank: Node Similarity in a Labeled Graph

Yang Wu[1], Ada Wai-Chee Fu[1], Cheng Long[2], Zitong Chen[1]
[1]The Chinese University of Hong Kong, [2]Nanyang Technological University

Abstract. The notion of node similarity is useful in many real-world applications. Many existing similarity measurements such as SimRank and its variants have been proposed. Among these measurements, most capture the structural information of a graph only, and thus they are not suitable for graphs with additional label information. We propose a new similarity measurement called LSimRank which measures the similarities among nodes by using both the structural information and the label information of a graph. Extensive experiments on datasets verify that LSimRank is superior over SimRank and other variants on labeled graphs.

Research Session 5: Data Mining 2

Time: 15:50-18:00, September 18, 2020, Friday
Chair: Saiful Islam, Griffith University

A Method for Decompensation Prediction in Emergency and Harsh Situations

Guozheng Rao[1], Shuying Zhao[1], Li Zhang[2], Qing Cong[1], Zhiyong Feng[1]
[1]Tianjin University, [2]Tianjin University of Science and Technology

Abstract. To save more lives, critically ill patients need to make timely decisions or predictive diagnosis and treatment in emergency and harsh conditions, such as earthquakes, medical emergencies, and hurricanes. However, in such circumstances, medical resources such as medical staff and medical facilities are short supply abnormally. So, we propose a method for decompensation prediction in emergency and harsh conditions. The method includes components

such as patient information collection, data selection, data processing, and decompensation prediction. Based on this, this paper demonstrates the method using MIMIC-III data. Firstly, we tried a series of machine learning models to predict physiological decompensation. Secondly, to detect patients whose condition deteriorates rap-idly under severe and limited circumstances, we try to reduce the essential physiological variables as much as possible for prediction. The experimental results show that the Bi-LSTM-attention method, combined with eleven essential physiological variables, can be used to predict the decompensation of severe ICUs patients. The AUC-ROC can reach 0.8509. Furthermore, these eleven physiological variables can be easily monitored without the need for complicated manual and massive, costly instruments, which meets the real requirements under emergency and harsh conditions. In summary, our decompensation prediction method can provide intelligent decision support for saving more lives in emergency and harsh conditions.

Bayes Classifier Chain based on SVM for Traditional Chinese Medical Prescription Generation

Chaohan Pei[1,4], Chunyang Ruan[2], Yanchun Zhang[3,4], Yun Yang[5]
[1]Fudan University, [2]Shanghai International Studies University, [3]Victoria University,
[4]Guangzhou University, [5]Department of Oncology and Longhua Hospital

Abstract. Traditional Chinese Medicine(TCM) plays an important role in the comprehensive treatment of lung cancer. However the quality of the prescriptions from TCM doctors depends on the doctor's personal experience, which leads to the TCM prescriptions are the lack of standardization. We apply the original clinical TCM prescriptions data to train a standardized prescription generating model for TCM therapy. Our model adopts the Bayes Classifier Chain(BCC) algorithm to solve the label correlation problem, whose basic classifier is cost-sensitive SVM targeted to the class imbalance of the label. The results of experiments on the prescription dataset demonstrated the effectiveness and practicability of the proposed model for a prescription generation.

Learning Ability Community for Personalized Knowledge Tracing

Juntao Zhang, Biao Li, Wei Song, Nanzhou Lin, Xiandi Yang, Zhiyong Peng
Wuhan University

Abstract. Knowledge tracing is an essential task that estimates students' knowledge state as they engage in the online learning platform. Several models have been proposed to predict the state of students' learning process to improve their learning efficiencies, such as Bayesian Knowledge Tracing, Deep Knowledge Tracing, and Dynamic Key-Value Memory Networks. However, these models fail to fully consider the influence of students' current knowledge state on knowledge growth, and ignore the current knowledge state of students is affected by forgetting mechanisms. Moreover, these models are a unified model that does not consider the use of group learning behavior to guide individual learning. To tackle these problems, in this paper, we first propose a model named Knowledge Tracking based on Learning and Memory Process (LMKT) to solve the effect of students' current knowledge state on knowledge growth and forgetting mechanisms. Then we propose the definition of learning capacity community and personalized knowledge tracking. Finally, we present a novel method called Learning Ability Community for Personalized Knowledge Tracing

(LACPKT), which models students' learning process according to group dynamics theory. Experimental results on public data sets show that the LMKT model and LACPKT model are effective. Besides, the LACPKT model can trace students' knowledge state in a personalized way.

A Pruned DOM-Based Iterative Strategy for Approximate Global Optimization in Crowdsourcing Microtasks

Lizhen Cui, Jing Chen, Wei He, Hui Li, Wei Guo
ShanDong University

Abstract. Crowdsourcing can solve many challenging problems for machines. The ability and knowledge background of employees on the Internet are unknown and different, the answers collected from the crowd are ambiguous. The choice of employee quality control strategy is really important to ensure the crowdsourcing results. In previous works, Expectation-Maximization(EM) was mainly used to estimate the real answer and quality of workers. Unfortunately, EM provides a local optimal solution, and the estimation results are often affected by the initial parameters. In this paper, an iterative optimization method based on EM local optimal results is designed to improve the quality estimation of workers for crowdsourcing micro-tasks (which has binary answers). The iterative search method works on the dominance ordering model(DOM) we proposed, which prunes the dominated task-response sequences while preserving the dominating ones, to iteratively search for the approximate global optimal estimation in a reduced space. We evaluate the proposed approach through extensive experiments on both simulated and realworld datasets, the experimental results illustrate that this strategy has higher performence than EM-based algorithm.

Predicting Adverse Drug-Drug Interactions via Semi-Supervised Variational Autoencoders

Meihao Hou, Fan Yang, Lizhen Cui, Wei Guo
Shandong University

Abstract. Adverse Drug-Drug Interactions (DDIs) are a very important risk factor in the medical process, which may lead to readmission or death. Although a part of DDIs can be obtained through in vitro or in vivo experiments in the drug development stage, a large number of new DDIs still appear after the market, more and more researchers be-gin to pay attention to the research related to drug molecules, such as drug discovery, drug target prediction, DDIs prediction, etc. In recent years, many computational methods for predicting DDIs have been proposed. However, most of them only used labeled data and neglect a lot of information hidden in unlabeled data. Moreover, they always focus on binary prediction instead of multiclass prediction, although the exact DDI type is very helpful for our reasonable choice of medication. In this paper, a Semi-Surpervised Variational Autoencoders (SPRAT) method for predicting DDIs is proposed, which is composed of a neural network classifier and a Variational autoencoders (VAE). Classifier is the core components, VAE plays a role of calibration. In the end, the predicted label is a multi-hot vector which indicates specific DDI types between drug pairs. Finally, the experiments on real world dataset demonstrate the effectiveness of the proposed method in this paper.

Mining Affective Needs from Online Opinions for Design Innovation

Danping Jia, Jian Jin
Beijing Normal University

Abstract. Innovative product features may gain higher brand reputation with lower cost for companies. Besides functional features, products having differential advantages on aesthetic design are acknowledged to be attractive in the market. As a result, exploring customer affective needs plays a critical role in product design innovation. In this paper, a hybrid method is proposed to reveal and classify customer affective needs from online opinions, including customer affective emotions and related product features. Firstly, inspired by Kansei engineering (KE), a knowledge-based method is presented to extract customer affective emotions. Then, enlightened by Kano model which determines the priorities of product features based on their abilities in satisfying customers, affective features are automatically extracted and classified into Kano categories. Finally, empirical studies are investigated to evaluate the effectiveness of the proposed framework. Compared with others, this method achieves higher F-measure scores in different domains. It highlights that a data-driven integration of KE and Kano model brings novel ideas and advanced suggestions for product design and marketing management in the view of designers and manager.

Joint Learning-based Anomaly Detection on KPI Data

Yongqin Huang, Yijie Wang, Li Cheng
National University of Defense Technology

Abstract. Unsupervised anomaly detection on KPI (Key Performance Indicator) is an important research problem that has broad industrial applications. The dynamic change and infinity of KPI data make it a challenging problem to estimate the outlierness of KPI data. Existing methods detect anomalies only from a single perspective, which fails to capture the dynamic change of KPI data, leading to unsatisfying performance. In this paper, we propose a Joint Learning-based Anomaly Detection algorithm (JLAD) which integrates a bias-based model and a similarity-based model from two perspectives at the same time. Specifically, the similarity-based model motivates a data driven abnormal ratio setting which automatically obtains abnormal ratio for bias-based model to avoid manual setting. In turn, based on the bias-based model,we develop an automatic anomaly templates adding method that adds anomaly templates for similarity-based model to timely improve the prior knowledge. Experiments on 13 public KPI datasets empirically confirm the superiority of our algorithm with a F1-Score improvement up to 20.5% on average.

Research Session 6: Security, Privacy, and Trust

Time: 15:50-18:00, September 18, 2020, Friday
Chair: Junhu Wang, Griffith University

On the vulnerability and generality of K-anonymity location privacy under continuous LBS requests

Hanbo Dai[1], Hui Li[2], Xue Meng[2], Yingxue Wang[3]
[1]Hubei University, [2]Xidian University,
[3]National Engineering Laboratory for Public Safety Risk Perception and Control by Big Data
Abstract. With the development of personal communication devices, location-based services have been widely used. However, the risk of lo-cation information leakage is a fundamental problem that prevents the success for these applications. Recently, some location-based privacy protection schemes have been proposed, among which K-anonymity scheme is the most popular one. However, as we empirically demonstrated, these schemes may not preserve satisfactory effect in trajectory-aware scenarios. In particular, we propose a new attack model using public navigation services. According to the empirical results, the attack algorithm correlates a series of snapshots associated with continuous queries, eliminating some of the less likely routes, and seriously undermining the anonymity of the query, thereby increasing the probability of attack. In order to defend against the proposed attacks, two enhanced versions of K-anonymity mechanism are proposed for this attack model, which further protects the user's trajectory privacy.

Instance-aware Evaluation of Sensitive Columns in Tabular Dataset

Zheng Gong[1], Kechun Zhao[1], Hui Li[1], Yingxue Wang[2]
[1]Xidian University, [2]National Engineering Laboratory for Public Safety Risk Perception and Control by Big Data
Abstract. Fully discovering knowledge from big data has to publish and share corresponding datasets whenever required. However, the risk for privacy leakage, i.e., record re-identification through some released columns, in the datasets is a fatal problem that prevents these tasks. Therefore, evaluating the sensitivity for different attributes is a prerequisite for dataset desensitization and anonymization, after which datasets can be published and shared in a privacy-preserving way. However, automatically evaluating the sensitivity for attributes is challenging and remains an open problem. In this work, we present a novel-but-simple technique for quantifying the sensitivity in structural database. It automatically evaluates the risks for re-identification for different columns according to Record-linkage Attack. Under the support of our scheme, the output sensitivity for the same attribute in different instances of a relational schema varies. Moreover, our scheme can quantify the risks of the columns no matter the semantics of columns are known or not. We also empirically show that the proposed scheme is effective in dataset sensitivity governance comparing with baselines.

FedSmart: An Auto Updating Federated Learning Optimization Mechanism

Anxun He, Jianzong Wang, Zhangcheng Huang, Jing Xiao
Ping An Technology (Shenzhen) Co., Ltd
Abstract. Federated learning has made an important contribution to data privacy-preserving. Many previous works are based on the assumption that the data are independently identically distributed (IID). As a result, the model performance on non-identically independently distributed (non-IID) data is beyond expectation, which is the concrete situation. Some existing methods of ensuring the model

robustness onnon-IID data, like the data-sharing strategy or pre-training, may lead to privacy leaking. In addition, there exist some participants who try to poison the model with low-quality data. In this paper, a performance-based parameter return method for optimization is introduced, we term it Federated Smart (FedSmart). It optimizes different model for each client through sharing global gradients, and it extracts the data from each client as a local validation set, and the accuracy that model achieves in round t determines the weights of the next round. The experiment results show that FedSmart enables the participants to allocate a greater weight to the ones with similar data distribution.

