The 26th International Conference on Database Systems for Advanced Applications

Online Conference

April 11-14, 2021

DASFAA 2021

Welcome Message

Welcome to DASFAA 2021, the 26th International Conference on Database Systems for Advanced Application, held from April 11 to April 14, 2021! The conference was originally planned to be held in Taipei, Taiwan. Due to the outbreak of the COVID-19 pandemic and the consequent health concerns and restrictions on international travel all over the world, this prestigious event eventually happens on-line as a virtual conference, thanks for the tremendous effort made by authors, participants, technical program committee, organization committee, and steering committee. While the traditional in-person, face-to-face research exchanges and social interactions in the DASFAA community is temporarily paused this year, the long and successful history of the events, which established DASFAA as a premier research conference in the database area, continues!

On behalf of the program committee, it is our great pleasure to present the proceedings of DASFAA 2021, which includes 135 papers in the research track, 8 papers in the industrial track, 8 demo papers, and 4 tutorials. In addition, the conference program included three keynote presentations by Prof. Beng Chin Ooi from National University of Singapore, Singapore, Prof. Jiawei Han from the University of Illinois at Urbana-Champaign, USA, and Dr. Eunice Chiu, Vice President of NVidia, Taiwan.

The highly selective papers in the DASFAA 2021 proceedings report the latest and most exciting research results from academia and industry in the general area of database systems for advanced applications. The quality of the accepted research papers at DASFAA 2021 is extremely high, owing to a robust and rigorous double-blind review process (supported by the Microsoft CMT system). This year, we received 490 excellent submissions, of which 98 full papers (acceptance ratio of 20%) and 37 short papers (acceptance ratio of 27.6%) were accepted. The selection process was competitive and thorough. Each paper received at least 3 reviews with some papers receiving as many as 4-5 reviews, followed by a discussion, and then further evaluated by a senior program committee (SPC) member. We, the TPC co-chairs, considered the recommendations from the SPC

members and looked into each submission as well as its reviews and discussions to make the final decisions, which took into account multiple factors such as depth and novelty of technical content and relevance to the conference. The most popular topic areas for the selected papers include information retrieval, search and recommendation techniques; RDF, knowledge graphs, semantic web, and knowledge management; and Spatial, temporal, sequence, and streaming data management, while the dominant keywords are network, recommendation, graph, learning, and model. These topic areas and keywords shed the light on the direction where the research in DASFAA is moving towards.

Five workshops are held in conjunction with DASFAA 2021, including International Workshop on Machine Learning and Deep Learning for Data Security Applications (MLDLDSA 2021), International Workshop on Mobile Data Management, Mining, and Computing on Social Networks (Mobisocial 2021), International Workshop on Big Data Quality Management (BDQM 2021), International Workshop on Mobile Ubiquitous Systems and Technologies (MUST 2021), and International Workshop on Graph Data Management and Analysis (GDMA 2021). The workshop papers are included in a separate volume of the proceedings, also published by Springer in its Lecture Notes in Computer Science series.

We would like to express our sincere gratitude to all of the 43 SPC members, 278 PC members, and many external reviewers for their hard work in providing us with comprehensive and insightful reviews and recommendations. Many thanks to all the authors for submitting their papers, which contributes significantly to the technical program and the success of the conference. We are grateful to the general chairs, Christian S Jensen, Ee-Peng Lim, and De-Nian Yang for their help. We wish to thank everyone who contribute to the proceedings, including the Jianliang Xu, Chia-Hui Chang and Wen-Chih Peng (workshop chairs), Xing Xie and Shou-De Lin (Industrial program chairs), Wenjie Zhang, Wook-Shin Han and Hung-Yu Kao (demonstration chairs), Ying Zhang and Mi-Yen Yeh (tutorial chairs), as well as the organizers of workshops, their respective PC members and reviewers.

We are also grateful to all the members of the Organizing Committee and many volunteers for their tireless work before and during the conference. Also, we would like to express our sincere thanks to Chih-Ya Shen and Jen-Wei Huang (proceedings chairs) for working with the Springer team to produce the proceedings. Special thanks go to Xiaofang Zhou and Sourav S Bhowmick for their guidance. Lastly, we acknowledge the generous financial support from various industrial companies and academic institutes.

We hope that you will enjoy the DASFAA 2021 conference, its technical program and the proceedings!

DASFAA 2021 Organizing Committee

February 2021

Organizing Committee

Honorary Chairs
Philip S. Yu, University of Illinois at Chicago, USA
Ming-Syan Chen, National Taiwan University, Taiwan
Masaru Kitsuregawa, University of Tokyo, Japan

General Chairs
Christian S Jensen, Aalborg University, Denmark
Ee-Peng Lim, Singapore Management University, Singapore
De-Nian Yang, Academia Sinica, Taiwan

Program Committee Chairs
Wang-Chien Lee, Pennsylvania State University, USA
Vincent S. Tseng, National Chiao Tung University, Taiwan
Vana Kalogeraki, Athens University of Economics and Business, Greece

Industrial Program Chairs
Xing Xie, Microsoft Research Asia, China
Shou-De Lin, Appier, Taiwan

Demo Chairs
Wenjie Zhang, The University of New South Wales, Australia
Wook-Shin Han, Pohang University of Science and Technology, Korea
Hung-Yu Kao, National Cheng Kung University, Taiwan

Tutorial Chairs
Ying Zhang, University of Technology, Sydney, Australia
Mi-Yen Yeh, Academia Sinica, Taiwan

Workshop Chairs
Chia-Hui Chang, National Central University, Taiwan
Jianliang Xu, Hong Kong Baptist University, Hong Kong
Wen-Chih Peng, National Chiao Tung University, Taiwan

Panel Chairs
Zi Huang, The University of Queensland, Australia
Takahiro Hara, Osaka University, Japan
Shan-Hung Wu, National Tsing Hua University, Taiwan

Ph.D Consortium
Lydia Chen, Technology University Delft, Netherlands
Kun-Ta Chuang, National Cheng Kung University, Taiwan

Publicity Chairs
Wen Hua, The University of Queensland, Australia
Yongxin Tong, Beihang University, China
Jiun-Long Huang, National Chiao Tung University, Taiwan

Proceedings Chairs
Jen-Wei Huang, National Cheng Kung University, Taiwan
Chih-Ya Shen, National Tsing Hua University, Taiwan

Registration Chair
Chuan-Ju Wang, Academia Sinica, Taiwan
Hong-Han Shuai, National Chiao Tung University, Taiwan

Sponsor Chair
Chih-Hua Tai, National Taipei University, Taiwan

Web Chair
Ya-Wen Teng, Academia Sinica, Taiwan
Yi-Cheng Chen, National Central University, Taiwan

Finance Chairs
Yi-Ling Chen, National Taiwan University of Science and Technology, Taiwan

Local Arrangement Chairs
Chien-Chin Chen, National Taiwan University, Taiwan
Chih-Chieh Hung, National Chung Hsing University, Taiwan

Program Schedule

11 Apr. 2021 – Workshops

09:00 – 12:00	MUST	BDQM	
12:00 – 14:00	Lunch		
14:00 – 17:00	GDMA	MobiSocial	MLDLDSA

12 Apr. 2021

09:00 – 09:30	Main Conference Opening				
09:30 – 10:30	Keynote 1				
10:30 – 11:00	Coffee Break				
11:00 – 12:40	Graph Data 1	Spatial/Temp-oral Data 1	Text and Unstructured Data 1	Recommen-dation 1	Emerging Applications 1
12:40 – 14:00	Lunch				
14:00 – 15:40	Data Mining 1	Machine Learning 1	Information Retrieval and Search	Tutorial 1-1	Tutorial 2-1
15:40 – 16:00	Coffee Break				
16:00 – 17:40	Big Data 1	Social Network	Demo-1	Tutorial 1-2	Tutorial 2-2

13 Apr. 2021

09:00 – 10:00	Keynote 2				
10:00 – 10:30	Coffee Break				
10:30 – 12:10	Graph Data 2	Spatial/Temp-oral Data 2	Text and Unstructured Data 2	Panel Discussion	
12:10 – 14:00	Lunch				
14:00 – 15:40	Machine Learning 2	Recommend-ation 2	Text and Unstructured Data 3	PhD Consortium	Industry-1
15:40 – 16:00	Coffee Break				
16:00 – 17:40	Emerging Applications 2	Recommend-ation 3	Data Mining 2	Demo-2	Industry-2

14 Apr. 2021

09:00 – 10:00	Keynote 3				
10:00 – 10:30	Coffee Break				
10:30 – 12:10	Graph Data 3	Spatial/Temp-oral Data 3	Recommend-ation 4		
12:10 – 14:00	Lunch				
14:00 – 15:40	Text and Unstructured Data 4	Spatial/Temp-oral Data 4	Recommend-ation 5	Tutorial 3-1	Tutorial 4-1
15:40 – 16:00	Coffee Break				
16:00 – 17:40	Graph Data 4	Big Data 2	Query Processing	Tutorial 3-2	Tutorial 4-2

Keynote Details

1. **What Can AI do for End-to-End Data Analytics?**
Speaker: *Beng Chin Ooi (Distinguished Professor, National University of Singapore)*
Time: Apr. 12, 2021, 9:30 a.m. – 10:30 a.m. (GMT+8 Taipei Time)

2. **Transforming Unstructured Text into Structured Knowledge: A Text-Mining Approach**
Speaker: *Jiawei Han (Michael Aiken Chair Professor, The University of Illinois at Urbana-Champaign)*
Time: Apr. 13, 2021, 9:00 a.m. – 10:00 a.m. (GMT+8 Taipei Time)

3. **AI Is Transforming Industries**
Speaker: *Eunice Chiu (VP at NVIDIA)*
Time: Apr. 14, 2021, 9:00 a.m. – 10:00 a.m. (GMT+8 Taipei Time)

Panel Details

Database Meets Artificial Intelligence: Challenges and Opportunities
Panelists: *Julia Stoyanovich (New York University, USA), Aaron Elmore (The University of Chicago, USA), Wei-Shinn Jeff Ku (Auburn University, USA), Makoto Onizuka (Osaka University, Japan), Haixun Wang (Instacart), and Dimitrios Gunopulos (National and Kapodistrian University of Athens, Greece)*
Time: Apr. 13, 2021, 10:30 a.m. – 12:10 p.m. (GMT+8 Taipei Time)

Tutorials Details

1. **Multi-Model Data, Query Languages and Processing Paradigms**
Speaker: *Qingsong Guo (University of Helsinki, Finland), Jiaheng Lu (University of Helsinki, Finland), and Chao Zhang (University of Helsinki, Finland)*
Date: Monday, Apr. 12, 2021 (GMT+8 Taipei Time)
Time: 2:00 p.m. - 3:40 p.m. (break) 4:00 p.m. - 5:40 p.m.

2. **Lightweight Deep Learning with Model Compression**
Speaker: *U Kang (Seoul National University, Seoul, South Korea)*
Date: Monday, Apr. 12, 2021 (GMT+8 Taipei Time)
Time: 2:00 p.m. - 3:40 p.m. (break) 4:00 p.m. - 5:40 p.m.

3. **Discovering Communities over Large Graphs: Algorithms, Applications, and Opportunities**
Speaker: *Chaokun Wang (Tsinghua University, China), Junchao Zhu (Tsinghua University, China), Zhuo Wang (Chinese Academy of Sciences, China), Yunkai Lou (Tsinghua University, China), Gaoyang Guo (Tsinghua University, China), and Binbin Wang (Tsinghua University, China)*
Date: Wednesday, Apr. 14, 2021 (GMT+8 Taipei Time)
Time: 2:00 p.m. - 3:40 p.m. (break) 4:00 p.m. - 5:40 p.m.

4. **AI Governance: Advanced Urban Computing on Dynamics Prediction and Route Planning**
Speaker: *Hsun-Ping Hsieh (National Cheng Kung University, Taiwan) and Fandel Lin (National Cheng Kung University, Taiwan)*
Date: Wednesday, Apr. 14, 2021 (GMT+8 Taipei Time)
Time: 2:00 p.m. - 3:40 p.m. (break) 4:00 p.m. - 5:40 p.m.

Session Details

Research Session 1: Graph Data 1
Session Chair: Rajeev Gupta (Microsoft)

1. Label Contrastive Coding based Graph Neural Network for Graph Classification
Yuxiang Ren, Jiyang Bai, and Jiawei Zhang
Abs. Graph classification is a critical research problem in many applications from different domains. In order to learn a graph classification model, the most widely used supervision component is an output layer together with classification loss (e.g.,cross-entropy loss together with softmax or margin loss). In fact, the discriminative information among instances are more fine-grained, which can benefit graph classification tasks. In this paper, we propose the novel Label Contrastive Coding based Graph Neural Network (LCGNN) to utilize label information more effectively and comprehensively. LCGNN still uses the classification loss to ensure the discriminability of classes. Meanwhile, LCGNN leverages the proposed Label Contrastive Loss derived from self- supervised learning to encourage instance-level intra-class compactness and inter-class separability. In order to power the contrastive learning, LCGNN introduces a dynamic label memory bank and a momentum updated encoder. Our extensive evaluations with 8 benchmark graph datasets demonstrate that LCGNN can outperform state-of-the-art graph classification models. Experimental results also verify that LCGNN can achieve competitive performance with less training data because LCGNN exploits label information comprehensively.