Smarter Smart Contracts: Efficient Consent Management in Health Data Sharing

Mira Shah, Chao Li, Ming Sheng, Yong Zhang and Chunxiao Xing
Tsinghua University

Abstract. The healthcare industry faces serious problems in data fragmentation and insufficient data sharing between patients, healthcare service providers and medical researchers. At the same time, patients' privacy must be protected, and patients should have authority over who can access their data. Researchers have proposed blockchain-based solutions to health data sharing, using blockchain for consent management. However, the implementation of the smart contracts that underpin these solutions has not been studied in detail. In this paper, we develop a blockchain-based framework for consent management in interorganizational health data sharing. We study the design of smart contracts that support the operation of our framework and evaluate its efficiency based on the execution costs on Ethereum. Our design improves on those previously proposed, lowering the computational costs of the framework significantly. This allows the framework to operate at scale and is more feasible for widespread adoption. Additionally, we introduce a novel contract that supports searching for patients in the framework that match certain criteria. This feature would be useful to medical researchers looking to obtain patient data.

DHBFT: Dynamic Hierarchical Byzantine Fault-Tolerant Consensus Mechanism Based on Credit

Fengqi Li, Kemeng Liu, Jing Liu, Yonggang Fan, Shengfa Wang
Dalian University of Technology

Abstract. It is significant to improve Practical Byzantine Fault Tolerance algorithm (PBFT) in consortium blockchain. At present,the serial verification process of transactions in the primary and backups greatly affects consensus efficiency. Meanwhile, the lack of reasonable valuation mechanism in PBFT makes it difficult to motivate existing reliable nodes. Moreover, consensus nodes work in an enclosed environment, where nodes cannot join and exit dynamically.
To solve the shortcomings stated above, we propose a dynamic hierarchical Byzantine fault-tolerant consensus mechanism based on credit (DHBFT). Firstly, we design a hierarchical-parallel scheme composed of consensus nodes, candidate nodes, and ordinary nodes. We realize parallel transaction logic verification in the primary and backups by delegating candidate nodes to verify the validity of transactions preliminarily. Secondly, we create a reward-punishment scheme. The consensus nodes

with better performances are assigned higher credit value and have higher probability to become the primary. Thirdly, we propose a dynamic promotion-demotion scheme. It enables faulty nodes to be excluded from the consensus set and reliable candidate nodes to join.
Experimental results show that DHBFT has better efficiency and higher stability. Compared with PBFT, the overall throughput of transactions is increased by 16%, and the average delay is reduced by 12%. Moreover, the proportion of abnormal nodes is basically 0 and much lower than that of PBFT.

MaSRChain: A Trusted Manuscript Submission and Review System Based on Blockchain
Fengqi Li, Kemeng Liu, Haoyu Wu, Xu Zhang
Dalian University of Technology
Abstract. Manuscript submission and review (MaSR) systems play an important role in scholarly publishing. However, there are some problems to be solved. Authors cannot gain an authoritative copyright certificate of manuscripts. Journals and conferences cannot achieve effective detection of multiple contributions with one manuscript. Reviewers may intentionally submit malicious evaluations due to competition. In this paper, we propose a trusted decentralized manuscript sub-mission and review system based on blockchain (MaSRChain) to solve problems above. At first, we use blockchain and Attribute-Based Encryption (ABE) to protect manuscript copyright and realize access control of manuscripts for authors. Secondly, we utilize blockchain to realize manuscript sharing that encrypted by Locality Sensitive Hash (LSH), which can achieve multiple contributions detection among different institutions. Thirdly, we apply Ring Signature to realize authentication of review evaluations, while providing some anonymity to reviewers. Finally, we conduct experiments based on Hyperledger Fabric and experimental results demonstrate the effectiveness and efficiency of the system.

Research Session 7: Neural Network Applications

Time: 15:50-18:00, September 18, 2020, Friday
Chair: Yingxia Shao, Beijing University of Posts and Telecommunications

Improved Brain Segmentation using Pixel Separation and Additional Segmentation Features
Afifa Khaled[1], Chung-Ming Own[1], Wenyuan Tao[1], Taher Ahmed Ghaleb[2]
[1]Tianjin University, [2]Queen's University
Abstract. Brain segmentation is key to brain structure evaluation for disease diagnosis and treatment. Much research has been invested to study brain segmentation. However, prior research has not considered separating actual brain pixels from the background of brain images. Not performing such separation may (a) distort brain segmentation models and (b) introduce overhead to the modeling performance. In this paper,we improve the performance of brain segmentation using3D, fully Convolutional Neural Network (CNN) models. We use (i) infant and adult datasets, (ii) a multi-instance loss method to separate actual brain pixels from the background and (iii) Gabor filter

banks and K-means clustering to provide additional segmentation features. Our model obtains dice coefficients of 87.4%-94.1% (i.e., an improvement of up to 11% to the results of five state-of-the-art models). Unlike prior studies, we consult experts in medical imaging to evaluate our segmentation results. We ob-serve that our results are fairly close to the manual reference. Moreover,we observe that our model is 1.2x-2.6x faster than prior models. We conclude that our model is more efficient and accurate in practice for both infant and adult brain segmentation.

Hyperthyroidism Progress Prediction with Enhanced LSTM

Haiqin Lu[1], Mei Wang[1], Weiliang Zhao[1], Tingwei Su[2], Jian Yang[3]
[1]Donghua University, [2]Ruijin Hospital, School of Medicine Shanghai Jiao Tong University, [3]Macquarie University

Abstract. In this work, we propose a method to predict the progress of the hyperthyroidism disease based on the sequence of the patient's blood test data in the early stage. Long-Short-Term-Memory (LSTM) network is employed to process the sequence information in the tests. We design an adaptive loss function for the LSTM learning. We set bigger weights to the blood test data samples which are nearby the range boundaries when judging the hyperthyroidism. We have carried out a set of experiments against a real world dataset from a hospital in Shanghai, China. The experimental results show that our method outperforms the traditional LSTM significantly.

Leveraging Explicit Unsupervised Information for Robust Graph Convolutional Neural Network Learning

Chu Zheng, Peiyun Wu, Xiaowang Zhang, Zhiyong Feng
Tianjin University

Abstract. Most existing graph convolutional networks focus on utilizing supervised information for training semi-supervised graph learning. However, the inherent randomness of supervised information can reduce the robustness of graph convolutional network models in some cases. To cope with this problem, in this paper, we propose a novel semi-supervised graph representation learning method RUGCN by leveraging explicit unsupervised information into training. We first propose a practical training method to ensure unsupervised information measurable by preserving both unsupervised (ranking smoothing) and semi-supervised (Laplacian smoothing) information. And then, we introduce a broadcast cross-entropy function to ensure ranking smoothing run in harmony with Laplacian smoothing. Experiments show that RUGCN achieves competitive results and stronger robustness.

A Spatial and Sequential Combined Method for Web Service Classification

Xin Wang[1], Jin Liu[1], Xiao Liu[2], Xiaohui Cui[1], Hao Wu[3]
[1]Wuhan University, [2]Deakin University, [3]Yunnan University

Abstract. With the growing prosperity of the Web service ecosystem, high-quality service classification has become an essential requirement. Web service description documents are semantic definitions of services,which is edited by service developers to include not only usage scenarios and

functions of services but also a lot of prior knowledge and jargons. However, at present, existing deep learning models cannot fully extract the heterogeneous features of service description documents, resulting in unsatisfactory service classification results. In this paper, we propose a novel deep neural network which integrates the Graph Convolutional Network (GCN) with Bidirectional Long Short-Term Memory (Bi-LSTM) network to automatically extract the features of function description documents for Web services. Specifically, we first utilize a two-layer GCN to extract global spatial structure features of Web services,which serves as a pre-training word embedding process. Afterwards, the sequential features of Web services learned from the Bi-LSTM model areintegrated for joint training of parameters. Experimental results demonstrate that our proposed method outperforms various state-of-the-art methods in classification performance.

KASR: Knowledge-Aware Sequential Recommendation

Qingqin Wang[1], Yun Xiong[1], Yangyong Zhu[1], Philip S. Yu[2]
[1]Fudan University, [2]University of Illinois at Chicago

Abstract. The goal of sequential recommendations is to capture the transitions of users' interests. Most existing methods utilize sequential neural networks to model interaction records, mapping items into latent vectors. Although such methods do explore the transitions of items in interaction sequences, they only capture the sequence dependencies of items, neglecting the deep semantic relevance between items. Such limited information contributes less to catching the complicated sequential behaviors of users accurately. In this paper, we propose a novel model Knowledge-Aware Sequential Recommendation (KASR), which captures sequence dependencies and semantic relevance of items simultaneously in an end-to-end manner. Specifically, we first convert the interaction records into a knowledge-transfer interaction sequence, which reflects the fine-grained transitions of users' interests. Next, we further recursively aggregate information in the knowledge graph based on a specific relation attention network, to explicitly capture the high-order relevance between items. A knowledge-aware GRU is later introduced to explore the sequential and semantic relevance between items automatically. We have conducted extensive experiments on three real datasets, and the results demonstrate that our method outperforms the state-of-the-art models.

High Order Semantic Relations-based Temporal Recommendation model by Collaborative Knowledge Graph Learning

Yongwei Qiao, Leilei Sun, Chunjing Xiao
Civil Aviation University of China

Abstract. Knowledge graph (KG) as the source of side information has been proven to be useful to alleviate the data sparsity and cold start. Existing methods usually exploit the semantic relations between entities by learning structural or semantic paths information. However, they ignore the difficulty of information fusion and network alignment when constructing knowledge graph from different domains, and do not take temporal context into account. To address the limitations of existing methods, we propose a novel High-order semantic Relations-based Temporal Recommendation (HRTR), which captures the joint effects of high-order semantic relations in

Collaborative Knowledge Graph (CKG) and temporal context. Firstly, it automatically extracts different order connectivities to represent semantic relations between entities from CKG. Then, we define a joint learning model to capture high-quality representations of users, items, and their attributes by employing TransE and recurrent neural network, which captures not only structural information, but also sequence information by encoding semantic paths, and take their representations as the users'/items' long-term static features. Next, we respectively employ LSTM and attention machine to capture the users' and items' short-term dynamic preferences. At last, the long-short term features are seamlessly fused into recommender system. We conduct extensive experiments on real-world datasets and the evaluation result shows that HRTR achieves significant superiority over several state-of-the-art baselines.

Spatio-Temporal Self-Attention Network for Next POI Recommendation

Jiacheng Ni[1], Pengpeng Zhao[1], Jiajie Xu[1], Junhua Fang[1], Zhixu Li[1], Xuefeng Xian[2], Zhiming Cui[3], Victor S. Sheng[4]

[1]Soochow University, [2]Suzhou Vocational University, [3]Suzhou University of Science and Technology, [4]Texas Tech University

Abstract. Next Point-of-Interest (POI) recommendation, which aims to recommend next POIs that the user will likely visit in the near future, has become essential in Location-based Social Networks (LBSNs). Various Recurrent Neural Network (RNN) based sequential models have been proposed for next POI recommendation and achieved state-of-theart performance, however RNN is difficult to parallelize which limits its efficiency. Recently, Self-Attention Network (SAN), which is purely based on the self-attention mechanism instead of recurrent modules, improves both performance and efficiency in various sequential tasks. However, none of the existing self-attention networks consider the spatio-temporal intervals between neighbor check-ins, which are essential for modeling user check-in behaviors in next POI recommendation. To this end, in this paper, we propose a new Spatio-Temporal Self-Attention Network (STSAN), which combines self-attention mechanisms with spatio-temporal patterns of users' check-in history. Specifically, time-specific weight matrices and distance-specific weight matrices through a decay function are used to model the spatio-temporal influence of POI pairs. Moreover, we introduce a simple but effective way to dynamically measure the importances of spatial and temporal weights to capture users' spatio-temporal preferences. Finally, we evaluate the proposed model using two real-world LBSN datasets, and the experimental results show that our model significantly outperforms the state-of-the-art approaches for next POI recommendation.

Research Session 8: Machine Learning 1

Time: 13:30-15:40, September 19, 2020, Saturday
Chair: Jiajie Xu, Soochow University

FHAN: Feature-level Hierarchical Attention Network for Group Event Recommendation

Guoqiong Liao[1], Xiaobin Deng[1,2], Xiaomei Huang[1], Changxuan Wan[1]
[1]Jiangxi University of Finance and Economics, [2]Jiangxi Water Resources Institute
Abstract. Recommending events to groups is different from to single-user in event-based social networks (EBSN), which involves various complex factors. Generally, group recommendation methods are either based on recommendation fusion or model fusion. However, most existing methods neglect the fact that user preferences change over time. Moreover, they believe that the weights of different factors that affect group decision-making are fixed in different periods. Recently, there are a few works using the attention mechanism for group recommendation. Although they take into account the dynamic variability of user preferences and the dynamic adjustment of user features weights, they haven't discussed more features of groups and events affecting group decision-making. To this end, we propose a novel Feature-level Hierarchical Attention Network (FHAN) for group event recommendation for EBSN. Specifically, group decision-making factors are divided into group-feature factors and event-feature factors, which are integrated into a two-layer attention network. The first attention layer is constructed to learn the influence weights of words of group topics and event topics, which generates better thematic features. The second attention layer is built to learn the weights of group-feature factors and event-feature factors affecting group decision-making, which results in better comprehensive representation of groups and events. All influence weights of different features in the model can be dynamically adjusted over time. Finally, we evaluate the suggested model on three real-world datasets. Extensive experimental results show that FHAN outperforms the state-of-the-art approaches.

Long Short-Term Memory with Sequence Completion for Cross-Domain Sequential Recommendation

Guang Yang[1], Xiaoguang Hong[1], Zhaohui Peng[1], Yang Xu[2]
[1]Shandong university, [2]Shandong Normal University
Abstract. As the emerging topic to solve the loss of time dimension information, sequential recommender systems (SRSs) has attracted increasing attention in recent years. Although SRSs can model the sequential user behaviors, the interactions between users and items, and the evolution of users' preferences and item popularity over time, the challenging issues of data sparsity and cold start are beyond our control. The conventional solutions based on cross-domain recommendation aims to matrix completion by means of transfer-ring explicit or implicit feedback from the auxiliary domain to the target do-main. But most existing transfer methods can't deal with temporal information. In this paper, we propose a Long Short-Term Memory with Sequence Completion (SCLSTM) model for cross-domain sequential recommendation. We first construct the sequence and supplement it in which two methods are proposed. The first method is to use the intrinsic features of users and items and the temporal features of user behaviors to establish similarity measure for sequence completion. Another method is to improve LSTM by building the connection between the output layer and the input layer of the next time step. Then we use LSTM to complete sequential recommendation. Experimental results on two real datasets extracted from Amazon transaction data demonstrate the superiority of our proposed models against other state-of-the-art

methods.

Turn-Level Recurrence Self-Attention for Joint Dialogue Action Prediction and Response Generation

Yanxin Tan, Zhonghong Ou, Kemeng Liu, Yanan Shi, Meina Song
Beijing University of Posts and Telecommunications

Abstract. In task-oriented dialogue systems, semantically controlled natural language generation is the procedure to generate responses based on current context information. Seq2seq models are widely used to generate dialogue responses and achieve favorable performance. Nevertheless,how to effectively obtain the dialogue's key information from history re-mains to be a critical problem. To overcome this problem, we propose a Turn-level Recurrence Self-Attention (TRSA) encoder, which effectively obtains progressive structural relationship in turn-level from conversation history. Moreover, we propose a novel model to predict dialogue actions and generate dialogue responses jointly, which is different from the separate training model used in previous studies. Experiments demonstrate that our model alleviates the problem of inaccurate attention in dialogue history and improves the degree of dialogue completion significantly. In the large-scale MultiWOZ dataset, we improve the performance by 3.9% of inform rate and 3.4% of success rate, which is significantly higher than the state-of-the-art.

IterG: An Iteratively Learning Graph Convolutional Network With Ontology Semantics

Xingya Liang, Fuxiang Zhang, Xin Liu, Yajun Yang
Tianjin University

Abstract. Knowledge reasoning aims to infer new triples based on existing triples, which is essential for the development of large knowledge graphs, especially for knowledge graph completion. With the development of neural networks, Graph Convolutional Networks (GCNs) in knowledge reasoning have been paid widespread attention in recent years. However, the GCN model only considers the structural information of knowledge graphs and ignores the ontology semantic information. In this paper, we propose a novel model named IterG, which is able to incorporate ontology semantics seamlessly into the GCN model. More specifically, IterG learns the embeddings of knowledge graphs in an unsupervised manner via GCNs and extracts the semantic ontology information via rule learning. The model is capable of propagating relation layer-wisely as well as combining both rich structural information in knowledge graphs and ontological semantics. The experimental results on five real-world datasets demonstrate that our method outperforms the state-of-the-art approaches, and IterG can effectively and efficiently fuse ontology semantics into GCNs.

SSMDL: Semi-supervised multi-task deep learning for transportation mode classification and path prediction with GPS trajectories

Asif Nawaz, Zhiqiu Huang, Senzhang Wang
Nanjing University of Aeronautics and Astronautics

Abstract. The advancement of positioning technology enables people to use GPS devices to record their location histories. The patterns and con-textual information hidden in GPS data opens variety of research issues including trajectory prediction, transportation mode detection, travel route recommendation and many more. Most existing studies have been performed to address these individual issues, but they have the following limitations. 1) Single task learning does not consider the correlations among the correlated tasks. 2) A large number of training samples are required to achieve better performance. In this paper, we propose a semi-supervised multi-task deep learning model to perform the tasks of transportation mode classification and path prediction simultaneously with GPS trajectories. Our model uses both labelled and unlabeled dataset for model training dataset, and concurrently perform both tasks in parallel. Experimental results over a large trajectory dataset collected in Beijing show that our proposal achieves significant performance improvement in terms of all evaluation metrics by comparison with baseline models.

Unsupervised Deep Hashing with Structured Similarity Learning

Xuanrong Pang[1], Xiaojun Chen[1], Shu Yang[1], Feiping Nie[2]
[1]Shenzhen University, [2]Northwestern Polytechnical University

Abstract. Hashing technology, one of the most efficient approximate nearest neighbor searching methods due to its fast query speed and low storage cost, has been widely used in image retrieval. Recently, unsupervised deep hashing methods have attracted more and more attention due to the lack of labels in real applications. Most unsupervised hashing methods usually construct a similarity matrix with the features extracted from the images, and then guide the hash code learning with this similarity matrix. However, in unsupervised scenario, such similarity matrix may be unreliable due to the affect of noise and irrelevant objects in images. In this paper, we propose a novel unsupervised deep hashing method called Deep Structured Hashing (DSH). In the new method, we first learn both continuous and binary structured similarity matrices with explicit cluster structure to better preserve the semantic structure, where the binary one preserves the coarse-grained semantic structure while the continuous one preserves the fine-grained semantic structure. And then jointly optimize three kinds of losses to learn high quality hash codes. Extensive experiments on three benchmark datasets show the superior retrieval performance of our proposed method.

A Framework for Learning Cross-lingual Word Embedding with Topics

Xiaoya Peng, Dong Zhou
Hunan university of science and technology

Abstract. Cross-lingual word embeddings have been served as fundamental components for many Web-based applications. However, current models learn cross-lingual word embeddings based on projection of two pre-trained monolingual embeddings based on well-known models such as word2vec. This procedure makes it indiscriminative for some crucial factors of words such as homonymy and polysemy. In this paper, we propose a novel framework for learning better cross-lingual word embeddings with latent topics. In this framework, we firstly incorporate latent topical representations into the Skip-Gram model to learn high quality monolingual word

embeddings. Then we use the supervised and unsupervised methods to train cross-lingual word embeddings with topical information. We evaluate our framework in the cross-lingual Web search tasks using the CLEF test collections. The results show that our framework outperforms previous state-of-the-art methods for generating cross-lingual word embeddings.

Research Session 9: Knowledge Graph

Time: 13:30-15:40, September 19, 2020, Saturday
Chair: Bohan Li, Nanjing University of Aeronautics and Astronautics

Knowledge Graph Attention Network Enhanced Sequential Recommendation

Xingwei Zhu[1], Pengpeng Zhao[1], Jiajie Xu[1], Junhua Fang[1], Lei Zhao[1], Xuefeng Xian[2], Zhiming Cui[3], Victor Sheng[4]
[1]Soochow University, [2]Suzhou Vocational University, [3]Suzhou University of Science and Technology
[4]Texas Tech University

Abstract. Knowledge graph (KG) has recently been proved effective and attracted a lot of attentions in sequential recommender systems. However, the relations between the attributes of different entities in KG, which could be utilized to improve the performance, remain largely unexploited. In this paper, we propose an end-to-end Knowledge Graph attention network enhanced Sequential Recommendation (KGSR) framework to capture the context-dependency of sequence items and the semantic information of items in KG by explicitly exploiting high-order relations between entities. Specifically, our method first combines the user-item bipartite graph and the KG into a unified graph and encodes all nodes of the unified graph into vector representations with TransR. Then, a graph attention network recursively propagates the information of neighbor nodes to refine the embedding of nodes and distinguishes the importance of neighbors with an attention mechanism. Finally, we apply recurrent neural network to capture the user's dynamic preferences by encoding user-interactive sequence items that contain rich auxiliary semantic information. Experimental results on two datasets demonstrate that KGSR outperforms the state-of-the-art sequential recommendation methods.

An Ontology-Aware Unified Storage Scheme for Knowledge Graphs

Sizhuo Li, Guozheng Rao, Baozhu Liu, Pengkai Liu, Sicong Dong, Zhiyong Feng
Tianjin University

Abstract. With the development of knowledge-based artificial intelligence, the scale of knowledge graphs has been increasing rapidly. The RDF graph and the property graph are two mainstream data models of knowledge graphs. On the one hand, with the development of the Semantic Web, there are a large number of RDF knowledge graphs. On the other hand, property graphs are widely used in the graph database community. However, different families of data management methods of RDF graphs and property graphs have been seperately developed in each community over a decade, which hinder the interoperability in managing large knowledge graph data. To address this problem, we propose a

unified storage scheme for knowledge graphs which can seamlessly accommodate both RDF and property graphs. Meanwhile, the concept of ontology is introduced to meet the need for RDF graph data storage and query load. Experimental results on the benchmark datasets show that the proposed ontology-aware unified storage scheme can effectively manage large-scale knowledge graphs and significantly avoid data redundancy.

Fine-grained Evaluation of Knowledge Graph Embedding Models in Downstream Tasks

Yuxin Zhang[1], Bohan Li[1], Han Gao[1], Ye Ji[1], Han Yang[2], Meng Wang[3]
[1]Nanjing University of Aeronautics and Astronautics, [2]Peking University, [3]Southeast University

Abstract. Knowledge graph (KG) embedding models are proposed to encode entities and relations into a low-dimensional vector space, in turn, can support various machine learning models on KG completion with good performance and robustness. However, the current entity ranking protocol about KG completion cannot adequately evaluate the impacts of KG embedding models in real-world applications. However, KG embeddings is not widely used as word embeddings. An asserted powerful KG embedding model may not be effective in downstream tasks. So in this paper, we commit to finding the answers by using downstream tasks instead of entity ranking protocol to evaluate the effectiveness of KG embeddings. Specifically, we conduct comprehensive experiments on different KG embedding models in KG based recommendation and question answering tasks. Our findings indicate that: 1) Modifying embeddings by considering more complex KG structural information may not achieve improvements in practical applications, such as updating TransE to TransR. 2) Modeling KG embeddings in non-euclidean space can effectively improve the performance of downstream tasks.