2. Which Node Pair and What Status? Asking Expert for Better Network Embedding
Longcan Wu, Daling Wang, Shi Feng, Kaisong Song, Yifei Zhang, and Ge Yu
Abs. In network data, the connection between a small number of node pair are observed, but for most remaining situations, the link status (i.e., connected or disconnected) of node pair can not be observed. If we can get more useful information hidden in node pairs with unknown link status, it will help improve the performance of network embedding. Therefore, how to model the network

with unknown link status actively and effectively remains an area for exploration. In this paper, we formulate a new network embedding problem, which is how to select valuable node pair (which node pair) to ask expert about their link status (what status) information for improving network embedding. To tackle this problem, we propose a novel active learning method called ALNE, which includes a proposed network embedding model AGCN, three active node pair selection strategies and an information evaluation module. In this way, we can obtain the real valuable link statuses information between node pairs and generate better node embeddings.Extensive experiments are conducted to show the effectiveness of ALNE.

3. Keyword-Centric Community Search over Large Heterogeneous Information Networks

Lianpeng Qiao, Zhiwei Zhang, Ye Yuan, Chen Chen, and Guoren Wang

Abs. Community search in heterogeneous information networks (HINs) has attracted much attention in recent years and has been widely used for graph analysis works. However, existing community search studies over heterogeneous information networks ignore the importance of keywords and cannot be directly applied to the keyword-centric community search problem. To deal with these problems, we propose $k\mathcal{KP}$-core, which is defined based on a densely-connected subgraph with respect to the given keywords set. A $k\mathcal{KP}$-core is a maximal set of $\mathcal{P}$-connected vertices in which every vertex has at least one $\mathcal{KP}$-neighbor and k path instances. We further propose three algorithms to solve the keyword-centric community search problem based on $k\mathcal{KP}$-core. When searching for answers, the basic algorithm Basic-$k\mathcal{KP}$-core will enumerate all paths rather than only the path instances of the given meta-path $\mathcal{P}$. To improve efficiency, we design an advanced algorithm $Advk\mathcal{KP}$-core using a new method of traversing the search space based on trees to accelerate the searching procedure. For online queries, we optimize the approach with a new index to handle the online queries of community search over HINs. Extensive experiments on HINs are conducted to evaluate both the effectiveness and efficiency of our proposed methods.

4. KGSynNet: A Novel Entity Synonyms Discovery Framework with Knowledge Graph

Yiying Yang, Xi Yin, Haiqin Yang, Xingjian Fei, Hao Peng, Kaijie Zhou, Kunfeng Lai, and Jianping Shen

Abs. Entity synonyms discovery is crucial for entity-leveraging applications. However, existing studies suffer from several critical issues: (1) the input mentions may be out-of-vocabulary (OOV) and may come from a different semantic space of the entities; (2) the connection between mentions and entities may be hidden and cannot be established by surface matching; and (3) some entities rarely appear due to the long-tail effect. To tackle these challenges, we facilitate knowledge graphs and propose a novel entity synonyms discovery framework, named KGSynNet. Specifically, we pre-train subword embeddings for mentions and entities using a large-scale domain-specific corpus while learning the knowledge embeddings of entities via a joint TransC-TransE model. More importantly, to obtain a comprehensive representation of entities, we employ a specifically designed fusion gate to adaptively absorb the entities' knowledge information into their semantic features. We conduct extensive experiments to demonstrate the effectiveness of our KGSynNet in leveraging the knowledge graph. The experimental results show that the KGSynNet improves the state-of-the-art methods by 14.7% in terms of hits@3 in the offline evaluation and outperforms the BERT model by 8.3% in the positive feedback rate of an online A/B test on the entity linking module of a question answering system.

5. Iterative Reasoning over Knowledge Graph

Liang Xu, and Junjie Yao

Abs. The concept reasoning is an essential task in text data management and understanding.Knowledge graphs have rich text information and connections. A semantic relation graph, is used to encode complex semantic relation between evidence and question. The nodes represent valuable information as clue entities and candidate answers in evidence and question, and the edges represent the logical rules between nodes. Recent methods usually capture shallow semantic features and cannot extend to multi-hop reasoning. In this paper, we propose a graph-based reasoning framework with iterative steps. The new approach iteratively infers the clue entities and candidate answers from the question and clue paragraphs to as new nodes to expand the semantic relation graph. Then we update the semantic representation of the questions and context via memory

network and apply the graph attention network to encode the logical rules of explicit paths in semantic relation graph. Extensive experiments on commonsense reasoning and multi-hop question answering verified the advantage and improvements of the proposed approach.

Research Session 2: Spatial/Temporal Data 1
Session Chair: Xiang Lian (Kent State University)

1. Online High-cardinality Flow Detection over Big Network Data Stream
Yang Du, He Huang, Yue Sun, An Liu, Guoju Gao, and Boyu Zhang
Abs. High-cardinality flow detection over the big network data stream plays an important role in many practical applications. To process large and fast data streams in real-time, most existing work uses compact data structures like sketches to fit themself in high-speed but small on-chip memory. However, this design suffers from expensive computation and thus only supports periodical high-cardinality flow detection. Although NDS can provide online flow cardinality estimation, it is designed to estimate all flows accurately. In contrast, high-cardinality flow detection only concerns whether a flow's cardinality exceeds a certain threshold. This paper complements the prior work by proposing an online high-cardinality flow detection method with high resource efficiency. Based on the on-chip/off-chip design, the proposed method reduces large flows' resource consumption by constructing a virtual bitmap sharing module over the physical bitmap. We evaluate the performance of the proposed method using the real-world Internet traces downloaded from CAIDA. The experimental results show that our method can save up to 65.8% on-chip memory when bounding the same constraints for false-positive rates and false-negative rates.

2. SCSG Attention: A Self-Centered Star Graph with Attention for Pedestrian Trajectory Prediction
Xu Chen, Shuncheng Liu, Zhi Xu, Yupeng Diao, Shaozhi Wu, Kai Zheng, and Han Su
Abs. Pedestrian trajectory prediction enables faster progress in autonomous driving and robot navigation where complex social and environmental interactions involve. Previous models use grid-based pooling or global attention

to measure social interactions and use Recurrent Neural Network (RNN) to generate sequences. However, these methods can not extract latent features from temporal and spatial information simultaneously. To address the limitation of previous work, we propose a Self-Centered Star Graph with Attention (SCSG Attention) framework. Firstly, pedestrians' historical trajectories are encoded. Then multi-head attention mechanism plays a role as enhancement of social interaction awareness and simulation of physical attention from human beings. Lastly, the self-centered star graph decoder can aggregate temporal and spatial features and make predictions. Experiments are conducted on public benchmark datasets and measured with uniform standards. Our results show an improvement over the state-of-the-art algorithms by 38% on average displacement error (ADE) and 19% on final displacement error (FDE). Furthermore, it is demonstrated that the star graph has better performance in efficiency of training convergence and ends up with better results in limited time.

3. Time Period-based Top-k Semantic Trajectory Pattern Query

Munkh-Erdene Yadamjav, Farhana Choudhury, Zhifeng Bao, and Baihua Zheng

Abs. The sequences of user check-ins form semantic trajectories that represent the movement of users through time, along with the types of POIs visited. Extracting patterns in semantic trajectories can be widely used in applications such as route planning and trip recommendation. Existing studies focus on the entire time duration of the data, which may miss some temporally significant patterns. In addition, they require thresholds to define the interestingness of the patterns. Motivated by the above, we study a new problem of finding top-k semantic trajectory patterns w.r.t. a given time period and categories by considering the spatial closeness of POIs. Specifically, we propose a novel algorithm, EC2M that converts the problem from POI-based to cluster-based pattern search and progressively consider pattern sequences with efficient pruning strategies at different steps. Two hashmap structures are proposed to validate the spatial closeness of the trajectories that constitute temporally relevant patterns. Experimental results on real-life trajectory data verify both the efficiency and effectiveness of our method.

4. Optimal Sequenced Route Query with POI Preferences

Wenbin Li, Huaijie Zhu, Wei Liu, Jian Yin, and Jianliang Xu

Abs. The optimal sequenced route (OSR) query, as a popular problem in route planning for smart cities, searches for a minimum-distance route passing through a number of POIs in a specific order from a starting position. In reality, POIs are usually rated, which helps users in making decisions. Existing OSR queries neglect the fact that the POIs in the same category could have different scores, which may affect users' route choices. In this paper, we study a novel variant of OSR query, namely Rating Constrained Optimal Sequenced Route query (RCOSR), in which the rating score of each POI in the optimal sequenced route should exceed the query threshold. To efficiently process RCOSR queries, we first extend the existing TD-OSR algorithm to propose a baseline method, called MTDOSR. To tackle the shortcomings of MTDOSR, we first try to adopt the dynamic programming to propose a new Optimal Subroute Expansion (OSE) Algorithm. To enhance the OSE algorithm, we propose a Reference Node Inverted Index (RNII) to accelerate the distance computation of POI pairs in OSE and quickly retrieve the POIs of each category. In addition, we develop a Greedy Merge (GM) strategy to determine the reference nodes for RNII. To make full use of the OSE and RNII, we further propose a new efficient RCOSR algorithm, called Recurrent Optimal Subroute Expansion (ROSE), which recurrently utilizes OSE to compute the current optimal route as the guiding path and update the distance of POI pairs to guide the expansion. The experimental results using real and synthetic datasets demonstrate that the proposed algorithm significantly outperforms the existing approaches.

5. Privacy-Preserving Polynomial Evaluation over Spatio-Temporal Data on An Untrusted Cloud Server

Wei Song, Mengfei Tang, Qiben Yan, Yuan Shen, Yang Cao, Qian Wang, and Zhiyong Peng

Abs. Nowadays, with the popularity of location-aware devices, multifarious applications based on the spatio-temporal data come forth in our lives. In these applications, the platform (enterprise) collects the users’ spatio-temporal data based on which it recommends the top-k most appropriate users to the registered service providers (drivers). Outsourcing the tremendous scale of spatio-temporal data to cloud provides an economical way for the enterprises to implement their applications. In this paradigm, the third-party cloud server is not completely trustworthy. The collected spatio-temporal data can hold users’ privacy, so it’s a

critical challenge to design a secure and efficient query mechanism for these applications, such as the ride-hailing and the ride sharing services. However, the existing solutions for the privacy-preserving kNN queries mainly focus on data privacy protection or computation complexity. There still lacks a practical privacy-preserving polynomial evaluation solution over the spatio-temporal data. In this paper, we propose a virtual road network structure to storage and index the spatio-temporal data in the road network and design a novel homomorphic encryption scheme based on Order-Revealing Encryption to achieve the privacy-preserving polynomial evaluation over the encrypted spatio-temporal data on an untrusted cloud server. We formally prove the security of the proposed scheme under the random oracle model. Extensive experiments on real world data demonstrate the effectiveness and efficiency of the proposed scheme over alternatives.

Research Session 3: Text and Unstructured Data 1
Session Chair: Sharma Chakravarthy (UT Arlington)

1. Multi-label Classification of Long Text Based on Key-sentences Extraction
Jiayin Chen, Xiaolong Gong, Ye Qiu, Xi Chen, and Zhiyi Ma
Abs. Most existing works on multi-label classification of long text task will perform text truncation preprocessing, which leads to the loss of label-related global feature information. Some approaches that split an entire text into multiple segments for feature extracting, which generates noise features of irrelevant segments. To address these issues, we introduce key-sentences extraction task with semi-supervised learning to quickly distinguish relevant segments, which added to multi-label classification task to form a multi-task learning framework. The key-sentences extraction task can capture global information and filter irrelevant information to improve multi-label prediction. In addition, we apply sentence distribution and multi-label attention mechanism to improve the efficiency of our model. Experimental results on real-world datasets demonstrate that our proposed model achieves significant and consistent improvements compared with other state-of-the-art baselines.

2. Automated Context-aware Phrase Mining from Text Corpora

Xue Zhang, Qinghua Li, Cuiping Li, and Hong Chen

Abs. Phrase mining aims to automatically extract high-quality phrases from a given corpus, which serves as the essential step in transforming unstructured text into structured information. Existing statistic-based methods have achieved the state-of-the-art performance of this task. However, such methods often heavily rely on statistical signals to extract quality phrases, ignoring the effect of contextual information. In this paper, we propose a novel context-aware method for automated phrase mining, ConPhrase, which formulates phrase mining as a sequence labeling problem with consideration of contextual information. Meanwhile, to tackle the global information scarcity issue and the noisy data filtration issue, our ConPhrase method designs two modules, respectively: 1) a topic-aware phrase recognition network that incorporates domain-related topic information into word representation learning for identifying quality phrases effectively. 2) an instance selection network that focuses on choosing correct sentences with reinforcement learning for further improving the prediction performance of phrase recognition network. Experimental results demonstrate that our ConPhrase outperforms the state-of-the-art approach.

3. Keyword-Aware Encoder for Abstractive Text Summarization

Tianxiang Hu, Jingxi Liang, Wei Ye, and Shikun Zhang

Abs. Text summarization aims to produce a brief statement covering main points. Human beings would intentionally look for key entities and key concepts when summarizing a text. Fewer efforts are needed to write a high-quality summary if keywords in the original text are provided. Inspired by this observation, we propose a keyword-aware encoder (KAE) for abstractive text summarization, which extracts and exploits keywords explicitly. It enriches word representations by incorporating keyword information and thus leverages keywords to distill salient information. We construct an attention-based neural summarizer equipped with KAE and evaluate our model extensively on benchmark datasets of various languages and text lengths. Experiment results show that our model generates competitive results comparing to state-of-the-art methods.