Learning to Answer Complex Questions with Evidence Graph

Gao Gu[1], Bohan Li[1], Han Gao[1], Meng Wang[2]
[1]Nanjing University of Aeronautics and Astronautics, [2]Southeast University

Abstract. Text-based end-to-end question answering (QA) systems have attracted more attention for their good robustness and excellent performance in dealing with complex questions. However, this kind of method lacks certain interpretability, which is essential for the QA system. For instance, the interpretability of answers is particularly significant in the medical field, in that interpretable answers are more credible and apt to acception. The methods based on knowledge graph (KG) can improve the interpretability, but suffer from the problems of incompleteness and sparseness of KG. In this paper, we propose a novel method (EGQA) to solve complex question answering via combining text and KG. We use Wikipedia as a text source to extract documents related to the question and extract triples from the documents to construct a raw graph (i.e., a small-scale KG). Then, we extract the evidence graphs from the raw graph and adopt Attention-based Graph Neural Network (AGNN) to embed them to find the answer. Our experiments conduct on a real medical dataset Head-QA, which shows that our approach can effectively improve the interpretability and performance of complex question answering.

Tail Entity Recognition and Linking for Knowledge Graphs

Dalei Zhang[1], Yang Qiang[2], Zhixu Li[1], Junhua Fang[1], Ying He[3], Xin Zheng[3], Zhigang Chen[3]
[1]Soochow University, [2]King Abdullah University of Science and Technology, [3]iFLYTEK CO., LTD

Abstract. This paper works on a new task - Tail Entity Recognition and Linking (TERL) for Knowledge Graphs (KG), i.e., recognizing ambiguous entity mentions from the tails of some relational triples, and linking these mentions to their corresponding KG entities. Although plenty of work has been done on both entity recognition and entity linking, the TERL problem in this specific scenario is untouched. In this paper, we work towards the TERL problem by fully leveraging KG information with two neural models for solving the two sub-problems, i.e., tail entity recognition and tail entity linking respectively. We finally solve the TERL problem end-to-end by proposing a joint learning mechanism with the two proposed neural models, which could further improve both tail entity recognition and linking results. To the best of our knowledge, this is the first effort working towards TERL for KG. Our empirical study conducted on real-world datasets shows that our models can effectively expand KG and improve the quality of KG.

Diversified Top-k Querying in Knowledge Graphs

Xintong Guo, Hong Gao, Yinan An, Zhaonian Zou
Harbin Institute of Technology

Abstract. The existing literatures of the query processing on knowledge graphs focus on an exhaustive enumeration of all matches, which is time-consuming. Users are often interested in diversified top-k matches, rather than the entire match set. Motivated by these, this paper formalizes the diversified top-k querying(DTQ) problem in the context of RDF/SPARQL and proposes a diversification function to balance importance and diversity. We first prove that the decision problem of DTQ is NP-complete, and give a baseline algorithm with an approximation ratio of 2. Secondly, an index-based algorithm with the early termination property is proposed. The index is adept in parallel diversified top-k selection in multicore architectures. Using real-world and synthetic data, we experimentally verify that our algorithms are efficient and effective in computing meaningful diversified top-k matches.

Research Session 10: Text Analysis

Time: 13:30-15:40, September 19, 2020, Saturday
Chair: Rize Jin, Tiangong University

Densely-connected Transformer with Co-attentive Information for Matching Text Sequences

Minxu Zhang[1], Yingxia Shao[3], Kai Lei[1], Yuesheng Zhu[1], Bin Cui[2],
[1]Peking University & Peking University Shenzhen Graduate School, [2]Peking University,
[3]Beijing University of Posts and Telecommunications

Abstract. Sentence matching, which aims to capture the semantic relationship between two sequences, is a crucial problem in NLP research. It plays a vital role in various natural language tasks

such as question answering, natural language inference and paraphrase identification. The state-of-the-art works utilize the interactive information of sentence pairs through adopting the general Compare-Aggregate framework and achieve promising results. In this study, we propose Densely connected Transformer to perform multiple matching processes with co-attentive information to enhance the interaction of sentence pairs in each matching process. Specifically, our model consists of multiple stacked matching blocks. Inside each block, we first employ a transformer encoder to obtain refined representations for two sequences, then we leverage multi-way co-attention mechanism or multi-head co-attention mechanism to perform word-level comparison between the two sequences, the original representations and aligned representations are fused to form the alignment information of this matching layer. We evaluate our proposed model on five well-studied sentence matching datasets and achieve highly competitive performance.

Contribution of Improved Character Embedding and Latent Posting Styles to Authorship Attribution of Short Texts

Wenjing Huang, Rui Su, Mizuho Iwaihara
Waseda University

Abstract. Text contents generated by social networking platforms tend to be short. The problem of authorship attribution on short texts is to determine the author of a given collection of short posts, which is more challenging than that on long texts. Considering the textual characteristics of sparsity and using informal terms, we propose a method of learning text representations using a mixture of words and character n-grams, as input to the architecture of deep neural networks. In this way we make full use of user mentions and topic mentions in posts. We also focus on the textual implicit characteristics and incorporate ten latent posting styles into the models. Our experimental evaluations on tweets show a significant improvement over baselines. We achieve a best accuracy of 83.6%, which is 7.5% improvement over the state-of-the-art. Further experiments with increasing number of authors also demonstrate the superiority of our models.

Enabling Efficient Multi-Keyword Search over Fine grained Authorized Healthcare Blockchain System

Yicheng Ding, Wei Song, Yuan Shen
Wuhan Unversity

Abstract. As a new emerging technology, blockchain is attracting the attention from academic and industry and has been widely exploited to build the large-scale data sharing and management systems, such as healthcare database or bank distributed database system. The health records contain a lot of sensitive information, so putting these health records into blockchain can solve the security and privacy issues while uploading them to an untrustworthy network. In a typical health record management system, there are escalating demands for users including the patients and the doctors to execute multi-keyword search over the huge scale of healthcare records. In the meantime, they can authorize some part of their personal treatments to others according to personalized needs of the patients. In literatures, there is not an existing blockchain solution can satisfy these two requirements

at the same time. These issues become prominent since it's more inconvenient to adjust a blockchain-based system to support efficient multi-keyword search and fine-grained authorization comparing to traditional RDBMS.To overcome the two challenges, we propose a novel multi-keyword searching scheme by establishing a set of Bloom Filters within the health record blockchain system to accelerate the searching process on service provider (SP). Moreover, we reduce the overhead of key derivation by proposing a Healthcare Data Key Derivation Tree (HDKDT) stored locally on the user's side. Putting our proposed scheme on the medical blockchain can speed up the multi-keyword search [3,12] processes and reduce the key storage space to certain extent. At the end of this article, we formally prove the security of the proposed scheme and implement a prototype system to evaluate its performance. The experimental results validate our proposed scheme in this paper is a secure and efficient approach for the health record management scenario.

Paperant: Key Elements Generation with New Ideas

Xin He[1], Jiuyang Tang[1], Zhen Tan[1], Zheng Yu[2], Xiang Zhao[1]
[1]National University of Defense Technology, [2]Mininglamp Technology

Abstract. Long text generation, leveraging one sentence to generate a meaningful paper, is an effective method to reduce repetitive works. Conventional text generation models utilize rule-based and plan-based methods to produce paper, such as SCIgen, which is hard to suit the complex sematic scene. Recently, several neural network-based models, such as Point Network and PaperRobot, were proposed to tackle the problem, and achieve state-of-the-art performance. However, most of them only try to generate part of the paper, and ignore the sematic information of each entity in input sentence. In this paper, we present a novel method named Paperant, which leverage not only multi-sentence features to describe latent features of each entity, but also hybrid structure to generate different parts of the paper. In experiment, Paperant was superior to other methods on each indicator.

A Method for Place Name Recognition in Tang Poetry Based on Feature Templates and Conditional Random Field

Yan Zhang, Yukun Li, Jing Zhang, Yunbo Ye
Tianjin University of Technology

Abstract. Tang poetry is an important aspect of ancient Chinese culture. Given that Tang poetry has unique features in text structure, how to use entity recognition, knowledge graph and other information processing technologies to research poetry is of great importance. However, the existing artificial neural network methods for named entity recognition require a large number of labeled training sets, while Chinese Tang poetry has not been labeled with a good training set. Besides, the grammatical structure of Tang poetry is far from modern Chinese. Therefore, for place name recognition in poetry, the existing neural network methods do not perform well. This article studies and analyzes the metrical form of Tang poetry, finds the metrical rules of place names, and summarizes the feature templates based on the metrical rules. According to the feature templates of Tang poetry, a method of combining feature templates with conditional random field is proposed. Experimental results prove the effectiveness of the proposed method.

WEKE: Learning Word Embeddings for Keyphrase Extraction

Yuxiang Zhang[1], Huan Liu[1], Bei Shi[2], Xiaoli Li[3], Suge Wang[4]
[1]Civil Aviation University of China, [2]Tencent AI Lab, [3]Institute for Infocomm Research, A*STAR, Singapore, [4]Shanxi University

Abstract. Traditional supervised keyphrase extraction models depend on the features of labeled keyphrases while prevailing unsupervised models mainly rely on global structure of the word graph, with nodes representing candidate words and edges/links capturing the co-occurrence between words. However, the local context information of the word graph can not be exploited in existing unsupervised graph-based key phrase ex-traction methods and integrating different types of information into a unified model is relatively unexplored. In this paper, we propose a new word embedding model specially for keyphrase extraction task, which can capture local context information and incorporate them with other types of crucial information into the low-dimensional word vector to help better extract keyphrases. Experimental results show that our method consistently outperforms 7 state-of-the-art unsupervised methods on three real datasets in Computer Science area for keyphrase extraction.

Research Session 11: Information Extraction and Retrieval

Time: 13:30-15:40, September 19, 2020, Saturday
Chair: Lin Li, Wuhan University of Technology

Multi-hop Reading Comprehension Incorporating Sentence-Based Reasoning

Lijun Huo[1], Bin Ge[1], Xiang Zhao[1,2]
[1]National University of Defense Technology,
[2]Collaborative Innovation Center of Geospatial Technology

Abstract. Multi-hop machine reading comprehension (MRC) requires models to mine and utilize relevant information from multiple documents to predict the answer to a semantically related question. Existing work resorts to either document-level or entity-level inference among relevant information, which can be too coarse or too subtle, resulting less accurate understanding of the texts. To mitigate the issue, this research proposes a sentence-based multi-hop reasoning approach named SMR. SMR starts with sentences of documents, and unites the question to establish several reasoning chains based on sentence-level representations. In addition, to resolve the complication of pronouns on sentence semantics, we concatenate two sentences, if necessary, to assist in constructing reasoning chains. The model then synthesizes the information existed in all the reasoning chains, and predicts a probability distribution for selecting the correct answer. In experiments, we evaluate SMR on two popular multi-hop MRC benchmark datasets - WikiHop and MedHop. The model achieves 68.3 and 62.9 in terms of accuracy, respectively, exhibiting a remarkable improvement over state-of-the-art option. Additionally, qualitative analysis also demonstrates the validity and interpretability of SMR.