4. Neural Adversarial Review Summarization with Hierarchical Personalized Attention

Hongyan Xu, Hongtao Liu, Wenjun Wang, and Pengfei Jiao

Abs. Review summarization aims to generate condensed text for online product reviews. Existing methods always focus on word-level representation of reviews and ignore different informativeness of different sentences in a review towards summary generation. In addition, the personalized information along with reviews (e.g., user/product and ratings) is also highly related to the quality of generated summaries. Hence, we propose a review summarization method with hierarchical personalized attention including a review encoder and a summary decoder.The encoder contains a sentence encoder to learn sentence representations with word-level attention, and a review encoder to learn review representations with sentence-level attention. Both the two attentions are of personalized paradigm whose attention vectors are derived from personalized information of input reviews instead of randomly initialized. Thus, our encoder could focus on important words and sentences in the input review. Then a summary decoder is employed to generate target summaries with hierarchical attention likewise, where the decoding scores are not only related to word information, but re-weighted by another sentence-level attention. We further design an adversarial discriminator which takes the generated summary and personalized information as inputs to force the generator adapting the generation policy accordingly. Extensive experimental results show the effectiveness of our method.

5. Generating Contextually Coherent Responses by Learning Structured Vectorized Semantics

Yan Wang, Yanan Zheng, Shimin Jiang, Yucheng Dong, Jessica Chen, and David Wang

Abs. Generating contextually coherent responses has been one of the most critical challenges in building intelligent dialogue systems. Key issues are how to appropriately encode contexts and how to make good use of them during the generation. Past works either directly use (hierarchical) RNN to encode contexts or use attention-based variants to further weight different words and utterances. They tend to learn dispersed focuses on overall contextual information, which contradicts the fact that humans tend to respond to certain concentrated

semantics of contexts. This leads to the results that generated responses only show semantically related to, but not precisely coherent with the given contexts. To this end, this paper proposes a contextually coherent dialogue generation (ConDial) method by first encoding contexts into structured semantic vectors using self-attention, and then adaptively choosing key semantic vectors to guide the response generation. Based on the structured semantics, it also develops a calibration mechanism with a dynamic vocabulary during decoding, which enhances exact coherent expressions by adjusting word distribution. According to the experiments, ConDial shows better generative performance than state-of-the-arts and is capable of generating responses that not only continue the topics but also keep coherent contextual expressions.

Research Session 4: Recommendation 1

Session Chair: Chih-Hua Tai (National Taipei University)

1. Gated Sequential Recommendation System with Social and Textual Information under Dynamic Contexts

Haoyu Geng, Shuodian Yu, and Xiaofeng Gao

Abs. Recommendation systems are undergoing plentiful practices in research and industry to improve consumers' satisfaction. In recent years, many research papers leverage abundant data from heterogeneous information sources to grasp diverse preferences and improve overall accuracy. Some noticeable papers proposed to extract users' preference from information along with ratings such as reviews or social relations. However, their combinations are generally static and less expressive without considerations on dynamic contexts in users' purchases and choices. In this paper, we propose Heterogeneous Information Sequential Recommendation System (HISR), a dual-GRU structure that builds the sequential dynamics behind the customer behaviors, and combines preference features from review text and social attentional relations under dynamics contexts. A novel gating layer is applied to dynamically select and explicitly combine two views of data. Moreover, in social attention module, temporal textual information is brought in as a clue to dynamically select friends that are helpful for contextual purchase intentions as an implicit combination. We validate our proposed method on two large subsets of real-world local business dataset Yelp, and our method outperforms the state of the art methods on related tasks including social, sequential and heterogeneous recommendations.

2. SRecGAN: Pairwise Adversarial Training for Sequential Recommendation

Guangben Lu, Ziheng Zhao, Xiaofeng Gao, and Guihai Chen

Abs. Sequential recommendation is essentially a learning-to-rank task under special conditions. Bayesian Personalized Ranking (BPR) has been proved its effectiveness for such a task by maximizing the margin between observed and unobserved interactions. However, there exist unobserved positive items that are very likely to be selected in the future. Treating those items as negative leads astray and poses a limitation to further exploiting its potential. To alleviate such problem, we present a novel approach, Sequential Recommendation GAN (SRecGAN), which captures latent users' interests and predicts the next item in a pairwise adversarial manner. It can be interpreted as playing a minimax game, where the generator would learn a similarity function and try to diminish the distance between the observed samples and its unobserved counterpart, whereas the discriminator would try to maximize their margin. This intense adversarial competition provides increasing learning difficulties and constantly pushes the boundaries of its performance. Extensive experiments on three real-world datasets demonstrate the superiority of our methods over some strong baselines and prove the effectiveness of adversarial training in sequential recommendation.

3. SSRGAN: A Generative Adversarial Network for Streaming Sequential Recommendation

Yao Lv, Jiajie Xu, Rui Zhou, Junhua Fang, and Chengfei Liu

Abs. Studying the sequential recommendation in streaming settings becomes meaningful because large volumes of user-item interactions are generated in a chronological order. Although a few streaming update strategies have been developed, they cannot be applied in sequential recommendation, because they can hardly capture the long-term user preference only by updating the model with random sampled new instances. Besides, some latent information is ignored because the existing streaming update strategies are designed for individual interactions, without considering the interaction subsequence. In this paper, we propose a Streaming Sequential Recommendation with Generative Adversarial Network (SSRGAN) to solve the streaming sequential recommendation problem. To maintain the long-term memory and keep sequential information,

we use the reservoir-based streaming strorage mechanism and exploit an active subsequence selection strategy to update model. Moreover, to improve the effectiveness and efficiency of online model training, we propose a novel negative sampling strategy based on GAN to generate the most informative negative samples and use Gumble-Softmax to overcome the gradient block problem. We conduct extensive experiments on two real-world datasets and the results shows the superiority of our approaches in streaming sequential recommendation.

4. Topological Interpretable Multi-Scale Sequential Recommendation
Tao Yuan, Shuzi Niu, and Huiyuan Li
Abs. Sequential recommendation attempts to predict next items based on user historical sequences. However, items to be predicted next depend on user's long, short or mid-term interest. The multi-scale modeling of user interest in an interpretable way poses a great challenge in sequential recommendation. Hence, we propose a topological data analysis based framework to model target items' explicit dependency on previous items or item chunks with different time scales, which are easily changed into sequential patterns. First, we propose a topological transformation layer to map each user interaction sequence into persistent homology organized in a multi-scale interest tree. Then, this multi-scale interest tree is encoded to represent natural inclusion relations across scales through an recurrent aggregation process, namely tree aggregation block. Next, we add this block to the vanilla transformer, referred to as recurrent tree transformer, and utilize this new transformer to generate a unified user interest representation. The last fully connected layer is utilized to model the interaction between this unified representation and item embedding. Comprehensive experiments are conducted on two public benchmark datasets. Performance improvement on both datasets is averagely 5% over state-of-the-art baselines.

5. SANS: Setwise Attentional Neural Similarity Method for Few-shot Recommendation
Zhenghao Zhang, Tun Lu, Dongsheng Li, Peng Zhang, Hansu Gu, and Ning Gu
Abs. Recommender systems generate personalized recommendations for users based on their historical data. However, if some users have few interactions in the training data, i.e., few-shot users, recommendations for them will be

inaccurate. In this paper, we propose a setwise attentional neural similarity method (SANS) for the few-shot recommendation problem. Unlike general recommender models, we eliminate direct representations of few-shot users. First, a neural similarity method is proposed to effectively estimate the correlation between item pairs. Then, we propose a setwise attention mechanism to obtain recommendation scores by aggregating the correlations between a candidate item and items in a candidate user's historical interactions. To facilitate model training in the few-shot scenario, training samples are generated by episode sampling, and each training sample is assigned with an adaptive weight to emphasize the importance of few-shot users. We simulate the few-shot recommendation problem on three datasets and extensive results show that SANS can outperform the state-of-the-art recommendation algorithms in few-shot recommendation.

Research Session 5: Emerging Applications 1
Session Chair: Chih-Ya Shen (National Tsing Hua University)

1. PEEP: A Parallel Execution Engine for Permissioned Blockchain Systems
Zhihao Chen, Xiaodong Qi, Xiaofan Du, Zhao Zhang, and Cheqing Jin
Abs. Unlike blockchain systems in public settings, the stricter trust model in permissioned blockchain opens an opportunity for pursuing higher throughput. Recently, as the consensus protocols are developed significantly, the existing serial execution manner of transactions becomes a key factor in limiting overall performance. However, it is not easy to extend the concurrency control protocols, widely used in database systems, to blockchain systems. In particular, there are two challenges to achieve parallel execution of transactions in blockchain as follows: (i) the final results of different replicas may diverge since most protocols just promise the effect of transactions equivalent to some serial order but this order may vary for every concurrent execution; and (ii) almost all state trees that are used to manage states of blockchain do not support fast concurrent updates. In the view of above challenges, we propose a parallel execution engine called PEEP, towards permissioned blockchain systems. Specifically, PEEP employs a deterministic concurrency mechanism to obtain a predetermined serial order for parallel execution, and offers parallel update operations on state tree, which can be implemented on any radix tree with

Merkle property. Finally, the extensive experiments show that PEEP outperforms existing serial execution greatly.

2. URIM: Utility-Oriented Role-Centric Incentive Mechanism Design for Blockchain-based Crowdsensing

Zheng Xu, Chaofan Liu, Peng Zhang, Tun Lu, and Ning Gu

Abs. Crowdsensing is a prominent paradigm that collects data by outsourcing to individuals with sensing devices. However, most existing crowdsensing systems are based on centralized architecture which suffers from poor data quality, high service charge, single point of failure, etc. Some studies have explored decentralized architectures and implementations for crowdsensing based on blockchain, while incentive mechanisms for worker participation and miner participation, which serve as a crucial role in blockchain-based crowdsensing systems (BCSs), are ignored. To address this issue, we propose an incentive mechanism design named URIM to maximize participants' utilities, which consists of worker-centric and miner-centric incentive mechanisms for BCSs. For the worker-centric incentive mechanism, we model it as a reverse auction, in which dynamic programming is utilized to select workers, and payments are determined based on the Vickrey-Clarke-Groves scheme. We also prove this incentive mechanism is computationally efficient, individually rational and truthful. For the miner-centric incentive mechanism, we model interactions among the task requester and miners as a Stackelberg game and adopt the backward induction to analyze its equilibrium at which the utilities of the task requester and miners are optimized. Finally, we demonstrate the significant performance of URIM through extensive simulations.

3. PAS: Enable Partial Consensus in the Blockchain

Zihuan XU, Siyuan HAN, and Lei Chen

Abs. Permissioned Blockchain enables distributed collaboration among organizations that may not trust each other. However, existing systems cannot efficiently support the ordering and execution of transactions in different workflows parallelly, which seriously affect system scalability and the performances in terms of throughput and latency.In this paper, we present a partial consensus mechanism named PAS to achieve fault tolerance and parallelism of transaction processing. In PAS, transactions in different

workflows only need to be confirmed by the involved subset of nodes, which significantly enhances the system performance and scalability. Specifically, we introduce a novel data structure, called hierarchical consensus tree (HCT). It is maintained in each node and used to coordinate the consensus process. HCT guarantees that the consistency reached in different sets of nodes is eventually agreed by all nodes without conflicts and rollbacks. Since there are many valid HCTs with different system improvements, we introduce an optimization problem, named OHCT, to obtain an HCT with respect to the optimal enhancement. We prove OHCT is NP-hard and propose a general framework with efficient algorithms to address it. Finally, we implement PAS on PBFT-based Hyperledger fabric and conduct extensive experiments to show the performance and scalability of PAS.

4. Redesigning the Sorting Engine for Persistent Memory
Yifan Hua, Kaixin Huang, Shengan Zheng, and Linpeng Huang
Abs. Emerging persistent memory (PM, also termed as non-volatile memory) technologies can promise large capacity, non-volatility, byte-addressability and DRAM-comparable access latency. Such amazing features have inspired a host of PM-based storage systems and applications that store and access data directly in PM. Sorting is an important function for many systems, but how to optimize sorting for PM-based systems has not been systematically studied yet. In this paper, we conduct extensive experiments for many existing sorting methods, including both conventional sorting algorithms adapted for PM and recently-proposed PM-friendly sorting techniques, on a real PM platform. The results indicate that these sorting methods all have drawbacks for various workloads. Some of the results are even counterintuitive compared to running on a DRAM-simulated platform in their papers. To the best of our knowledge, we are the first to perform a systematic study on the sorting issue for persistent memory. Based on our study, we propose an adaptive sorting engine, namely SmartSort, to optimize the sorting performance for different conditions. The experimental results demonstrate that SmartSort remarkably outperforms existing sorting methods in a variety of cases.

5. ImputeRNN: Imputing Missing Values in Electronic Medical Records
Jiawei Ouyang, Yuhao Zhang, Xiangrui Cai, Ying ZHANG, and Xiaojie Yuan

Abs. Electronic Medical Records (EMRs), which record visits of patients to the hospital, are the main resources for medical data analysis. However, plenty of missing values in EMRs limit the model capability for various researches in healthcare. Recently, many imputation methods have been proposed to address this challenging problem, but they fail to take medical bias into account. Medical bias is a ubiquitous phenomenon that the missingness of medical data is missing not at random because doctors prone to measure features related to the disease of patients. It reflects the physical conditions of patients, which helps impute missing data with accurate and practical values. In this paper, we propose a novel joint recurrent neural network (RNN) model called ImputeRNN, which considers medical bias for EMR imputation. We model the medical bias by an additional RNN based on a mask (missing or not) matrix, whose hidden vectors are incorporated into the model as contexts by a fusion layer. Extensive experiments on two real-world EMR datasets demonstrate that ImputeRNN outperforms state-of-the-art methods on imputation and downstream prediction tasks.

Research Session 6: Data Mining 1
Session Chair: Varun Chandola (University at Buffalo)

1. Consistency- and Inconsistency-aware Multi-view Subspace Clustering
Guangyu Zhang, Xiaowei Chen, Yuren Zhou, Chang-Dong Wang, and Dong Huang
Abs. Multi-view subspace clustering has emerged as a crucial tool to solve the multi-view clustering problem. However, many of the existing methods merely focus on the consistency issue when learning the multi-view representations, failing to capture the latent inconsistency across different views (which can be caused by the view-specificity or diversity). To tackle this issue, we therefore develop a Consistency- and Inconsistency-aware Multi-view Subspace Clustering for robust clustering. In the proposed method, we decompose the multi-view representations into a view-consistent representation and a set of view-inconsistent representations, through which the multi-view consistency as well as multi-view inconsistency can be well explored. Meanwhile, our method aims to suppress the redundancy and determine the importance of different views by introducing a novel view weighting strategy. Then a unified objective

function is constructed, upon which an efficient optimization algorithm based on ADMM is further performed. Additionally, we design a new way to compute the affinity matrix from both consistent and inconsistent perspectives, which makes sure that the learned affinity matrix comprehensively fit the inherent properties of multi-view data. Experimental results on multiple multi-view data sets confirm the superiority of our method.

2. Discovering Collective Converging Groups of Large Scale Moving Objects in Road Networks

Jinping Jia, Ying Hu, Bin Zhao, Genlin Ji, and Richen Liu

Abs. Group pattern mining based on spatio-temporal trajectories have gained significant attentions due to the prevalence of location acquisition devices and tracking technologies. Representative work includes convoy, swarm, travelling companion, gathering, and platoon. However, these works based on Euclidean space cannot handle group pattern discovery in non-planar space, such as urban road networks. In this paper, we propose a new group pattern, named converging, and its mining method in road networks. Unlike the aforementioned group patterns, a converging indicates that a group of moving objects converge from different directions for a certain time period. Motivated by this, we formalize the concept of a converging based on cluster containment relationship. Since the process of discovering convergings over large scale road network constrained trajectories is quite lengthy, we propose a density clustering algorithm based on road networks (DCRN) and a cluster containment join (CCJ) algorithm to improve the performance. Specifically, DCRN adopts the well-known filter-refinement-verification framework for effciently identifying core points, which utilizes the upper bound property for ε-neighbourhood of point set on an edge to dramatically reduce the candidate core points. To process the neighbourhood queries effciently, we further develop a vertex-neighbourhood based index, which precomputes the ε-neighbourhoods of all vertices, to facilitate neighbourhood queries of all points in road networks. In addition, to process the CCJ effciently, we develop a signature tree based on road network partition index to organize the clusters in road networks hierarchically, which enable us to prune enormous unqualified candidates in an effcient way. Finally, extensive experiments with real and synthetic datasets

show that our proposed methods achieve superior performance and good scalability.

3. Efficient Mining of Outlying Sequential Behavior Patterns
Xu Yifan, Lei Duan, Guicai Xie, Min Fu, Longhai Li, and Jyrki Nummenmaa
Abs. Sequential patterns play an important role when observing behavior. For instance,the daily routines and practices of people can be characterized by sequences of activities.These activity sequences, in turn, can be used to find exceptional and changed behavior. Observing students' behavior changes is an effective approach to find indications of mental health problems, and changes in an elderly person's daily activities may indicate a weakening health condition. With the availability of behaviour sequential events, outlierness analysis of behavior sequences has been established as a meaningful research problem. This paper considers the mining of outlying behavior patterns (OBP) from sequential behaviors. After discussing the challenges of OBP mining, we present OBP-Miner, a heuristic method that computes OBPs by incorporating various pruning techniques.Empirical studies on two real-world datasets demonstrate that OBP-Miner is effective and efficient.

4. Clustering Mixed-Type Data With Correlation-Preserving Embedding
Luan V Tran, Liyue Fan, and Cyrus Shahabi
Abs. Mixed-type data that contains both categorical and numerical features is prevalent in many real-world applications. Clustering mixed-type data is challenging, especially because of the complex relationship between categorical and numerical features. Unfortunately, widely adopted encoding methods and existing representation learning algorithms fail to capture these complex relationships. In this paper, we propose a new correlation-preserving embedding framework, COPE, to learn the representation of categorical features in mixed-type data while preserving the correlation between numerical and categorical features. Our extensive experiments with real-world datasets show that COPE generates high-quality representations and outperforms the state-of-the-art clustering algorithms by a wide margin.

5. Beyond Matching: Modeling Two-Sided Multi-Behavioral Sequences For Dynamic Person-Job Fit

Bin Fu, Hongzhi Liu, Yao Zhu, Yang Song, Tao Zhang, and Zhonghai Wu

Abs. Online recruitment aims to match right talents with right jobs (Person-Job Fit, PJF) online by satisfying the preferences of both persons (job seekers) and jobs (recruiters). Recently, some research tried to solve this problem by deep semantic matching of curriculum vitaes and job postings. But those static profiles don't (fully) reflect users' personalized preferences. In addition, most existing preference learning methods are based on users' matching behaviors. However, matching behaviors are sparse due to the nature of PJF and not fine-grained enough to reflect users' dynamic preferences.With going deep into the process of online PJF, we observed abundant auxiliary behaviors generated by both sides before achieving a matching, such as click, invite/apply and chat. To solve the above problems, we propose to collect and utilize these behaviors along the timeline to capture users' dynamic preferences. We design a Dynamic Multi-Key Value Memory Network to capture users' dynamic preferences from their multi-behavioral sequences. Furthermore, a Bilateral Cascade Multi-Task Learning framework is designed to transfer two-sided preferences learned from auxiliary behaviors to the matching task with consideration of their cascade relations. Offline experimental results on two real-world datasets show our method outperforms the state-of-the-art methods.

Research Session 7: Machine Learning 1
Session Chair: Alfredo Cuzzocrea (University of Calabria)

1. Partial Modal Conditioned GANs for Multi-modal Multi-label Learning with Arbitrary Modal-missing

Yi Zhang, Jundong Shen, Zhecheng Zhang, and Chongjun Wang

Abs. Multi-modal multi-label (MMML) learning serves an important framework to learn from objects with multiple representations and annotations. Previous MMML approaches assume that all instances are with complete modalities, which usually does not hold for real-world MMML data. Meanwhile, most existing works focus on data generation using GAN, while few of them explore the downstream tasks, such as multi-modal multi-label learning. The major challenge is how to jointly model complex modal correlation and label correlation in a mutually beneficial way, especially under the arbitrary modal-missing pattern. Aim at addressing the aforementioned research

challenges, we propose a novel framework named Partial Modal Conditioned Generative Adversarial Networks (PMC-GANs) for MMML learning with arbitrary modal-missing. The proposed model contains a modal completion part and a multi-modal multi-label learning part. Firstly, in order to strike a balance between consistency and complementary across different modalities,PMC-GANs incorporates all available modalities during training and generates high-quality missing modality in an efficient way. After that, PMC-GANs exploits label correlation by leveraging shared information from all modalities and specific information of each individual modality. Empirical studies on 3 MMML datasets clearly show the superior performance of PMC-GANs against other state-of-the-art approaches.

2. Cross-domain error minimization for unsupervised domain adaptation
Yuntao Du, Yinghao Chen, Fengli Cui, Xiaowen Zhang, and Chongjun Wang
Abs. Unsupervised domain adaptation aims to transfer knowledge from a labeled source domain to an unlabeled target domain for improving the performance in the target domain. Previous methods focus on learning domain-invariant features to decrease the discrepancy between the feature distributions as well as minimizing the source error and have made remarkable progress. However, a recently proposed theory reveals that such a strategy is not sufficient for a successful domain adaptation. It is suggested that besides a small source error, both the discrepancy between the feature distributions and the discrepancy between the labeling functions should be small across domains. The discrepancy between the labeling functions is essentially the **cross-domain error** which is ignored by existing methods. To overcome this issue, in this paper, a novel method is proposed to integrate all the three objectives into a unified optimization framework. Moreover, the pseudo labels are widely used in domain adaptation methods, but incorrect pseudo labels can lead to error accumulation during learning. To alleviate this problem, the pseudo labels are obtained by utilizing structural information of the target domain besides source classifier and we propose a curriculum learning based strategy to select the target samples with more accurate pseudo-labels during training. Comprehensive experiments are conducted, and the results validate that our approach outperforms state-of-the-art methods.

3. Unsupervised domain adaptation with unified joint distribution alignment

Yuntao Du, Tan Zhiwen, Xiaowen Zhang, Yirong Yao, Hualei Yu, and Chongjun Wang

Abs. Unsupervised domain adaptation aims at transferring knowledge from a labeled source domain to an unlabeled target domain. Recently, domain-adversarial learning has become an increasingly popular method to tackle this task, which bridges the source domain and target domain by adversarially learning domain-invariant representations. Despite the great success in domain-adversarial learning, these methods fail to achieve the invariance of representations at a class level, which may lead to incorrect distribution alignment. To address this problem, in this paper, we propose a method called *domain adaptation with Unified Joint Distribution Alignment* (UJDA) to perform both domain-level and class-level alignments simultaneously in a unified learning process. Instead of adopting the classical domain discriminator, two novel components named joint classifiers, which are provided with both domain information and label information in both domains, are adopted in UJDA. Single joint classifier plays the min-max game with the feature extractor by the joint adversarial loss to align the class-level alignment. Besides, two joint classifiers as a whole also play the min-max game with the feature extractor by the disagreement loss to achieve the domain-level alignment. Comprehensive experiments on two real-world datasets verify that our method outperforms several state-of-the-art domain adaptation methods.

4. Relation-aware Alignment Attention Network for Multi-view Multi-label Learning

Yi Zhang, Jundong Shen, Cheng Yu, and Chongjun Wang

Abs. Multi-View Multi-Label (MVML) learning refers to complex objects represented by multi-view features and associated with multiple labels simultaneously. Modeling flexible view consistency is recently demanded, existing approaches cannot fully exploit the complementary information across multiple views and meanwhile preserve view-specific properties. Additionally, each label has heterogeneous features from multiple views, and probably correlates with other labels via common views. Traditional strategy tends to select features that are distinguishable for all labels. However, globally shared knowledge cannot handle the label heterogeneity. Furthermore, previous studies

model view consistency and label correlations independently, where interactions between views and labels are not fully exploited. In this paper, we propose a novel MVML learning approach named Relation-aware Alignment attentIon Network (RAIN), where three types of relationships are considered. Specifically, 1) view interactions: capture diverse and complementary information for deep correlated subspace learning; 2) label correlations: adopt multi-head attention to learn semantic label embeddings; 3) label-view dependence: dynamically extracts label-specific representation with the guidance of learned label embeddings. Experiments on various MVML datasets demonstrate effectiveness of RAIN compared with state-of-the-arts. We also experiment on one real-world herbs dataset, which shows promising results for clinical decision support.

5. BIRL: Bidirectional-Interaction Reinforcement Learning Framework for Joint Relation and Entity Extraction
Yashen Wang, Huanhuan Zhang
Abs. Joint relation and entity extraction is a crucial technology to construct a knowledge graph. However, most existing methods (i) can not fully capture the beneficial connections between relation extraction and entity extraction tasks, and (ii) can not combat the noisy data in the training dataset. To overcome these problems, this paper proposes a novel Bidirectional-Interaction Reinforcement Learning (BIRL) framework, to extract entities and relations from plain text. Especially, we apply a relation calibration RL policy to (i) measure relation consistency and enhance the bidirectional interaction between entity mentions and relation types; and (ii) guide a dynamic selection strategy to remove noise from training dataset. Moreover, we also introduce a data augmentation module for bridging the gap of data-efficiency and generalization. Empirical studies on two real-world datasets confirm the superiority of the proposed model.

Research Session 8: Information Retrieval and Search
Session Chair: Xiang Zhao (National University of Defense Technology)

1. Quantum-Inspired Keyword Search on Multi-Model Databases
Gongsheng Yuan, Jiaheng Lu, and Peifeng Su
Abs. With the rising applications implemented in different domains, it is inevitable to require databases to adopt corresponding appropriate data models

to store and exchange data derived from various sources. To handle these data models in a single platform, the community of databases introduces a multi-model database. And many vendors are improving their products from supporting a single data model to being multi-model databases. Although this brings benefits, spending lots of enthusiasm to master one of the multi-model query languages for exploring a database is unfriendly to most users. Therefore, we study using keyword searches as an alternative way to explore and query multi-model databases. In this paper, we attempt to utilize quantum physics's probabilistic formalism to bring the problem into vector spaces and represent events (e.g., words) as subspaces. Then we employ a density matrix to encapsulate all the information over these subspaces and use density matrices to measure the divergence between query and candidate answers for finding top-k the most relevant results. In this process, we propose using pattern mining to identify compounds for improving accuracy and using dimensionality reduction for reducing complexity. Finally, empirical experiments demonstrate the performance superiority of our approaches over the state-of-the-art approaches.

2. ZH-NER: Chinese Named Entity Recognition with Adversarial Multi-Task Learning and Self-Attentions

Peng Zhu, Dawei Cheng, Fangzhou Yang, Yifeng Luo, Weining Qian, and Aoying Zhou

Abs. NER is challenging because of the semantic ambiguities in academic literature, especially for non-Latin languages. Besides, recognizing Chinese named entities needs to consider word boundary information, as words contained in Chinese texts are not separated with spaces. Leveraging word boundary information could help to determine entity boundaries and thus improve entity recognition performance. In this paper, we propose to combine word boundary information and semantic information for named entity recognition based on multi-task adversarial learning. We learn common shared boundary information of entities from multiple kinds of tasks, including Chinese word segmentation (CWS), part-of-speech (POS) tagging and entity recognition, with adversarial learning. We learn task-specific semantic information of words from these tasks, and combine the learned boundary information with the semantic information to improve entity recognition, with multi-task learning. We

conduct extensive experiments to demonstrate that our model achieves considerable performance improvements.