Discriminative Multi-label Model Reuse for Multi-label Learning

Yi Zhang, Zhecheng Zhang, Yinlong Zhu, Lei Zhang, Chongjun Wang
Nanjing University

Abstract. Traditional Chinese Medicine (TCM) with diagnosis scales is a holistic way for diagnosing Parkinson's Disease, where symptoms can be represented as multiple labels. To solve this problem, multi-label learning provides a framework for handling such task and has exhibited excellent performance. Besides, it is a challenging issue of how to effectively utilize label correlations in multi-label learning. In this paper,we propose a novel algorithm named Discriminative Multi-label Model Reuse (DMLMR) for multi-label learning, which exploits label correlations with model reuse, instance distribution adaptation and label distribution adaptation. Experiments on real-world dataset of Parkinson's disease demonstrate the superiority of DMLMR for diagnosing PD. To prove the effectiveness of the proposed DMLMR, extensive experiments on four benchmark multi-label datasets show that DMLMR significantly outperforms other state-of-the-art multi-label learning algorithm.

Debiasing Learning to Rank Models with Generative Adversarial Networks

Hui Cai, Chengyu Wang, Xiaofeng He
East China Normal University

Abstract. Unbiased learning to rank aims to generate optimal orders for candidates utilizing noisy click-through data. To deal with such problem, most models treat the biased click labels as combined supervision of relevance and propensity, which pay little attention to the uncertainty of implicit user feedback. We propose a semi-supervised framework to ad-dress this issue, namely ULTRGAN (Unbiased Learning To Rank with Generative Adversarial Networks). The unified framework regards the task as semi-supervised learning with missing labels, and employs adversarial training to debias click-through datasets. In ULTRGAN, the generator samples potential negative examples combined with true positive examples for the discriminator. Meanwhile, the discriminator challenges the generator for better performances. We further incorporate pairwise debiasing to generate unbiased labels diffusing from the discriminator to the generator. Experimental results over both synthetic and real-world datasets show the effectiveness and robustness of ULTRGAN

Joint Reasoning of Events, Participants and Locations for Plot Relation Recognition

Shengguang Qiu[1], Botao Yu [1], Lei Qian[2], Qiang Guo[2], Wei Hu[1,2]
[1]Nanjing University, [2]State Key Laboratory of Mathematical Engineering and Advanced Computing

Abstract. Event information is of great value, but the exploitation of it generally relies on not only extracting events from the text, but also figuring out the relations among events and organizing them accordingly.In this paper, based on a more flexible and practical type of event relation called the plot relation, we study the method of automatic event relation recognition. Specifically, we propose a local prediction method by using diversified linguistic and temporal features. Furthermore, we design a joint reasoning framework, in which we leverage the information of participants and locations, and add global constraints to further improve the performance. Finally, we transform the proposed model

into integer linear programming (ILP) to obtain the global optimum. Our experiments demonstrate that our method significantly outperforms all the existing methods.

Multi-view Clustering via Multiple Auto-Encoder

Guowang Du[1], Lihua Zhou[1], Yudi Yang[1], Kevin Lü[2], Lizhen Wang[1]
[1]Yunnan University, [2]Brunel University

Abstract. Multi-view clustering (MVC), which aims to explore the underlying structure of data by leveraging heterogeneous information of different views, has brought along a growth of attention. Multi-view clustering algorithms based on different theories have been proposed and extended in various applications. However, existing of most MVC algorithms are shallow models. They learn structure information of multi-view data by mapping multi-view data to low-dimensional representation space directly, which ignore the Non-linear structure information hidden in each view. This weakens the performance of multi-view clustering to a certain extent. In this paper, we propose a multi-view clustering algorithm based on multiple Auto-Encoder, named MVC-MAE, to cluster multi-view data. MVC-MAE algorithm adopts Auto-Encoder to capture the non-linear structure information of each view in a layer-wise manner. To exploit the consistent and complementary information contained in different views, we also incorporate the local invariance within each view and consistent and complementary information between any two views. Besides, we integrate the representation learning and clustering into a unified step, which jointly optimizes these two steps. Extensive experiments on three real-world datasets demonstrate a superior performance of our algorithm compared with 13 baseline algorithms in terms of two evaluation metrics.

Dynamic Multi-hop Reasoning

Liang Xu[1], Junjie Yao[1], Yingjie Zhang[2]
[1]East China Normal University.
[2]Shanghai Electric Vehicle Public Data Collecting, Monitoring and Research Center

Abstract. Multi-hop reasoning is an essential part of the current reading comprehension and question answering areas. The reasoning methods have been extensively studied, and most of them are generally focused on the pre-retrieval based inference, with the help of a few paragraphs. These methods are fixed and unable to cope with dynamic and complex questions. Here, we propose to utilize the dynamic graph reasoning network for multi-hop reading comprehension question answering.

Specifically, the new approach continuously infers the clue entities and candidate answers based on the question and clue paragraphs. The clue entities and candidate answers extracted at each hop are used as new nodes to expand the dynamic graph. Then we iteratively update the semantic representation of the questions via dynamic question memory, and apply the graph attention network to encode the information of inference paths. Extensive experiments on two datasets verify the advantage and improvements of the proposed approach.

Research Session 12: Machine Learning 2

Time: 15:50-18:00, September 19, 2020, Saturday
Chair: Caie Xu, University of Yamanashi

D-GHNAS for Joint Intent Classification and Slot Filling

Yanxi Tang, Jianzong Wang, Xiaoyang Qu, Nan Zhang, Jing Xiao
Ping An Technology (Shenzhen) Co., Ltd.

Abstract. Intent classification and slot filling are two classical problems for spoken language understanding and dialog systems. The existing works, either accomplishing intent classification or slot filling separately or using a joint model, are all human-designed models with trial and error. In order to explore the variety of network architecture and to find whether there exist possible network architectures with better results, we proposed the D-GHNAS (Deep deterministic policy gradient based Graph Hypernetwork Neural Architecture Search) to accomplish intent classification and slot filling via a NAS (Neural Architecture Search)method. NAS based techniques can automatically search for network architectures without experts' trial and error. Different from early NAS methods with hundreds of GPU days to find an ideal neural architecture that takes too much computation resource, in this work, hypernetwork is used to decrease the computation cost. Experimental results demonstrate that our model improves intent classification and slot filling results on public benchmark datasets ATIS and SNIPS compared with other joint models for these tasks.

Utilizing BERT Pretrained Models with Various Fine tune Methods for Subjectivity Detection

Hairong Huo, Mizuho Iwaihara
Waseda University

Abstract. As an essentially antecedent task of sentiment analysis, subjectivity detection refers to classifying sentences to be subjective ones containing opinions, or objective and neutral ones with out bias. In the situations where impartial language is required, such as Wikipedia, subjectivity detection could play an important part. Recently, pretrained language models have proven to be effective in learning representations, profoundly boosting the performance among several NLP tasks. As a state-of-art pretrained model, BERT is trained on large unlabeled data with masked word prediction and next sentence prediction tasks. In this paper, we mainly explore utilizing BERT pretrained models with several combinations of fine-tuning methods, holding the intention to enhance performance in subjectivity detection task. Our experimental results reveal that optimum combinations of fine-tune and multi-task learning surplus the state-of-the-art on subjectivity detection and related tasks.

Hylo: Hybrid Layer-Based Optimization to Reduce Communication in Distributed Deep Learning

Wenbin Jiang, Jing Peng, Pai Liu, Hai Jin
Huazhong University of Science and Technology

Abstract. In distributed deep learning training, the synchronization of gradients usually brings huge

network communication overhead. Although many methods have been proposed to solve the problem, limited effectiveness has been obtained, since these methods do not fully consider the differences of diverse layers. We propose a novel hybrid layer-based optimization approach named Hylo to reduce the communication over-head. Two different strategies are designed for gradient compression of two types of layers (convolution layer and fully-connected layer). For convolution layers, only some important convolution kernels are chosen for gradient transmission. For fully-connected layers, all gradients are quantized to 2 bits with an adaptive gradient threshold. The experimental results show that Hylo brings obvious accelerations for distributed deep learning systems, while with little accuracy loss. It achieves training speedups up to 1.31x compared to state-of-the-art works.

Global and Local Attention Embedding Network for Few-Shot Fine-Grained Image Classification

Jiayuan Hu, Chung-Ming Own, Wenyuan Tao
Tianjin University

Abstract. Few-shot fine-grained image recognition aims to classify fine-grained images with limited training samples. Nowadays exist in a majority of few-shot fine-grained image classification methods the following problems: local information loss and ignoring pivotal parts. To solve the above problems, this paper proposes a new embedding module, called GLAE. The author designs a hierarchical structure and combines the first-order and second-order information to reduce the local information loss. Besides, this paper proposes an attention mechanism to obtain the vital parts by the attention mask. On the StanfordCars dataset, GLAE achieves an accuracy of 91.18% which is the best result in the field of few-shot fine-grained image recognition.

Predicting Human Mobility with Self-Attention and Feature Interaction

Jingang Jiang, Shuo Tao, Defu Lian, Zhenya Huang, Enhong Chen
University of Science and Technology of China

Abstract. Mobility prediction plays an important role in a wide range of location-based applications and services. However, two important challenges are not well addressed in existing literature: 1) explicit high-order interactions of spatio-temporal features are not systemically modeled; 2) most existing algorithms place attention mechanisms on top of recurrent network, so they can not allow for full parallelism and are inferior to self-attention for capturing long-range dependence. To this end, we propose MoveNet, a self-attention based sequential model, to predict each user's next destination based on her most recent visits and historical trajectory. MoveNet first introduces a cross based learning framework for modeling feature interactions. With self-attention on both the most recent visits and historical trajectory, MoveNet can use an attention mechanism to capture user's long-term regularity in a more efficient and effective way. We evaluate MoveNet with three real-world mobility datasets, and show that MoveNet outperforms the state-of-the-art mobility predictor by around 10% in terms of accuracy, and simultaneously achieves faster convergence and over 4x training speedup.

Active Classification of Cold-start Users in Large Sparse Datasets

Xiang Li[1], Xiao Li[2], Wang Tao[2]
[1]Academy of Military Science, [2]National University of Defense Technology

Abstract. Many applications need to perform classification on large sparse datasets.Classifying the cold-start users who have very few feedbacks is still a challenging task. Previous work has applied active learning to classification with partially observed data. However, for large and sparse data, the number of feedbacks to be queried is huge and many of them are invalid. In this paper, we develop an active classification framework that can address these challenges by leveraging online Matrix Factorization models. We first identify a step-wise data acquisition heuristic which is useful for active classification. We then use the estimations of online Probabilistic Matrix Factorization to compute this heuristic function. In order to reduce the number of invalid queries, we further estimate the probability that a query can be answered by the cold-start user with online Poisson Factorization. During active learning, a query is selected based on the current knowledge learned in these two online factorization models. We demonstrate with real-world movie rating datasets that our framework is highly effective. It not only gains better improvement in classification, but also reduces the number of invalid queries.

Multi-task Learning for Low-resource Second Language Acquisition Modeling

Yong Hu[1,2], Heyan Huang[1], Tian Lan[1], Xiaochi Wei[3], Yuxiang Nie[1], Jiarui Qi[1], Liner Yang[4], Xian-Ling Mao[1]
[1]Beijing Institute of Technology, [2]CETC Big Data Research Institute Co,. Ltd.,
[3]Baidu Inc., [4]Beijing Language and Culture University

Abstract. Second language acquisition (SLA) modeling is to predict whether second language learners could correctly answer the questions according to what they have learned, which is a fundamental building block of the personalized learning system. However, as far as we know, almost all existing methods cannot work well in low-resource scenarios due to lacking of training data. Fortunately, there are some latent common patterns among different language-learning tasks, which gives us an opportunity to solve the low-resource SLA modeling problem. Inspired by this idea, we propose a novel SLA modeling method, which learns the latent common patterns among different language-learning datasets by multi-task learning and are further applied to improving the prediction performance in low-resource scenarios. Extensive experiments show that the proposed method performs much better than the state-of-the-art baselines in the low-resource scenario. Meanwhile, it also obtains improvement slightly in the non-low-resource scenario.

Research Session 13: Recommender System

Time: 15:50-18:00, September 19, 2020, Saturday
Chair: Weiguo Zheng, Fudan University

Few-Shot Representation Learning for Cold-Start Users and Items

Bowen Hao, Jing Zhang, Cuiping Li, Hong Chen
Renmin University of China

Abstract. Existing recommendation algorithms suffer from cold-start issues as it is challenging to learn accurate representations of cold-start users and items. In this paper, we formulate learning the representations of cold-start users and items as a few-shot learning task, and address it by training a representation function to predict the target user (item) embeddings based on limited training instances. Specifically, we propose a novel attention-based encoder serving as the neural function, with which the K training instances of a user (item) are viewed as the interactive context information to be further encoded and aggregated. Experiments show that our proposed method significantly outperforms existing baselines in predicting the representations of the cold-start users and items, and improves several downstream tasks where the embeddings of users and items are used.

Dual Role Neural Graph auto-encoder for CQA Recommendation

Xing Luo, Yuanyuan Jin, Tao Ji, Xiaoling Wang
East China Normal University

Abstract. Matching between questions and suitable users is an appealing and challenging problem in the research area of community question answering (CQA). Usually, different from the traditional recommendation systems where a user has only a single role, each user in CQA can play two different roles (dual roles) simultaneously: as a requester and as an answerer. For different roles, users usually have varying interests and expertise in different topics and knowledge domains, which is rarely addressed in the previous methods. Besides, based on an explicit single link between two users, existing methods cannot capture implicit associations between their possibly similar roles. Therefore, in this paper, we propose the structure of a dual role graph and employ the link prediction approach to make CQA recommendation on the graph. Moreover, we develop a Dual Role Neural Graph auto-encoder (DRNGae) framework, which can: 1) encode the dual role graph structure to capture the implicit dual role correlation by propagating high-order information embeddings of graph neural network; 2) learn variable weights with the dual role feature preferences from dual role content information by self-attention mechanism; 3) reconstruct the graph structure to predict the possible interaction links. Experimental studies on real-world datasets verify our design and prove that our model achieves significantly better performance than baselines in link prediction (95.3% AUC, 96.2% AP on Citeseer dataset) and CQA recommendation (79.5% recall@25, 76.7% ndcg@25 on Yahoo! answer dataset).

Joint Cooperative Content Caching and Recommendation in Mobile Edge-Cloud Networks

Zhihui Ke[1], Meng Cheng[2], Xiaobo Zhou[1], Keqiu Li[1], Tie Qiu[1]
[1]Tianjin University, [2]Japan Advanced Institute of Science and Technology

Abstract. In mobile edge-cloud networks, multiple edge nodes form a mesh network to cooperate with each other. To maximize the benefit of resource-limited edge nodes, the content providers jointly optimize the content caching and recommendation decisions. However, the cooperation

between edge nodes complicates both the content caching and recommendation decisions. To solve this problem, in this paper, we propose an efficient joint cooperative content caching and recommendation scheme in edge-cloud networks. Specifically, we formulate the joint cooperative content caching and recommendation problem as an integer-linear programming problem to minimize the average download delay, with controllable user preference distortion tolerance. We propose an efficient heuristic algorithm to solve the formulated problem due to its NP-hardness. We evaluate the performance of the proposed scheme with the MovieLens dataset. The simulation results demonstrate that the proposed scheme can decrease the average download latency by up to 37% and improve average cache hit rate by up to 24%, as compared with state-of-the-art solutions.

Graph Attentive Network for Region Recommendation with POI- and ROI-Level Attetnion

Hengpeng Xu[1], Jinmao Wei[1], Zhenglu Yang[1], Jun Wang[2]
[1]Nankai University, [2]Ludong University

Abstract. Due to the prevalence of human activity in urban space, recommending ROIs (region-of-interest) to users becomes an important task in social networks. The fundamental problem is how to aggregate users' preferences over POIs (point-of-interest) to infer the users' region-level mobility patterns. We emphasize two facts in this paper: (1) there simultaneously exists ROI-level and POI-level implicitness that blurs the users' underlying preferences; and (2) individual POIs should have non-uniform weights and more importantly, the weights should vary across different users. To address these issues, we contribute a novel solution, namely GANR2 (Graph Attentive Neural Network for Region Recommendation), based on the recent development of attention network and Neural Graph Collaborative Filtering (NGCF). Specifically, to learn the user preferences over ROIs, we provide a principled neural network model equipped with two attention modules: the POI-level attention module, to select informative POIs of one ROI, and the ROI-level attention module, to learn the ROI preferences. Moreover, we learn the interactions between users and ROIs under the NGCF framework. Extensive experiments on two real-world datasets demonstrate the effectiveness of the proposed framework.

Generalized Collaborative Personalized Ranking For Recommendation

Bin Fu[1], Hongzhi Liu[1], Yang Song[2], Tao Zhang[3], Zhonghai Wu[1]
[1]Peking University, [2]BOSS Zhipin NLP Center, [3]BOSS Zhipin

Abstract. Data sparsity is a common problem in collaborative ranking for personalized recommendation with implicit feedback. Several previous work tried to 'borrow' feedback information from users' neighborhood as their prior preferences to alleviate this problem. However, they emphasize the overlapping interests of users and their neighborhood while de-emphasize the importance of users' own specific taste, which leads to under-personalization. In addition, they ignore the collaborative influence among items which is also important for preference learning.
To solve these problems, we propose an effective collaborative ranking method Generalized Collaborative Personalized Ranking (GCPR), which utilizes the collaborative influence among users and items in a unified framework. It strengthens the specific taste of users by using inner-basket

influence of items to enhance the personalization. In addition, it utilizes cross-basket influence of items to dig more collaborative items to further alleviate the sparsity problem. Then, we utilize generalized AUC to learn a confidence-based listwise preference, and propose a post-training based on self-paced learning to solve the top-biased problem of the generalized AUC. Experimental results on four public real-world datasets show that GCPR achieves better performance than traditional collaborative filtering (CF) methods and state-of-the-art collaborative ranking methods.

KGWD: Knowledge Graph Based Wide & Deep Framework for Recommendation

Kemeng Liu, Zhonghong Ou, Yanxin Tan, Kai Zhao, Meina Song
Beijing University of Posts and Telecommunications

Abstract. Knowledge Graph (KG) contains rich real-world auxiliary information, which can be leveraged to improve the performance of recommender systems. Nevertheless, existing recommender systems usually sample and aggregate neighbor entities and relations that link to target items to enrich the representations of items or users, whereas ignoring combinatorial features among different neighbor entities and relations. To resolve the problem mentioned above, we propose an end-to-end Knowledge Graph based Wide & Deep (KGWD) framework to leverage combinatorial features effectively. At the wide level, KGWD introduces a novel Triplet Compressed Interaction Network (TriCIN) to generate high-order combinatorial features among different triplets associated with the target item automatically. At the deep level, KGWD discovers users' potential long-distance preferences by mining multi-hop neighbor information over the KG. We conduct experiments on three real-world datasets, i.e., Yelp2018, Last-FM, and Amazon-book, to evaluate the performance of KGWD. Experimental results demonstrate that KGWD outperforms state-of-the-art schemes significantly. Specifically, in all three datasets, KGWD improves the F1-score by more than 5% over the state-of-the-art.

Frequent Semantic Trajectory Sequence Pattern Mining in Location-Based Social Networks

Zhen Zhang[1], Jing Zhang[1], Fuxue Li[1], Xiangguo Zhao[2], Xin Bi[2]
[1]Yingkou Institute of Technology, [2]Northeastern University

Abstract. At present, more and more researchers have focused on the study of the frequent trajectory sequence pattern mining in location-based social network (LBSN), in which the trajectories of contributing frequent patterns in users' trajectory database must have same or similar the location coordinates and conform to the semantics and time constraints. In this paper, we focus on the study of users' daily frequent mobile pattern. Excessive limitations on location information may limit the results of mining users' frequent mobile pattern. Therefore, based on the frequent trajectory sequence pattern mining in LBSNs, we first define a new frequent semantic trajectory sequence pattern mining (FSTS-PM) problem that focuses on the study of mining users' frequent mobile pattern. FSTS-PM problem does not consider the location coordinates of the trajectory points, but uses the distance and time constraints among the trajectory points in a trajectory sequence to optimize the user's frequent mobile pattern mining results. Then, we propose the modified Pre-fixSpan (MP) algorithm which integrates the distance and time filtering mechanism based on the original PrefixSpan to find

frequent semantic trajectory sequence pattern. Finally, the extensive experiments verify the performance of MP algorithm.

Research Session 14: Social Networks

Time: 15:50-18:00, September 19, 2020, Saturday
Chair: Jianxin Li, Deakin University

Multiple Local Community Detection via High-Quality Seed Identification

Jiaxu Liu, Yingxia Shao, Sen Su
Beijing University of Posts and Telecommunications

Abstract. Local community detection aims to find the communities that a given seed node belongs to. Most existing works on this problem are based on a very strict assumption that the seed node only belongs to a single community, but in real-world networks, nodes are likely to belong to multiple communities. In this paper, we introduce a novel algorithm, HqsMLCD, that can detect multiple communities for a given seed node. HqsMLCD first finds the high-quality seeds which can detect better communities than the given seed node with the help of network representation, then expands the high-quality seeds one-by-one to get multiple communities, probably overlapping. Experimental results on real-world networks demonstrate that our new method HqsMLCD outperforms the state-of-the-art multiple local community detection algorithms.

Efficient Personalized Influential Community Search in Large Networks

Yanping Wu, Jun Zhao, Renjie Sun, Chen Chen, Xiaoyang Wang
Zhejiang Gongshang University

Abstract. Community search, which aims to retrieve important communities (i.e., subgraphs) for a given query vertex, has been widely studied in the literature. In the recent, plenty of research is conducted to detect influential communities, where each vertex in the network is associated with an influence value. Nevertheless, there is a paucity of work that can support personalized requirement. In this paper, we propose a new problem, i.e., maximal personalized influential community (MPIC) search. Given a graph G, an integer k and a query vertex u, we aim to obtain the most influential community for u by leveraging the k-core concept. To handle larger networks efficiently, two algorithms, i.e., top-down algorithm and bottom-up algorithm, are developed. To further speedup the search, an index-based approach is proposed. We conduct extensive experiments on 6 real-world networks to demonstrate the advantage of proposed techniques.

DeepStyle: User Style Embedding for Authorship Attribution of Short Texts

Zhiqiang Hu[1], Roy Ka-Wei Lee[2], Lei Wang[3], Ee-peng Lim[3], Bo Dai[1]
[1]University of Electronic Science and Technology of China, [2]University of Saskatchewan, [3]Singapore Management University,

Abstract. Authorship attribution (AA), which is the task of finding the owner of a given text, is an important and widely studied research topic with many applications. Recent works have shown that

deep learning methods could achieve significant accuracy improvement for the AA task. Nevertheless, most of these proposed methods represent user posts using a single type of features (e.g., word bi-grams) and adopt a text classification approach to address the task. Furthermore, these methods offer very limited explainability of the AA results. In this paper, we ad-dress these limitations by proposing DeepStyle, a novel embedding-based framework that learns the representations of users' salient writing styles.We conduct extensive experiments on two real-world datasets from Twitter and Weibo. Our experiment results show that DeepStyle outperforms the state-of-the-art baselines on the AA task.