3. Drug-Drug Interaction Extraction via Attentive Capsule Network with an Improved Sliding-margin Loss

Dongsheng Wang, Hongjie Fan, and Junfei Liu

Abs. Relation extraction (RE) is an important task in information extraction. Drug-drug interaction (DDI) extraction is a subtask of RE in the biomedical field. Existing DDI extraction methods are usually based on recurrent neural network (RNN) or convolution neural network (CNN) which have finite feature extraction capability. Therefore, we propose a new approach for addressing the task of DDI extraction with consideration of sequence features and dependency characteristics. A sequence feature extractor is used to collect features between words, and a dependency feature extractor is designed to mine knowledge from the dependency graph of sentence. Moreover, we use an attention-based capsule network for DDI relation classification, and an improved sliding-margin loss is proposed to well learn relations. Experiments demonstrate that incorporating capsule network and improved sliding-margin loss can effectively improve the performance of DDI extraction.

4. Span-based Nested Named Entity Recognition with Pretrained Language Model

Chenxu Liu, Hongjie Fan, and Junfei Liu

Abs. Named Entity Recognition (NER) is generally regarded as a sequence labeling task, which faces a serious problem when the named entities are nested. In this paper, we propose a span-based model for nested NER, which enumerates all possible spans as potential entity mentions in a sentence and classifies them with pretrained BERT model. In view of the phenomenon that there are too many negative samples in all spans, we propose a multi-task learning method, which divides NER task into entity identification and entity classification task. In addition, we propose the entity IoU loss function to focus our model on the hard negative samples. We evaluate our model on three standard nested NER datasets: GENIA, ACE2004 and ACE2005, and the results show that our model outperforms other state-of-the-art models with the same pretrained language model, achieving 79.46, 87.30 and 85.24 respectively in terms of F1 score.

5. Poetic Expression Through Scenery: Automatic Chinese Classical Poetry Generation from Images

Haotian Li, Jiatao Zhu, Sichen Cao, Xiangyu Li, Jiajun Zeng, and Peng Wang

Abs. Automatic poetry generation is a hot topic in natural language processing. In recent years, some methods have been proposed for generating Chinese classical poetry. However, most of them only accept texts or user-specified words as input, which contradicts with the fact that ancient Chinese wrote poems inspired by visions, hearings and feelings. This paper proposes a method to generate sentimental Chinese classical poetry automatically from images based on convolutional neural networks and the language model. First, our method extracts visual information from the image and maps it to initial keywords by two parallel image classification models, then filters and extends these keywords to form a keywords set which is finally input into the poetry generation model to generate poems of different genres. A bi-directional generation algorithm and two fluency checkers are proposed to ensure the diversity and quality of generated poems, respectively. Besides, we constrain the range of optional keywords and define three sentiment-related keywords dictionary to avoid modern words that lead to incoherent content as well as ensure the emotional consistency with given images. Both human and automatic evaluation results demonstrate that our method can reach a better performance on quality and diversity of generated poems.

Research Session 9: Big Data 1

Session Chair: Yang Cao (Kyoto University)

1. Learning the Implicit Semantic Representation on Graph-Structured Data

Likang Wu, Zhi Li, Hongke Zhao, Qi Liu, Jun Wang, Mengdi Zhang, and Enhong Chen

Abs. Existing representation learning methods in graph convolutional networks are mainly designed by describing the neighborhood of each node as a perceptual whole, while the implicit semantic associations behind highly complex interactions of graphs are largely unexploited. In this paper, we propose a Semantic Graph Convolutional Networks (SGCN) that explores the implicit semantics by learning latent semantic-paths in graphs. In previous work, there

are explorations of graph semantics via meta-paths. However, these methods mainly rely on explicit heterogeneous information that is hard to be obtained in a large amount of graph-structured data. SGCN first breaks through this restriction via leveraging the semantic-paths dynamically and automatically during the node aggregating process. To evaluate our idea, we conduct mainstream experiments on several standard datasets, the empirical results show the superior performance of the proposed model. We will open our source code if the paper is lucky enough to be accepted.

2. Multi-job merging framework and scheduling optimization for Apache Flink
Hangxu Ji, Gang WU, Yuhai Zhao, Ye Yuan, and Guoren Wang
Abs. With the popularization of big data technology, distributed computing systems are constantly evolving and maturing, making substantial contributions to the query and analysis of massive data. However, the insufficient utilization of system resources is an inherent problem of distributed computing engines. Particularly, when more jobs lead to execution blocking, the system schedules multiple jobs on a first-come-first-executed (FCFE) basis, even if there are still many remaining resources in the cluster. Therefore, the optimization of resource utilization is key to improving the efficiency of multi-job execution. We investigated the field of multi-job execution optimization, designed a multi-job merging framework and scheduling optimization algorithm, and implemented them in the latest generation of a distributed computing system, Apache Flink. In summary, the advantages of our work are highlighted as follows: (1) the framework enables Flink to support multi-job collection, merging and dynamic tuning of the execution sequence, and the selection of these functions are customizable. (2) with the multi-job merging and optimization, the total running time can be reduced by 31% compared with traditional sequential execution. (3) the multi-job scheduling optimization algorithm can bring 28% performance improvement, and in the average case can reduce the cluster idle resources by 61%.

3. CIC-FL: Enabling Class Imbalance-aware Clustered Federated Learning over Shifted Distributions
Xuefeng Liu, ShaoJie Tang, JianWei Niu, ZhangMin Huang, and YaNan Fu

Abs. Federated learning (FL) is a distributed training framework where decentralized clients collaboratively train a model. One challenge in FL is concept shift, i.e. that the conditional distributions of data in different clients are disagreeing. A natural solution is to group clients with similar conditional distributions into the same cluster. However, methods following this approach leverage features extracted in federated settings (e.g. model weights or gradients) which intrinsically reflect the joint distributions of clients. Considering the difference between conditional and joint distributions, they would fail in the presence of class imbalance (i.e. that the marginal distributions of different classes vary in a client's data). Although adopting sampling techniques or cost-sensitive algorithms can alleviate class imbalance, they either skew the original conditional distributions or lead to privacy leakage.To address this challenge, we propose CIC-FL, a class imbalance-aware clustered federated learning method. CIC-FL iteratively bipartitions clients by leveraging a particular feature sensitive to concept shift but robust to class imbalance. In addition, CIC-FL is privacy-preserving and communication efficient. We test CIC-FL on benchmark datasets including Fashion-MNIST, CIFAR-10 and IMDB. The results show that CIC-FL outperforms state-of-the-art clustering methods in FL in the presence of class imbalance.

4. vRaft: Accelerating the Distributed Consensus under Virtualized Environments

Yangyang Wang, and YunPeng Chai

Abs. In recent years, Raft has been gradually widely used in many distributed systems (e.g., Etcd, TiKV, PolarFS, etc.) to ensure the distributed consensus because it is effective and easy to implement. However, because the performance of the virtual node in cloud environments usually heterogeneous and fluctuant due to the “noisy neighbor” problem and the cost efficiency, the strong leader mechanism makes the Raft protocol encounter a serious performance challenge. Specifically, when the performance of the leader node declines temporarily, the whole system performance will descend accordingly since both the write and the read requests serving will be blocked by the slow leader processing. Aiming to solve this problem, we proposed a modified version of Raft specially optimized for virtualized environments, i.e., vRaft. It breaks Raft’s strong leader mechanism and can fully utilize the temporarily fast

followers to accelerating both the write and the read requests processing in a virtualized cloud environment, without affecting the linearizability guarantee of Raft. The experiments based on the virtual nodes in Tencent Cloud indicate that vRaft improves the throughput by up to 64.2%, reduces average latency by 38.1%, and shortens the tail latency by 88.5% in a typical read/write-balanced workload compared with Raft.

5. Secure and Efficient Certificateless Provable Data Possession for Cloud-Based Data Management Systems
Jing Zhang, Jie Cui, Hong Zhong, Chengjie Gu, and Lu Liu
Abs. Cloud computing provides important data storage, processing and management functions for data owners who share their data with data users through cloud servers. Although cloud computing brings significant advantages to data owners, the data stored in the cloud also faces many internal/external security attacks. Existing certificateless data provider schemes have the following two common shortcomings, i.e., most of which use plaintext to store data and use the complex bilinear pairing operation. To address such shortcomings, this scheme proposes secure and efficient certificateless provable data possession for cloud-based data management systems. In our solution, the data owners and cloud servers need to register with the key generation center only once. To ensure the integrity of encrypted data, we use the public key of the cloud server to participate in signature calculation. Moreover, the third-party verifier can audit the integrity of ciphertext without downloading the whole encrypted data. Security analysis shows that our proposed scheme is provably secure under the random oracle model. An evaluation of performance shows that our proposed scheme is efficient in terms of computation and communication overheads.

Research Session 10: Social Network
Session Chair: Long Yuan (Nanjing University of Science and Technology)

1. SCHC: Incorporating Social Contagion and Hashtag Consistency for Topic-oriented Social Summarization
Ruifang He, huanyu Liu, and Liangliang Zhao

Abs. The boom of social media platforms like Twitter brings the large scale short, noisy and redundant messages, making it difficult for people to obtain essential information. We study extractive topic-oriented social summarization to help people grasp the core information on social media quickly. Previous methods mainly extract salient content based on textual information and shallow social signals. They ignore that user generated messages propagate along the social network and affect users on their dissemination path, leading to user-level redundancy. Besides, hashtags on social media are a special kind of social signals, which can be regarded as keywords of a post and contain abundant semantics. In this paper, we propose to leverage social theories and social signals (i.e. multi-order social relations and hashtags) to address the redundancy problem and extract diverse summaries. Specifically, we propose a novel unsupervised social summarization framework which considers Social Contagion and Hashtag Consistency (SCHC) theories. To model relations among tweets, two relation graphs are constructed based on user-level and hashtag-level interaction among tweets. These social relations are further integrated into a sparse reconstruction framework to alleviate the user-level and hashtag-level redundancy respectively. Experimental results on the CTS dataset prove that our approach is effective.

2. Image-Enhanced Multi-Modal Representation for Local Topic Detection from Social Media

Junsha Chen, Neng Gao, Yifei Zhang, and Chenyang Tu

Abs. Detecting local topics from social media is an important task for many applications, ranging from event tracking to emergency warning. Recent years have witnessed growing interest in leveraging multi-modal social media information for local topic detection. However, existing methods suffer great limitation in capturing comprehensive semantics from social media and fall short in bridging semantic gaps among multi-modal contents, i.e., some of them overlook visual information which contains rich semantics, others neglect indirect semantic correlation among multi-modal information. To deal with above problems, we propose an effective local topic detection method with two major modules, called IEMM-LTD. The first module is an image-enhanced multi-modal embedding learner to generate embeddings for words and images, which can capture comprehensive semantics and preserve both direct and

indirect semantic correlations. The second module is an embedding based topic model to detect local topics represented by both words and images, which adopts different prior distributions to model multi-modal information separately. We evaluate IEMM-LTD on two real-world tweet datasets, the experimental results show that IEMM-LTD has achieved the best performance compared to the existing state-of-the-art methods.

3. Personality Traits Prediction Based on Sparse Digital Footprints via Discriminative Matrix Factorization

Shipeng Wang, Daokun Zhang, Lizhen Cui, Xudong Lu, Lei Liu, and Qingzhong Li

Abs. Identifying individuals' personality traits from their digital footprints has been proved able to improve the service of online platforms. However, due to the privacy concerns and legal restrictions, only some sparse, incomplete and anonymous digital footprints can be accessed, which seriously challenges the existing personality traits identification methods. To make the best of the available sparse digital footprints, we propose a novel personality traits prediction algorithm through jointly learning discriminative latent features for individuals and a personality traits predictor performed on the learned features. By formulating a discriminative matrix factorization problem, we seamlessly integrate the discriminative individual feature learning and personality traits predictor learning together. To solve the discriminative matrix factorization problem, we develop an alternative optimization based solution, which is efficient and easy to be parallelized for large-scale data. Experiments are conducted on the real-world Facebook like digital footprints. The results show that the proposed algorithm outperforms the state-of-the-art personality traits prediction methods significantly.

4. A Reinforcement Learning Model for Influence Maximization in Social Networks

Chao Wang, Yiming Liu, Xiaofeng Gao, and Guihai Chen

Abs. Social influence maximization problem has been widely studied by the industrial and theoretical researchers over the years. However, with the skyrocketing scale of networks and growing complexity of application scenarios, traditional approximation approaches suffer from weak approximation

guarantees and bad empirical performances. What's more, they can't be applied to new users in dynamic network. To tackle those problems, we introduce a social influence maximization algorithm via graph embedding and reinforcement learning. Nodes are represented in the graph with their embedding, and then we formulate a reinforcement learning model where both the states and the actions can be represented with vectors in low dimensional space. Now we can deal with graphs under various scenarios and sizes, just by learning parameters for the deep neural network. Hence, our model can be applied to both large-scale and dynamic social networks. Extensive real-world experiments show that our model significantly outperforms baselines across various data sets, and the algorithm learned on small-scale graphs can be generalized to large-scale ones.

5. A multilevel inference mechanism for user attributes over social networks
Hang Zhang, Yajun Yang, Xin Wang, Hong Gao, Qinghua Hu, and Dan Yin
Abs. In a real social network, each user has attributes for self-description called user attributes which are semantically hierarchical. With these attributes, we can implement personalized services such as user classification and targeted recommendations. Most traditional approaches mainly focus on the flat inference problem without considering the semantic hierarchy of user attributes which will cause serious inconsistency in multilevel tasks. To address these issues, in this paper, we propose a cross-level model called IWM. It is based on the theory of maximum entropy which collects attribute information by mining the global graph structure. Meanwhile, we propose a correction method based on the predefined hierarchy to realize the mutual correction between different layers of attributes. Finally, we conduct extensive verification experiments on the DBLP data set and it has been proved that compared with other algorithms, our method has a superior effect.