Content Sharing Prediction for Device-to-Device (D2D)-based Offline Mobile Social Networks by Network Representation Learning

Qing Zhang[1], Xiaoxu Ren[1], Yifan Cao[1], Hengda Zhang[1], Xiaofei Wang[1], Victor Leung[2]
[1]Tianjin University, [2]The University of British Columbia

Abstract. With the explosion of cellular data, the content sharing in proximity among offline Mobile Social Networks (MSNs) has received significant attention. It is necessary to understand the face-to-face (e.g. Device-to-Device, D2D) social network structure and to predict content propagation precisely, which can be conducted by learning the low-dimensional embedding of the network nodes, called Network Representation Learning (NRL). However, most existing NRL models consider each edge as a binary or continuous value, neglecting rich information between nodes. Besides, many traditional models are almost based on small-scale datasets or online Internet services, severely confining their applications in D2D scenarios. Therefore, we propose ResNel, a RESCAL-based network representation learning model, which aims to regard the multi-dimensional relations as a probability in third-order (3D) tensor space and achieve more accurate predictions for both discovered and undiscovered relations in the D2D social network. Specifically, we consider the Global Positioning System (GPS) information as a critical relation slice to avoid the loss of potential information. Experiments on a realistic large-scale D2D dataset corroborate the advantages of improving forecast accuracy.

Fruited-Forest: A Reachability Querying Method Based on Spanning Tree Modelling of Reduced DAG

Liu Yang[1], Tingxuan Chen[1], Junyu Zhang[1], Jun Long[1], Zhigang Hu[1], Victor S. Sheng[2]
[1]Central South University, [2]Texas Tech University

Abstract. A reachability query is a fundamental graph operation in real graph applications, which answers whether a node can reach another node through a path in a graph. However, the increasingly large amounts of real graph data make it more challenging for query efficiency and scalability. In this paper, we propose a Fruited-Forest (FF) approach to accelerate reachability queries in large graphs by constructing four kinds of fruited-forests from a reduced DAG in different traversal orders. We build different binary-label schemes for the four kinds of fruited-forests to cover reachability between nodes as much as possible, and create a corresponding index for the deleted edges which are deleted during the construction of fruited-forests. Our experimental results on 18 large real graph

datasets show that our FF approach requires less index construct time and a smaller index size, which is more scalable to answer reachability queries compared with other existing works.

IASR: an Item-level Attentive Social Recommendation Model for Personalized Ranking
Tianyi Tao, Yun Xiong, Guosen Wang, Yao Zhang, Peng Tian, Yangyong Zhu
Fudan University
Abstract. Most recommender systems provide recommendations by listing the most relevant items to a user. Such recommendation task can be viewed as a personalized ranking problem. Previous works have found it advantageous to improve recommendation performance by incorporating social information. However, most of them have two primary defects. First, in order to model interaction between users and items, existing works still resort to biased inner production, which has proved less expressive than neural architectures. Second, they do not delicately allocate weights of social neighbor influence based on the user feature or the item feature in a recommendation task. To address the issues, we propose an Item-level Attentive Social Recommendation model, IASR for short, in this paper. It employs an item-level attention mechanism to adaptively allocate social influences among trustees in the social network and gives more accurate predictions with a neural collaborative filtering framework. Extensive experiments on three real-world datasets are conducted to show our proposed IASR method out-performs the state-of-the-art baselines. Additionally, our method shows effectiveness in the cold-start scenario.

Aligning Users Across Social Networks via Intra and Inter Attentions
Zhichao Huang, Xutao Li, Yunming Ye
Harbin institute of Technology, Shenzhen
Abstract. In recent years, aligning users across different social networks receives a significant attention. Previous studies solve the problem based on attributes or topology structure approximation. However, most of them suffer from error propagation or the noise from diverse neighbors. To address the drawback, we design intra and inter attention mechanisms to model the influence of neighbors in local and across networks. In addition, to effectively incorporate the topology structure information, we leverage neighbors from labeled pairs instead of these in original networks, which are termed as matched neighbors. Then we treat the user alignment problem as a classification task and predict it upon a deep neural network. We conduct extensive experiments on six real-world datasets, and the results demonstrate the superiority of the proposed method against state-of-the-art competitors.

Research Session 15: Spatial-Temporal Databases

Time: 15:50-18:00, September 19, 2020, Saturday
Chair: Lu Chen, Zhejiang University

TKGFrame: A Two-Phase Framework for Temporal-Aware Knowledge Graph Completion
Jiasheng Zhang[1,2], Yongpan Sheng[2], Zheng Wang[2,3], Jie Shao[2,4]

[1]Guizhou University, [2]University of Electronic Science and Technology of China
[3]Institute of Electronic and Information Engineering of UESTC,
[4]Sichuan Artificial Intelligence Research Institute
Abstract. In this paper, we focus on temporal-aware knowledge graph (TKG) completion, which aims to automatically predict missing links in a TKG by making inferences from the existing temporal facts and the temporal information among the facts. Existing methods conducted on this task mainly focus on modeling temporal ordering of relations contained in the temporal facts to learn the low-dimensional vector space of TKG. However, these models either ignore the evolving strength of temporal ordering relations in the structure of relational chain, or discard more consideration to the revision of candidate prediction results produced by the TKG embeddings. To address these two limitations, we propose a novel two-phase framework called TKGFrame to boost the final performance of the task. Specifically, TKGFrame employs two major models. The first one is a relation evolving enhanced model to enhance evolving strength representations of pairwise relations pertaining to the same relational chain, resulting in more accurate TKG embeddings. The second one is a refinement model to revise the candidate predictions from the embeddings and further improve the performance of predicting missing temporal facts via solving a constrained optimization problem. Experiments conducted on three popular datasets for entity prediction and relation prediction demonstrate that TKGFrame achieves more accurate prediction results as compared to several state-of-the-art baselines.

Fine-grained Urban Flow Prediction via a Spatio-Temporal Super-Resolution Scheme
Rujia Shen, Jian Xu, Qing Bao, Wei Li, Hao Yuan, Ming Xu
Hangzhou Dianzi University
Abstract. Urban flow prediction plays an essential role in public safety and traffic scheduling for a city. By mining the original granularity flow data, current research methods could predict the coarse-grained region flow. However, the prediction of a more fine-grained region is more important for city management,which means cities could derive more details from the original granularity flow data. In this paper, given the future weather information, we aim to predict the fine-grained region flow. We design Weather-affected Fine-grained Region Flow Predictor (WFRFP) model based on the super-resolution scheme. Our model consists of three modules: 1) Key flow maps selection module selects key flow maps from massive historical data as the input instance according to temporal property and weather similarity; 2) Weather condition fusion module processes the original weather information and extracts weather features; 3) Fine-grained flow prediction module learns the spatial correlations by wide activation residual blocks and predicts the fine-grained region flow by the upsampling operation. Extensive experiments on a real-world dataset demonstrate the effectiveness and efficiency of our method, and show that our method outperforms the state-of-the-art baselines.

Detecting Abnormal Congregation through the Analysis of Massive Spatio-Temporal Data
Tianran Chen, Yongzheng Zhang, Yupeng Tuo, Weiguang Wang
Institute of Information Engineering, Chinese Academy of Sciences

Abstract. The pervasiveness of location-acquisition technologies leads to large amounts of spatio-temporal data, which brings researchers opportunities to discover interesting group patterns like congregation. Typically, a congregation is formed by a certain number of individuals within an area during a period of time. Previous work focused on discovering various congregations based on real-life scenarios to help in monitoring unusual group activities. However, most existing research didn't further analyze these results due to the consideration that the congregation is an unusual event already. In this article, firstly, we propose a group pattern to capture a variety of congregations from trajectory data. Secondly, congregations are separated into unexpected congregations and periodic congregations by extracting spatio-temporal features from historical trajectories. Thirdly, we further investigate the intensity of periodic congregation and combine environmental factors to dynamically identify anomalies within it, together with previously obtained unexpected congregations to form abnormal congregations. Moreover, incremental update techniques are utilized to detect abnormal congregations over massive-scale trajectory streams online, which means it can immediately respond to the updated trajectories. Finally, based on real cellular net-work dataset and real taxi trajectory dataset, our approach is evaluated through extensive experiments which demonstrate its effectiveness.

Efficient Semantic Enrichment Process for Spatiotemporal Trajectories in Geospatial Environment

Jingjing Han, Mingyu Liu, Genlin Ji, Bin Zhao, Richen Liu, Ying Li
NanJing Normal University

Abstract. The existing semantic enrichment process approaches which can produce semantic trajectories, are generally time consuming. In this paper, we propose a semantic enrichment process framework for spatiotemporal trajectories in geospatial environment. It can derive new semantic trajectories through the three phases: pre-annotated semantic trajectories storage, spatiotemporal similarity measurement, and semantic information matching. Having observed the common trajectories in the same geospatial object scenes,we propose an algorithm to match semantic information in pre-annotated semantic trajectories to new spatiotemporal trajectories. Finally, we demonstrate the effectiveness and efficiency of our proposed approach by using the real dataset.

An Effective Constraint-based Anomaly Detection Approach on Multivariate Time Series

Zijue Li[1], Xiaoou Ding[1], Hongzhi Wang[1,2]
[1]Harbin Institute of Technology, [2]Peng Cheng Laboratory

Abstract. With the development of IoT, various sensors are deployed in industry applications. Sensors produce multivariate time series, while error data and abnormal values often exist in the data. Correlation in multivariate time series can be used to identify such anomaly. In this paper, we propose an efficient method to utilize the correlation between multivariate time series with constraint-based anomaly detection. We develop a DP algorithm to execute the detection process, and optimize the algorithm efficiency with 2D range tree. Experiments on real IIoT dataset demonstrate the superiority of our proposed method compared to the prediction based models.

Sorting-based Interactive Regret Minimization

Jiping Zheng, Chen Chen
Nanjing University of Aeronautics and Astronautics

Abstract. As an important tool for multi-criteria decision making in database systems, the regret minimization query is shown to have the merits of top-k and skyline queries: it controls the output size while does not need users to provide any preferences. Existing researches verify that the regret ratio can be much decreased when interaction is available. In this paper, we study how to enhance current interactive regret minimization query by sorting mechanism. Instead of selecting the most favorite point from the displayed points for each interaction round, users sort the displayed data points and send the results to the system. By introducing sorting mechanism, for each round of interaction the utility space explored will be shrunk to some extent. Further the candidate points selection for following rounds of interaction will be narrowed to smaller data spaces thus the number of interaction rounds will be reduced. We propose two effective sorting-based algorithms namely Sorting-Simplex and Sorting-Random to find the maximum utility point based on Simplex method and randomly selection strategy respectively. Experiments on synthetic and real datasets verify our Sorting-Simplex and Sorting-Random algorithms outperform current state-of-art ones.

Seamless Incorporation of Appointment-based Requests on Taxi-Rider Match Scheduling

Yongxuan Lai, Shipeng Yang, Anshu Xiong, Fan Yang
Xiamen University

Abstract. Rider demand responsive systems (RDRS) makes a match between numerous requests and vehicles, it is a challenging problem to make the maximal match as soon as the real-time requests pop up in the RDRS. Much research has been addressed on this issue. However, there is still not much work on handling the appointment-based requests. In this paper, we propose an algorithm called BMF (Bipartite Minimal-cost Flow) to solve the taxi-rider match scheduling problem with appointment-based rider requests on a time-dependent road network. Riders and vehicles are modeled as vertices in a bipartite graph, and the maximal utility calculation is transformed to the minimal cost flow problem that could be solved efficiently. Experimental results show that the proposed scheme can effectively decrease the average waiting time of riders (> 44% reduction) at the cost of acceptable increase on the running time.

Demo Session

Time: 15:50-18:00, September 18, 2020, Friday
Chair: Hongzhi Wang, Harbin Institute of Technology
Xin Huang, Hong Kong Baptist University

A Meta-Search Engine Ranking based on Webpage Information Quality Evaluation

Yukun Li, Yunbo Ye, Wenya Xu
Tianjin University of Technology

Abstract. This paper demonstrates a meta-search engine developed by the authors, which ranks the results based on web page information quality evaluation algorithm. The web page information quality score is calculated based on the title of the web page, the abstract of the web page and the source of the web page. The quality of web page can be evaluated by these factors. When a user submits an input, the proposed meta-search engine system collects the results from some general search engines like Baidu, Bing, Sogou and so on, and rank the web pages according to their information quality scores. Because we do not need a local database to store a large amount of data, all operations are completed in the cache, which greatly reduces system consumption.The system is evaluated by three kinds of representative queries, and the results show that its search accuracy and user experience are obviously better than the current general search engines.