Research Session 11: Graph Data 2
Session Chair: Xiaofei Zhang (University of Memphis)

1. Spatial-Temporal Attention Network for Temporal Knowledge Graph Completion
Jiasheng Zhang, Shuang Liang, Zhiyi Deng, and Jie Shao

Abs. Temporal knowledge graph completion, which aims to predict missing links in temporal knowledge graph (TKG), is an important research task due to the incompleteness of TKG. Recently, TKG embedding methods have proved to be effective for this task. However, most of existing methods regard TKG as a set of independent facts and consequently ignore the implicit relevance among facts. Actually, as a kind of dynamic heterogeneous graph, the evolving graph structure of TKG is able to reflect a wealth of information. To this end, in this paper we regard temporal knowledge graph as heterogeneous and discrete spatial-temporal resource, and propose a novel spatial-temporal attention network to learn TKG embeddings by modeling spatial-temporal property of TKG while considering its special characteristics. Specifically, our model employs a Multi-Faceted Graph Attention Network (MFGAT) to extract rich structural information from the egocentric network of each entity. Additionally, an Adaptive Temporal Attention Mechanism (ADTAT) is utilized to flexibly model the correlation of entity representations in the time dimension. Finally, by combing our obtained representations with existing static KG completion methods, they can be extended to spatial-temporal versions to predict missing links in TKG while considering its inherent graph structure and time-evolving property. Experimental results on three real-world datasets demonstrate the superiority of our model over the state-of-the-art methods.

2. Ranking Associative Entities in Knowledge Graph by Graphical Modeling of Frequent Patterns

Jie Li, Kun Yue, Liang Duan, and Jianyu Li

Abs. Ranking associative entities in Knowledge Graph (KG) is critical for entity-oriented tasks like entity recommendation and associative inference. Existing methods benefit from explicit linkages in KG w.r.t. exactly two query entities via the closely appearing co-occurrences. Given a query including one or more entities in KG, it is necessary to obtain the implicit associative entities and uncover the strength of associations from data. To this end, we leverage KG with Web resources and propose an approach to ranking associative entities based on frequent pattern mining and graph embedding. First, we construct an entity dependency graph from the frequent patterns of entities generated from both KG and Web resources. Thus, the existence and strength of associations between entities could be depicted effectively in a holistic way. Second, we embed the

dependency graph into a lower-dimensional space and consequently fulfill entity ranking on the embedding. Finally, we conduct an extensive experimental study on real-life datasets, and verify the effectiveness of our proposed approach compared to competitive baselines.

3. A Novel Embedding Model for Knowledge Graph Completion Based on Multi-Task Learning

Jiaheng Dou, Bing Tian, Yong Zhang, and Chunxiao Xing

Abs. Knowledge graph completion is the task of predicting missing relationships between entities in Knowledge Graphs. State-of-the-art knowledge graph completion methods are known to be primarily knowledge embedding based models, which are broadly classified as translational models and neural network models. However, both kinds of models are single-task based models and hence fail to capture the underlying inter-semantic and structural relationships that are inherently present in different knowledge graphs. To this end, in this paper we combine the translational and neural network methods and propose a novel multi-task learning embedding framework (TransMTL) that can jointly learn multiple knowledge graph embeddings simultaneously. Specifically, in order to transfer structural knowledge between different KGs, we devise a global relational graph attention network which is shared by all knowledge graphs to obtain the global representation of each triple element.Such global representations are then integrated into task-specific translational embedding models of each knowledge graph to preserve its transition property.We conduct an extensive empirical evaluation of multi-version TransMTL based on different translational models on two benchmark datasets WN18RR and FB15k-237.Experiments show that TransMTL outperforms the corresponding single-task based models by an obvious margin and obtains the comparable performance to state-of-the-art embedding models.

4. Gaussian Metric Learning for Few-Shot Uncertain Knowledge Graph Completion

Jiatao Zhang, Tianxing Wu, and Guilin Qi

Abs. Recent advances in relational information extraction have allowed to automatically construct large-scale knowledge graphs (KGs). Nevertheless, an automatic process entails that a significant amount of uncertain facts are

introduced into KGs. Uncertain knowledge graphs (UKGs) such as NELL and Probase model this kind of uncertainty as confidence scores associated to facts for providing more precise knowledge descriptions. Existing UKG completion methods require sufficient training examples for each relation. However, most relations only have few facts in real-world UKGs. To solve the above problem, in this paper, we propose a novel method to complete few-shot UKGs based on Gaussian metric learning (GMUC) which could complete missing facts and confidence scores with few examples available. By employing a Gaussian-based encoder and metric function, GMUC could effectively capture uncertain semantic information. Extensive experiments conducted over various datasets with different uncertainty levels demonstrate that our method consistently outperforms baselines.

5. Towards Entity Alignment in the Open World: An Unsupervised Approach
Weixin Zeng, Xiang Zhao, Jiuyang Tang, Xinyi Li, Minnan Luo, and Qinghua Zheng

Abs. Entity alignment (EA) aims to discover the equivalent entities in different knowledge graphs (KGs). It is a pivotal step for integrating KGs to increase knowledge coverage and quality. Recent years have witnessed a rapid increase of EA frameworks. However, state-of-the-art solutions tend to rely on labeled data for model training. Additionally, they work under the closed-domain setting and cannot deal with entities that are unmatchable. To address these deficiencies, we offer an unsupervised framework that performs entity alignment in the open world. Specifically, we first mine useful features from the side information of KGs. Then, we devise an unmatchable entity prediction module to filter out unmatchable entities and produce preliminary alignment results. These preliminary results are regarded as the pseudo-labeled data and forwarded to the progressive learning framework to generate structural representations, which are integrated with the side information to provide a more comprehensive view for alignment. Finally, the progressive learning framework gradually improves the quality of structural embeddings and enhances the alignment performance by enriching the pseudo-labeled data with alignment results from the previous round. Our solution does not require labeled data and can effectively filter out unmatchable entities. Comprehensive experimental evaluations validate its superiority.

Research Session 12: Spatial/Temporal Data 2
Session Chair: Shiyu Yang (Guangzhou University)

1. Exploiting Multi-Source Data for Adversarial Driving Style Representation Learning
Zhidan Liu, Junhong Zheng, Zengyang Gong, Haodi Zhang, and Kaishun Wu
Abs. Characterizing human driver's driving behaviors from GPS trajectories is an important yet challenging trajectory mining task. Previous works heavily rely on high-quality GPS data to learn such driving style representations through deep neural networks. However, they have overlooked the driving contexts that greatly govern drivers' driving activities and the data sparsity issue of practical GPS trajectories collected at a low-sampling rate. To address the limitations of existing works, we present an adversarial driving style representation learning approach, named Radar. In addition to summarizing statistic features from raw GPS data, Radar also extracts contextual features from three aspects of road condition, geographic semantic, and traffic condition. We further exploit the advanced semi-supervised generative adversarial networks to construct our learning model. By jointly considering statistic features and contextual features, the trained model is able to efficiently learn driving style representations even from sparse trajectories. Experiments on two benchmark applications, i.e., driver number estimation and driver identification, with a large real-world GPS trajectory dataset demonstrate that Radar can outperform the state-of-the-art approaches by learning more effective and accurate driving style representations.

2. MM-CPred: A Multi-task Predictive Model for Continuous-time Event Sequences with Mixture Learning Losses
Li Lin, Zan Zong, Lijie Wen, Chen Qian, Shuang Li, and Jianmin Wang
Abs. Sequence prediction is a well-defined problem with a proliferation of applications, such as recommendation systems, social media monitor, economic analysis, etc. Recently, RNN-based methodologies have shown their superiority in time-series data analysis and sequence modeling. The question of which event would happen next is not difficult to answer anymore, but the prediction of when it would happen is still a mountain to climb. In this paper, we propose a Multi-task model to predict both events and their continuous timestamps at the

same time. Specifically, 1) we design a two-layer RNN encoder for event sequences and a CNN encoder for time sequences, both equipped with multi-head self-attention to align history features; 2) we form multiple generative adversarial models for predicting future time sequences to solve the problem of multi-modal time distribution; 3) Mixture learning losses are adopted to conduct a 3-step learning strategy for training our model, the cross-entropy loss for events, Huber loss and adversarial classification loss which induces the Wasserstein distance for times. Due to these characteristics, we name it MM-CPred. The experiments on 4 real-life datasets confirmed its improvements compared with the baselines.

3. Modeling Dynamic Social Behaviors with Time-Evolving Graphs for User Behavior Predictions

Tianzi Zang, Yanmin Zhu, Chen Gong, Haobing Liu, and Bo Li

Abs. The full coverage of Wi-fi signals and the popularization of intelligent card systems provide a large number of data that contain human mobility patterns. Effectively utilizing such data to make user behavior predictions finds useful applications such as predictive behavior analysis, personalized recommendation, and location-aware services. Existing methods for user behavior predictions merely capture temporal dependencies within individual historical records. We argue that user behaviors are largely affected by friends in their social circles and such influences are dynamic due to users' dynamic social behaviors. In this paper, we propose a model named SDSIM which consists of three independent and complementary modules to jointly model the influences of user dynamic social behaviors, user demographics similarities, and individual-level behavior patterns. We construct time-evolving graphs in days to indicate user dynamic social behaviors and design a novel component named DSBcell which captures not only the social influences but also the regularity and periodicity of user social behaviors. We also construct a graph based on user similarities in demographics and generate a representation for each user. Experiments on two real-world datasets for multiple user behavior-related prediction tasks prove the effectiveness of our proposed model compared with state-of-the-art methods.

4. Memory-Efficient Storing of Timestamps for Spatio-Temporal Data Management in Columnar In-Memory Databases

Keven Richly

Abs. Vast amounts of spatio-temporal data are continuously accumulated through the wide distribution of location-acquisition technologies. Concerning the increased performance requirements of spatio-temporal data mining applications, in-memory database systems are used to store and process the data. As DRAM capacities are limited and expensive, the efficient utilization of the available resources is necessary. In contrast to storing the positions of moving objects, there is less focus on optimized storage concepts for the temporal component. However, it has a significant impact on the memory footprint and the overall system performance. Especially for columnar databases, the memory-efficient storing of timestamps is challenging as numerous compression approaches are optimized for contradicting data characteristics (e.g., low number of distinct values, sequences of equal values). In this paper, we present and compare different data layouts for columnar in-memory databases to store timestamps. Additionally, we propose an optimized approach for range queries with standard access ranges that uses multiple columns. We evaluate the memory consumption and performance of different compression techniques for specific access patterns. Based on the results, we introduce a workload-aware heuristic approach for the selection of performance and cost balancing data layouts. Further, we demonstrate that workload-driven optimizations for timestamps can significantly reduce the data footprint and increase the performance of spatio-temporal data management.

5. Personalized POI Recommendation: Spatio-Temporal Representation Learning with Social Ties

Shaojie Dai, Yanwei Yu, Hao Fan, and Junyu Dong

Abs. Recommending a limited number of Point-of-Interests (POIs) a user will visit next has become increasingly important to both users and POI holders for Location-Based Social Networks (LBSNs). However, POI recommendation is a challenging task since complex sequential patterns and rich contexts are contained in extremely sparse user check-in data. Recently studies show that embedding techniques effectively incorporate POI contextual information to alleviate the data sparsity issue, and Recurrent Neural Network (RNN) has been successfully employed for sequential prediction. Nevertheless, existing POI recommendation approaches are still limited in capturing user personalized

preference due to separate embedding learning or network modeling. To this end, we propose a novel unified spatio-temporal neural network framework, named PPR, which leverages users' check-in records and social ties to recommend personalized POIs for querying users by joint embedding learning and sequential modeling. Specifically, PPR first learns user and POI representations by joint modeling User-POI relation, sequential patterns, geographical influence, and social ties in a heterogeneous graph, and then models user personalized sequential patterns using the designed spatio-temporal network based on LSTM model for the personalized POI recommendation. Finally, extensive experiments on three real-world datasets demonstrate that our model significantly outperforms state-of-the-art baselines for successive POI recommendation in terms of Accuracy, Precision, Recall and NDCG.

Research Session 13: Text and Unstructured Data 2
Session Chair: Xuequn Shang (Northwestern Polytechnical University)

1. Latent Graph Recurrent Network for Document Ranking
Qian Dong, and Shuzi Niu
Abs. BERT based ranking models are emerging for its superior natural language understanding ability. The attention matrix learned through BERT captures all the word relations in the input text. However, neural ranking models focus only on the text matching between query and document. To solve this problem, we propose a recurrent graph neural network based model to refine word representations from BERT for document ranking, referred to as Latent Graph Recurrent Network (LGRe for short). For each query and document pair, word representations are learned through transformer layer. Based on these word representations, we propose masking strategies to construct a bipartite-core word graph to model the matching between the query and document. Word representations will be further refined by recurrent graph neural network to enhance word relations in this graph. The final relevance score is computed from refined word representations through fully connected layers. Moreover, we propose a triangle distance loss function for embedding layers as an auxiliary task to obtain discriminative representations. It is optimized jointly with pairwise ranking loss for ad hoc document ranking task. Experimental results on

public benchmark TREC Robust04 and WebTrack2009-12 test collections show that LGRe outperforms state-of-the-art baselines more than 2%.

2. Discriminative Feature Adaptation via Conditional Mean Discrepancy for Cross-domain Text Classification

Bo Zhang, Xiaoming Zhang, Yun Liu, and Lei Cheng

Abs. This paper concerns the problem of Unsupervised Domain Adaptation (UDA) in text classification, aiming to transfer the knowledge from a source domain to a different but related target domain. Previous methods learn the discriminative feature of target domain in terms of noisy pseudo labels, which inevitably produces negative effects on training a robust model. In this paper, we propose a novel criterion Conditional Mean Discrepancy (CMD) to learn the discriminative features by matching the conditional distributions across domains. CMD embeds both the conditional distributions of source and target domains into tensor-product Hilbert space and computes Hilbert-Schmidt norm instead. We shed a new light on discriminative feature adaptation: the collective knowledge of discriminative features of different domains is naturally discovered by minimizing CMD. We propose Aligned Adaptation Networks (AAN) to learn the domain-invariant and discriminative features simultaneously based on Maximum Mean Discrepancy (MMD) and CMD. Meanwhile, to trade off between the marginal and conditional distributions, we further maximize both MMD and CMD criterions using adversarial strategy to make the features of AAN more discrepancy-invariant. To the best of our knowledge, this is the first work to definitely evaluate the shifts in the conditional distributions across domains. Experiments on cross-domain text classification demonstrate that AAN achieves better classification accuracy but less convergence time compared to the state-of-the-art deep methods.

3. Discovering Protagonist of Sentiment with Aspect Reconstructed Capsule Network

Chi Xu, Hao Feng, Guoxin Yu, Min Yang, Xiting Wang, Yan Song, and Xiang Ao

Abs. Most existing aspect-term level sentiment analysis (ATSA) approaches combined neural networks with attention mechanisms built upon given aspect to generate refined sentence representation for better predictions. In these methods, aspect terms are always provided in both training and testing process which may

degrade aspect-level analysis into sentence-level prediction. However, the annotated aspect term might be unavailable in real-world scenarios which may challenge the applicability of the existing methods. In this paper, we aim to improve ATSA by discovering the potential aspect terms of the predicted sentiment polarity when the aspect terms of a test sentence are unknown. We access this goal by proposing a capsule network based model named CAPSAR. In CAPSAR, sentiment categories are denoted by capsules and aspect term information is injected into sentiment capsules through a sentiment-aspect reconstruction procedure during the training. As a result, coherent patterns between aspects and sentimental expressions are encapsulated by these sentiment capsules. Experiments on three widely used benchmarks demonstrate these patterns have potential in exploring aspect terms from test sentence when only feeding the sentence to the model. Meanwhile, the proposed CAPSAR can clearly outperform SOTA methods in standard ATSA tasks.

4. Discriminant Mutual Information for Text Feature Selection

Jiaqi Wang, and Li Zhang

Abs. In text classification tasks, the high dimensionality of data would result in a high computational complexity and decrease the classification accuracy because of high correlation between features; so, it is necessary to execute feature selection.In this paper, we propose a Discriminant Mutual Information (DMI) criterion to select features for text classification tasks.DMI measures the discriminant ability of features from two aspects. One is the mutual information between features and the label information. The other is the discriminant correlation degree between a feature and a target feature subset based on the label information, which could be used for judging whether a feature is redundant in the target feature subset. Thus, DMI is a de-redundancy text feature selection method considering discriminant information.In order to prove the superiority of DMI, we compare it with the state-of-the-art filter methods for text feature selection and conduct experiments on two datasets: Reuters-21578 and WebKB.K-Nearest Neighbor (KNN) and Support Vector Machine (SVM) are taken as the subsequent classifiers. Experimental results shows that the proposed DMI has significantly improved the classification accuracy and F1-score of both Reuters-21578 and WebKB.

5. CAT-BERT: A Context-Aware Transferable BERT Model for Multi-Turn Machine Reading Comprehension

Cen Chen, Xinjing Huang, Feng Ji, Chengyu Wang, Minghui Qiu, Jun Huang, and Yin Zhang

Abs. Machine Reading Comprehension (MRC) is an important NLP task with the goal of extracting answers to user questions from background passages. For conversational applications, modeling the contexts under the multi-turn setting is highly necessary for MRC, which has drawn great attention recently. Past studies on multi-turn MRC usually focus on a single domain, ignoring the fact that knowledge in different MRC tasks are transferable. To address this issue, we present a unified framework to model both single-turn and multi-turn MRC tasks which allows knowledge sharing from different source MRC tasks to help solve the target MRC task. Specifically, the Context-Aware Transferable Bidirectional Enoceder Representations from Transformers (CAT-BERT) model is proposed, which jointly learns to solve both single-turn and multi-turn MRC tasks in a single pre-trained language model. In this model, both history questions and answers are encoded into the contexts for the multi-turn setting. To capture the task-level importance of different layer outputs, a task-specific attention layer is further added to the CAT-BERT outputs, reflecting the positions that the model should pay attention to for a specific MRC task. Extensive experimental results and ablation studies show that CAT-BERT achieves competitive results in multi-turn MRC tasks, outperforming strong baselines.

Research Session 14: Machine Learning 2
Session Chair: Chuan-Ju Wang (Academia Sinica, Taiwan)

1. DFILAN: Domain-based Feature Interactions Learning via Attention Networks for CTR Prediction

Yongliang Han, Yingyuan Xiao, Hongya Wang, Wenguang Zheng, and Ke Zhu

Abs. Click-Through Rate (CTR) prediction has become an important part of many enterprise applications, such as recommendation systems and online advertising. In recent years, some models based on deep learning have been applied to the CTR prediction systems. Although the accuracy is improving, the complexity of the model is constantly increasing. In this paper, we propose a novel model called Domain-based Feature Interactions Learning via Attention

Networks (DFILAN), which can effectively reduce model complexity and automatically learn the importance of feature interactions. On the one hand, the DFILAN divides the input features into several domains to reduce the time complexity of the model in the interaction process. On the other hand, the DFILAN interacts at the embedding vector dimension level to improve the feature interactions effect and leverages the attention network to automatically learn the importance of feature interactions. Extensive experiments conducted on the two public datasets show that DFILAN is effective and outperforms the state-of-the-art models.

2. Double Ensemble Soft Transfer Network for Unsupervised Domain adaptation

Manliang Cao, Xiangdong Zhou, Lan Lin, and Bo Yao

Abs. Domain adaptation aims to transfer the enriched label knowledge from large amounts of source data to unlabeled target data. Recent methods start to solve the class-wise domain adaptation problem by incorporating the soft labels to each target data. Although the soft label strategy could alleviate the negative influence caused by the hard label strategy to some extent, the improper propagation sequence ignoring the labeling difficulties of different target examples will lead to confusing probabilities problem. Moreover, the instability of a single propagation model in dealing with various data may also hinder the performance of target label inference. To address these limitations, we propose a Double Ensemble Soft Transfer Network (DESTN) to jointly optimize the class-wise adaptation and learn the discriminative domain-invariant features with clear soft target labels. Our motivation is to construct a Label Propagation Ensemble (LPE) model by various feature subspaces so as to get robust and clear soft target labels for class-wise domain adaptation. Meanwhile, the other Classifiers Ensemble Framework (CEF) is trained on the labeled source samples and the reliable pseudo-labeled target samples for learning the discriminative features during the iteration. Extensive experiments show that DESTN significantly outperforms several state-of-the-art methods.

3. Attention-based Multimodal Entity Linking with High-Quality Images

Li Zhang, Zhixu Li, and Qiang Yang

Abs. Multimodal entity linking (MEL) is an emerging research field which uses both textual and visual information to map an ambiguous mention to an entity in a knowledge base (KB). However, images do not always help, which may also backfire if they are irrelevant to the textual content at all. Besides, the existing efforts mainly focus on learning a representation of both mentions and entities from their textual and visual contexts, without considering the negative impact brought by noisy irrelevant images, which happens frequently with social media posts. In this paper, we propose a novel MEL model, which not only removes the negative impact of noisy images, but also uses multiple attention mechanism to better capture the connection between mention representation and its corresponding entity representation. Our empirical study on a large real data collection demonstrates the effectiveness of our approach.

4. Learning to Label with Active Learning and Reinforcement Learning
Xiu Tang, Sai Wu, Gang Chen, Ke Chen, and Lidan Shou
Abs. Training data labelling is financially expensive in domain-specific learning applications, which heavily relies on the intelligence from domain experts. Thus, with budget constraint, it is important to judiciously select high-quality training data for labelling in order to prevent over-fitting. In this paper, we propose a learning-to-label (L2L) framework leveraging active learning and reinforcement learning to iteratively select data to label for Name Entity Recognition (NER) task. Experimental results show that our approach is more effective than strong previous methods using heuristics and reinforcement learning. With the same number of labeled data, our approach improves the accuracy of NER by 11.91%. Our approach is superior to state-of-the-art learning-to-label method, with an improvement of accuracy by 6.49%.

5. Entity Resolution with Hybrid Attention-based Networks
Chenchen Sun, and Derong Shen
Abs. Entity resolution (ER) is an important step of data preprocessing. Deep learning based entity resolution is a growing topic in research communities. Considering that record structure is hierarchical: token, attribute, record, we propose a hybrid attention-based network framework for entity resolution. It synthesizes information from different abstract levels of record hierarchy. Systematic attention mechanisms are exploited in several aspects of ER:

self-attention for internal dependency capture, inter-attention for alignments, and multi-dimensional weight attention for importance discrimination. Also attribute order is taken into account in ER learning for better similarity representations. Moreover, we tackle ER over low-quality data by hybrid soft token alignments. Extensive experiments on 4 datasets are conducted, and the results show that our approach surpasses existing ER approaches.

Research Session 15: Recommendation 2
Session Chair: Yi-Ling Chen (National Taiwan University of Science and Technology)

1. Semi-Supervised Factorization Machines for Review-Aware Recommendation
Junheng Huang, Fangyuan Luo, and Jun Wu
Abs. Textual reviews, as a useful supplementary of the interaction data, has been widely used to enhance the performance of recommender systems, especially when the interaction data is sparse. However, existing solutions to review-aware recommendation only focus on learning more informative features from reviews, yet ignore the insufficient number of training examples, resulting in limited performance improvements. To this end, we propose a co-training style semi-supervised review-aware recommendation model, called Collaborative Factorization Machines (CoFM), to augment the training dataset as well as increase its informativeness. Our CoFM employs two FMs as base predictors, each of which labels unlabeled examples for its peer predictor in the learning process. Specifically, a user-leaded FM and an item-leaded FM are separately built using different reviews to increase the diversity between two predictors. Furthermore, to exploit unlabeled data safely, the labeling confidence is estimated through validating the influence of the labeling of unlabeled examples on the labeled ones. The final prediction is made by linearly blending the outputs of two predictors. Extensive experiments on three real-world benchmarks demonstrate the superiority of CoFM over several state-of-the-art review-aware and semi-supervised recommendation schemes.

2. DCAN: Deep Co-Attention Network by Modeling User Preference and News Lifecycle for News Recommendation
Lingkang Meng, Chongyang Shi, Shufeng Hao, and Xiangrui Su

Abs. Personalized news recommendation systems aim to alleviate information overload and provide users with personalized reading suggestions. In general, each news has its own lifecycle that is depicted by a bell-shaped curve of clicks, which is highly likely to influence users' choices. However, existing methods typically depend on capturing user preference to make recommendations while ignoring the importance of news lifecycle. To fill this gap, we propose a Deep Co-Attention Network DCAN by modeling user preference and news lifecycle for news recommendation. The core of DCAN is a Co-Attention Net that fuses the user preference attention and news lifecycle attention together to model the dual influence of users' clicked news. In addition, in order to learn the comprehensive news representation, a Multi-Path CNN is proposed to extract multiple patterns from the news title, content and entities. Moreover, to better capture user preference and model news lifecycle, we present a User Preference LSTM and a News Lifecycle LSTM to extract sequential correlations from news representations and additional features. Extensive experimental results on two real-world news datasets demonstrate the significant superiority of our method and validate the effectiveness of our Co-Attention Net by means of visualization.

3. Considering Interaction Sequence of Historical Items for Conversational Recommender System

Xintao Tian, Yongjing Hao, Pengpeng Zhao, deqing wang, Yanchi Liu, and Victor S. Sheng

Abs. Different from the traditional recommender systems with content-based and collaborative filtering, conversational recommender systems (CRS) can dynamically dialogue with users to capture fine-grained preferences. Although several efforts have been made for CRS, they neglect the importance of interaction sequences, which seek to capture the `context' of users' activities based on actions they have performed recently. Therefore, we propose a framework that considers interaction Sequence of historical items for Conversational Recommendation (SeqCR). Specifically, SeqCR first scores candidate items through the sequence which users interact with. Then it can generate the recommendation list and attributes to be asked based on the scores. We restrict candidate attributes to the ones with high-scoring (high-relevance) items, which effectively reduces the search space of attributes and leads to user preferences that can be hit more quickly and accurately. Finally, SeqCR utilizes

the policy network to decide whether to recommend or ask. We conduct extensive experiments on two datasets from MovieLens 10M and Yelp in multi-round conversational recommendation scenarios. Empirical results demonstrate our SeqCR significantly outperforms the state-of-the-art methods.