Automatic Document Data Storage System Based on Machine Learning

Yu Yan[1], Hongzhi Wang[1,2], Jian Zou[1], Yixuan Wang[1]
[1]Harbin Institute of Technology, [2]Peng Cheng Laboratory

Abstract. Document storage management plays a significant role in the field of database. With the advent of big data, making storage management manually becomes more and more difficult and inefficient. There are many researchers to develop algorithms for automatic storage management (ASM). However, at present, no automatic systems or algorithms related to document data has been developed. In order to realize the ASM of document data, we firstly propose an automatic document data storage system (ADSML) based on machine learning, a user-friendly management system with high efficiency for achieving storage selection and index recommendation automatically. In this paper, we present the architecture and key techniques of ADSML, and describe three demo scenarios of our system.

Epidemic Guard: A COVID-19 Detection System for Elderly People

Wenqi Wei, Jianzong Wang, Ning Cheng, Yuanxu Chen, Bao Zhou, Jing Xiao
Ping An Technology (Shenzhen) Co., Ltd

Abstract. The global outbreak of the COVID-19 in the worldwide has drawn lots of attention recently. The elderly are more vulnerable toCOVID-19 and tend to have severe conditions and higher mortality as their immune function decreased and they are prone to having multiple chronic diseases. Therefore, avoiding viral infection, early detection and treatment of viral infection in the elderly are important measures to protect the safety of the elderly. In this paper, we propose a real-time robot-based COVID-19 detection system: Epidemic Guard. It combines speech recognition, keyword detection, cough classification, and medical services to convert real-time audio into structured data to record the user's real condition. These data can be further utilized by the rules engine to provide a basis for real-time supervision and medical services. In addition, Epidemic Guard comes with a powerful pretraining model to effectively customize the user's health status.

A New CPU-FPGA Heterogeneous gStore System

Xunbin Su, Yinnian Lin, Lei Zou

Peking University

Abstract. In this demonstration, we present a new CPU-FPGA heterogeneous gStore system. The previous gStore system is based on CPU and has low join query performance when the data size is too big. We implement a FPGA-based join module to speed up join queries. Furthermore,we design a FPGA-friendly data structure called FFCSR to facilitate it. We compare our new system with the previous one on the LUBM2Bdataset. Experimental results demonstrate that the new CPU-FPGA heterogeneous system performs better than the previous one based on CPU.

Euge: Effective Utilization of GPU Resources for Serving DNN-based Video Analysis

Qihang Chen, Guangyao Ding, Chen Xu, Weining Qian, Aoying Zhou
East China Normal University

Abstract. Deep Neural Network (DNN) has been widely adopted in video analysis application. The computation involved in DNN is more efficient on GPUs than on CPUs. However, recent serving systems involve the low utilization of GPU, due to limited process parallelism and storage overhead of DNN model. We propose Euge, which introduces multi-process service (MPS) and model sharing technology to support effective utilization of GPU. With MPS technology, multiple processes overcome the obstacle of GPU context and execute DNN-based video analysis on one GPU in parallel. Furthermore, by sharing the DNN-based model among threads within a process, Euge reduces the GPU memory overhead. We implement Euge on Spark and demonstrate the performance of vehicle detection workload.

Blockchain PG: Enabling Authenticated Query and Trace Query in Database

Qingxing Guo, Sijia Deng, Lei Cai, Yanchao Zhu, Zhao Zhang, Cheqing Jin
East China Normal University

Abstract. Blockchain comes under the spotlight for successfully implementing a tamper-resistant ledger among multiple untrusted participants. The widespread adoption of blockchain in data-intensive applications has led to the demand for querying data stored in blockchain databases. However, compared to traditional databases, current blockchain systems cannot offer efficient queries. Moreover, migrating the original business supported by traditional databases to blockchain systems takes a long time and costs a lot. Motivated by this, we design and implement Blockchain PG, a novel data management system built on a legacy system. The system architecture applies to most databases system. By establishing a trusted relationship between the database and the client, Blockchain PG not only guarantees that the existing legacy system will not be greatly affected, but also achieves the data integrity and correctness by authenticated query, and the trace ability of data by trace query.

PHR:A Personalized Hidden Route Recommendation System based on Hidden Markov Model

Yundan Yang, Xiao Pan, Xin Yao, Shuhai Wang, Lihua Han
Shijiazhuang Tiedao University

Abstract. Route recommendation based on users' historical trajectories and behavior preferences is

one of the important research problems. However, most of the existing work recommends a route based on the similarity among the routes in historical trajectories. As a result, hidden routes that also meet the users' requirements cannot be explored. To solve this problem, we developed a system PHR that can recommend hidden routes to users employing the Hidden Markov Model, where a route recommendation problem is transformed to a point-of-in-terested (POI) sequence prediction. The system can return the top-k results including both explicit and hidden routes considering the personalized category sequence, route length, POI popularity, and visiting probabilities. The real check-in data from Foursquare is employed in this demo. The research can be used for travel itinerary plan or routine trip plan.

JoyDigit NexIoT: An Open IoT Data Platform for Senior Living

Kai Zhao[1], Peibiao Yang[1], Peng Zhang[1], Sufang Wang[1], Feng Wang[1], Xu Liu[1], Hongyan Deng[2]
[1]JoyDigit, [2]Zaozhuang University

Abstract. The senior care service plays an important role in senior living industry.With the unprecedented increasing of the seniors and the demanding activities of daily living (ADL) service,caregiver's workload and community's operational cost grow dramatically which lead tremendous challenges.From both caregivers and community operators' perspective,we developed an open IoT data platform for senior living named Joy-Digit NexIoT to solve above problems. Multidimensional open IoT data from smart home,health care equipment and security monitoring devices are integrated and processed in real-time on the platform to build dynamical profiles for both seniors and communities,and it also provides open data APIs for customized senior living applications. Together with the novel JoyDigit Intelligent Action Adviser(JIAA), the platform out-puts the recommended service or management actions to caregivers and operators, which could greatly reduce the caregiver's daily workload and increase community operator's management efficiency. In this paper, we describe JoyDigit NexIoT platform's features,architecture and JIAA, and also present the practical scenarios to be demonstrated.

Workshops

The Third International Workshop on Knowledge Graph Management and Applications (KGMA 2020)

Time: 9:00-12:00, September 20, 2020, Sunday
Co-Chairs: Zhuoming Xu, Hohai University, China
Saiful Islam, Griffith University, Australia
Xin Wang, Tianjin University, China

Invited Talk: Towards Knowledge Graph Federations: Issues and Technologies
Abstract: In enterprises at scale, knowledge bases tend to be constructed by different departments along with the development of their own information systems, which hinders the use of knowledge at the enterprise level. To overcome the limit, on top of solitary knowledge graphs, we present the concept of knowledge graph federation with the aim of uniting knowledge graphs to boost intelligent services at the enterprise level. In this talk, we will discuss the notion of knowledge graph federation, as well as current issues in the direction. Particularly, relating to the issues, we will share some recent progress on constructing domain knowledge graphs from noisy Chinese text, and benchmarking the performance of knowledge graphs fusion via entity alignment.

Xiang Zhao
Professor, National University of Defense Technology
Speaker Bio: Dr. Xiang Zhao is an associate professor at College of Systems Engineering, National University of Defense Technology, China, where he also serves as the head of the direction of knowledge systems engineering. His research focus is to find effective and efficient solutions for managing, integrating and analyzing very large amount of complex data for business, scientific and personal applications. He has been working in the area of graph management and mining, knowledge graph construction, and recommendation. He is a regular invited reviewer for journals including IEEE TKDE, World Wide Web Journal, Information Sciences, and serves on the program committees including VLDB, ICDE, AAAI and CIKM. He was a recipient of ACM SIGMOD China Rising Star award, Young Talents Lift-up Program for China Association for Science and Technology, and Hunan Provincial Natural Science Fund for Distinguished Young Scholars.

Accepted Papers:

- **Method for Re-finding Mobile Phone Documents Based on Feature Knowledge Graph**

Jing Zhang[1], Yukun Li[1,2], Yan Zhang[1], Yunbo Ye[1]
[1]Tianjin University of Technology, China
[2]Tianjin Key Laboratory of Intelligence Computing and Novel Software Technology, China

• **Knowledge Graph Embedding with Relation Constraint**

Chunming Yang[1,4], Xinghao Song[3], Hui Zhang[2], Bo Li[1]
[1]School of Computer Science and Technology, Southwest University of Science and Technology, China
[2]School of Science, Southwest University of Science and Technology, China
[3]Sichuan Branch of China Telecom Co.,Ltd., China
[4]Sichuan Big Data and Intelligent System Engineering Technology Research Center, China

• **Knowledge-driven Multi-dimensional Dialogue Rewriting Model**

Xiangwei Guo[1], Yongli Wang[1], Gang Xiao[2], Feifei Ma[3]
[1]Nanjing University of Science and Technology, China
[2]Science and Technology on Complex Systems Simulation Laboratory, China
[3]Nanjing Power Supply Branch of State Grid Jiangsu Electric Power Co., Ltd., China

The Second International Workshop on Semi-structured Big Data Management and Applications (SemiBDMA 2020)

Time: 9:00-12:00, September 20, 2020, Sunday
Chairs: Qun Chen, Northwestern Polytechnical University, China
Jianxin Li, Deakin University, Australia

Accepted Papers:

- **A Learning Interests Oriented Model for Cold Start Recommendation**

Yuefeng Du, Tuoyu Yan, Xiaoguang Li, Baoyan Song
Liaoning University

- **Second-degree branch chain efficient polymorphic storage structure**

Junlu Wang, Qiang Liu, Linlin Ding, Baoyan Song
Liaoning University

- **A Composite Chain Structure Blockchain Storage Method Based on Blockchain Technology**

Junlu Wang, Su Li, Linlin Ding, Baoyan Song
Liaoning University

- **Event Arrangement Method Based on Multi-factor Features and User Feedback in EBSN**

Xiaohuan Shan, Xinao Qi, Zhiguo Zhang and Baoyan Song
Liaoning University

- **Curriculum-Oriented Multi-Goal Agent for Adaptive Learning**

Jieyue Ma[1], Xiaoguang Li[1], Xin Zhang[1], Tingting Liu[1], Yuefeng Du[1], Tie Li[2]
[1]Liaoning University, [2]Shenyang AeroTech Co. Ltd.

- **Distributed Storage and Query for Domain Knowledge Graphs**

Xiaohuan Shan, Xiyi Shi and Baoyan Song
Liaoning University

The 1st International Workshop on Deep Learning in Large-scale Unstructured Data Analytics (DeepLUDA 2020)

Time: 14:00-17:00, September 20, 2020, Sunday
Organizers: Tae-Sun Chung, Ajou University, Korea
Rize Jin, Tiangong University, China
Chair: Joon-Young Paik, Tiangong University, China

Invited Talk:
Prof. Jianming Wang
• Director, Tianjin International Joint Research and Development Center of Autonomous Intelligence Technology and Systems
• Dean, School of Computer Science & Technology, Tiangong University

Accepted Papers:
• **Label Propagation Algorithm Based on Topological Potential**
Wang Guocheng, Xia Zhengyou

• **LBNet: A Model for Judicial Reading Comprehension**
Hao Liu, Jungang Xu

• **Deep Semantic Hashing for Large-scale Image Retrieval**
Yang Yulin, Jin Rize, Xu Cai