4. Knowledge-aware Hypergraph Neural Network for Recommender Systems
Binghao Liu, Pengpeng Zhao, Fuzhen Zhuang, Xuefeng Xian, Yanchi Liu, and Victor S. Sheng
Abs. Knowledge graph (KG) has been widely studied and employed as auxiliary information to alleviate the cold start and sparsity problems of collaborative filtering in recommender systems. However, most of the existing KG-based recommendation models suffer from the following drawbacks, i.e, insufficient modeling of high-order correlations among users, items, and entities, and simple aggregation strategies which fail to preserve the relational information in the neighborhood. In this paper, we propose a Knowledge-aware Hypergraph Neural Network (KHNN) framework to tackle the above issues. First, the knowledgeaware hypergraph structure, which is composed of hyperedges, is employed for modeling users, items, and entities in the knowledge graph with explicit hybrid high-order correlations. Second, we propose a novel knowledge-aware hypergraph convolution method to aggregate different knowledge-based neighbors in hyperedge efficiently. Moreover, it can conduct the embedding propagation of high-order correlations explicitly and efficiently in knowledge-aware hypergraph. Finally, we apply the proposed model on three real-world datasets, and the empirical results demonstrate that KHNN can achieve the best improvements againstother state-of-the-art methods

5. Personalized Dynamic Knowledge-aware Recommendation with Hybrid Explanations
Hao Sun, Zijian Wu, Yue Cui, Liwei Deng, Yan Zhao, and Kai Zheng
Abs. Explainable recommendation is attracting more and more attention in both industry and research communities. While some existing models utilize reviews for improving the performance of recommender systems, most of them assume that user's preference is static and each review's importance is user-independent. However, it is intuitive that user's preference is always dynamically changing and reviews from similar users should be given more importance as they share

similar tastes. Moreover, they achieve model explainability at either feature level that is too concise or review level that is too redundant. To deal with these problems, we propose a Personalized Dynamic Knowledge-aware Recommender (PDKR) for dynamic user modeling and personalized item modeling. In particular, we model user's preference with defined entities and relations in sequential knowledge graphs and capture its dynamics with a novel interval-aware Gated Recurrent Unit (GRU). Furthermore, by leveraging self-attention mechanism, we can not only learn each review's user-specific importance, but also provide tailored explanations for each user at both feature level and review level. We conduct extensive experiments on three benchmark datasets from Amazon and Yelp and the results show that PDKR outperforms all the state-of-the-art recommendation approaches in rating prediction task while providing more effective explanations simultaneously.

Research Session 16: Text and Unstructured Data 3
Session Chair: Iouliana Litou (Athens University of Economics and Business)

1. Unpaired Multimodal Neural Machine Translation via Reinforcement Learning
Yijun Wang, Tianxin Wei, Qi Liu, and Enhong Chen
Abs. End-to-end neural machine translation (NMT) heavily relies on parallel corpora for training. However, high-quality parallel corpora are usually costly to collect. To tackle this problem, multimodal content, especially image, has been introduced to help build an NMT system without parallel corpora. In this paper, we propose a reinforcement learning (RL) method to build an NMT system by introducing a sequence-level supervision signal as a reward. Based on the fact that visual information can be a universal representation to ground different languages, we design two different rewards to guide the learning process, i.e., (1) the likelihood of generated sentence given source image and (2) the distance of attention weights given by image caption models. Experimental results on the Multi30K, IAPR-TC12, and IKEA datasets show that the proposed learning mechanism achieves better performance than existing methods.

2. Multimodal Named Entity Recognition with Image Attributes and Image Knowledge

Dawei Chen, Zhixu Li, Binbin Gu, and Zhigang Chen

Abs. Multimodal named entity extraction is an emerging task which uses both textual and visual information to detect named entities and identify their entity types. The existing efforts are often flawed in two aspects. Firstly, they may easily ignore the natural prejudice of visual guidance brought by the image. Secondly, they do not further explore the knowledge contained in the image. In this paper, we novelly propose a novel neural network model which introduces both image attributes andimage knowledge to help improve named entity extraction. While the image attributes are high-level abstract information of an image that could be labelled by a pre-trained model based on ImageNet, the image knowledge could be obtained from a general encyclopedia knowledge graph with multi-modal information such as DBPedia and Yago. Our emperical study conducted on real-world data collection demonstrates the effctiveness of our approach comparing with several state-of-the-art approaches.

3. Multi-task Neural Shared Structure Search: A Study based on Text Mining

Jiyi Li, and Fumiyo Fukumoto

Abs. Multi-task techniques are effective for handling the problem of small size of the datasets. They can leverage additional rich information from other tasks for improving the performance of the target task. One of the problems in the multi-task based methods is which resources are proper to be utilized as the auxiliary tasks and how to select the shared structures with an effective search mechanism. We propose a novel neural-based multi-task Shared Structure Encoding (SSE) to define the exploration space by which we can easily formulate the multi-task architecture search. For the search approaches, because these existing Network Architecture Search (NAS) techniques are not specially designed for the multi-task scenario, we propose two original search approaches, i.e., m-Sparse Search approach by Shared Structure encoding for neural-based Multi-Task models (m-S4MT) and Task-wise Greedy Generation Search approach by Shared Structure encoding for neural-based Multi-Task models (TGG-S3MT). The experiments based on the real text datasets with multiple text mining tasks show that SSE is effective for formulating the multi-task architecture search. Moreover, both m-S4MT and TGG-S3MT have better performance on the target aspects than the single-task method, multi-label method, naive multi-task methods and the variant of the NAS approach from the

existing works. Especially, 1-S4MT with a sparse assumption on the auxiliary tasks has good performance with very low computation cost.

4. A Semi-structured Data Classification Model with Integrating Tag Sequence and Ngram

Lijun Zhang, Ning Li, Wei Pan, and Zhanhuai Li

Abs. Many collaboratively building resources, such as Wikipedia, Weibo and Quora, exist in the form of semi-structured data and semi-structured data classification plays an important role in many data analysis applications. In addition to content information, semi-structured data also contain structural information. Thus, combining the structure and content features is a crucial issue in semi-structured data classification. In this paper, we propose a supervised semi-structured data classification approach that utilizes both the structural and content information. In this approach, generalized tag sequences are extracted from the structural information, and nGrams are extracted from the content information. Then the tag sequences and nGrams are combined into features called TSGram according to their link relation, and each semi-structured document is represented as a vector of TSGram features. Based on the TSGram features, a classification model is devised to improve the performance of semi-structured data classification. Because TSGram features retain the association between the structural and content information, they are helpful in improving the classification performance. Our experimental results on two real datasets show that the proposed approach is effective.

5. Inferring Deterministic Regular Expression with Unorder and Counting

Xiaofan Wang

Abs. Schema inference has been an essential task in database management, and can be reduced to learning regular expressions from sets of positive finite-sample. In this paper, schemata are inferred from unordered XML documents. We extend the single-occurrence regular expressions (SOREs) to single-occurrence regular expressions with unorder and counting (SOREUCs), and give an inference algorithm for SOREUCs. First, we present a finite automaton with unorder and counting (FAUC). Then, we construct an FAUC for recognizing the given finite sample. Next, the FAUC runs on the given finite sample to obtain counting operators. Finally we transform the FAUC to a

SOREUC by introducing unorder concatenations and counting operators. Experimental results demonstrate that, SOREUCs have stronger expressive power for modeling unordered schemata, and our algorithm can efficiently infer a concise SOREUC with better generalization ability.

Research Session 17: Emerging Applications 2
Session Chair: Zhaojing Luo (National University of Singapore)

1. Susceptible Temporal Patterns Discovery for Electronic Health Records via Adversarial Attack
Rui Zhang, Wei Zhang, Ning LIU, and Jianyong Wang
Abs. The recent advancements in deep neural networks (DNNs) are revolutionizing the healthcare domain. Although many studies try to build medical DNNs model based on historical Electronic Health Records (EHR) and have achieved promising performance in many clinical prediction tasks, recent studies show that DNNs are vulnerable to adversarial attacks. Much of the interest in adversarial examples has stemmed from their ability to shed light on possible limitations of DNNs. However, related research has been receiving sustained attention in computer vision community, how to design adversarial examples for EHR data remains a rarely investigated. To figure out this problem, we propose a novel approach for generating EHR adversarial examples, named as TSAttack, which explores temporal structure contained in EHR to achieve an effective and efficient attack. Based on the generated EHR adversarial examples, we further propose a procedure to discover susceptible temporal patterns (STP) in a patient's medical records, which provide clinical decision support for dynamic monitoring. Extensive experiments on the real-world longitudinal EHR database MIMIC-III have demonstrated the effectiveness of our approach is yielding better performance in adversarial settings.

2. A decision support system for heart failure risk prediction based on weighted naive Bayes
Kehui SONG, Shenglong Yu, Haiwei Zhang, Ying ZHANG, Xiangrui Cai, and Xiaojie Yuan
Abs. Heart failure (HF) affects the health of millions of people worldwide and the early detection of HF risk plays a vital role in prevention and prompt

treatment. Various decision support systems based on machine learning have been presented recently to predict HF. However, the existing systems usually assumed that all features add equal weight to the prediction result, which could not properly simulate the diagnostic status. In this study, a decision support system is proposed for HF prediction using MSE Back Propagation Method (MSEBPM) and weighted naive Bayes. First, the feature selection method eliminates irrelevant features to improve accuracy and decrease computational times. Second, the proposed MSEBPM computes a weight vector for features based on their contributions, trying to minimize the MSE loss of the predicted class probabilities. Finally, the trained weight vector is applied to the weighted naive Bayes model for HF risk prediction. The proposed system is evaluated with a published dataset of 899 patients, and compared with conventional data mining techniques and other state-of-the-art systems. The results show that our proposed system leads to 82.96% accuracy in HF risk prediction, which suggests that it could be used to early detect HF in the clinic.

3. Inheritance-guided Hierarchical Assignment for Clinical Automatic Diagnosis
Yichao Du, Pengfei Luo, Xudong Hong , University of Science and Technology of China, Tong Xu, Zhe Zhang, Chao Ren, Yi Zheng, and Enhong Chen
Abs. Clinical diagnosis, which aims to assign diagnosis codes for a patient based on the clinical note, plays an essential role in clinical decision-making. Considering that manual diagnosis could be error-prone and time-consuming, many intelligent approaches based on clinical text mining have been proposed to perform automatic diagnosis. However, these methods may not achieve satisfactory results due to the following challenges. First, most of the diagnosis codes are rare, and the distribution is extremely unbalanced. Second, existing methods are challenging to capture the correlation between diagnosis codes. Third, the lengthy clinical note leads to the excessive dispersion of key information related to codes. To tackle these challenges, we propose a novel framework to combine the inheritance-guided hierarchical assignment and co-occurrence graph propagation for clinical automatic diagnosis. Specifically, we propose a hierarchical joint prediction strategy to address the challenge of unbalanced codes distribution. Then, we utilize graph convolutional neural networks to obtain the correlation and semantic representations of medical ontology. Furthermore, we introduce multi attention mechanisms to extract

crucial information. Finally, extensive experiments on MIMIC-III dataset clearly validate the effectiveness of our method.

4. BPTree: An Optimized Index with Batch Persistence on Optane DC PM
Chenchen Huang, Huiqi Hu, and Aoying Zhou
Abs. Intel Optane DC Persistent Memory (PM) is the first commercially available PM product. Although it meets many hypothesises about PM in previous studies, some other design considerations are observed in subsequent tests. For instance, 1) the internal data access granularity in Optane DC PM is 256B, accesses smaller than 256B will cause read/write amplification; 2) the locking overhead will be amplified when the PM operations are included in the critical area or the lock is added on PM. In this paper, we propose a novel persistent index called BPTree to fit with these new features. The core idea of BPTree is to buffer multiple writes in DRAM first, and later persist them in batches to PM to reduce the write amplification. We add a buffer layer in BPTree to enable the batch persistence, and design a GC-friendly log structure on PM to guarantee the buffer's durability.To improve the scalability, we also implement a hybrid concurrency control strategy to ensure most of the operations on PM are lock-free, and move the lock from PM to DRAM for the operations that must be locked. Our experiments on Optane DC PM show that BPTree reduces 256B PM writes by a factor of 1.95-2.48x compared to the state-of-the-art persistent indexes. Moreover, BPTree has better scalability in the concurrent environment.

5. An improved dummy generation approach for enhancing user location privacy
Shadaab Siddiqie, Anirban Mondal, and Krishna Reddy P
Abs. Location-based services (LBS), which provide personalized and timely information, entail privacy concerns such as unwanted leak of current user locations to potential stalkers. Existing works have proposed dummy generation techniques by creating a cloaking region (CR) such that the user's location is at a fixed distance from the center of CR. Hence, if the adversary somehow knows the location of the center of CR, the user's location would be vulnerable to attack. We propose an improved dummy generation approach for facilitating improved location privacy for mobile users. Our performance study

demonstrates that our proposed approach is indeed effective in improving user location privacy.

Research Session 18: Recommendation 3
Session Chair: Meng-Fen Chiang (The University of Auckland)

1. Graph Attention Collaborative Similarity Embedding for Recommender System
Jinbo Song, Chao Chang, Fei Sun, Zhenyang Chen, Guoyong Hu, and Peng Jiang
Abs. We present Graph Attention Collaborative Similarity Embedding (GACSE), a new recommendation framework that exploits collaborative information in the user-item bipartite graph for representation learning. Our framework consists of two parts: the first part is to learn explicit graph collaborative filtering information such as user-item association through embedding propagation with attention mechanism, and the second part is to learn implicit graph collaborative information such as user-user similarities and item-item similarities through auxiliary loss. We design a new loss function that combines BPR loss with adaptive margin and similarity loss for the similarities learning. Extensive experiments on three benchmarks show that our model is consistently better than the latest state-of-the-art models.

2. Learning Disentangled User Representation Based on Controllable VAE for Recommendation
Yunyi Li, Pengpeng Zhao, deqing wang, Xuefeng Xian, Yanchi Liu, and Victor S. Sheng
Abs. User behaviour on purchasing is always driven by complex latent factors, which are highly disentangled in the real world. Learning latent factorized representation of users can uncover user intentions behind the observed data (i.e. user-item interaction) and improve the robustness and interpretability of the recommender system. However, existing collaborative filtering methods learning disentangled representation face problems of balancing the trade-off between reconstruction quality and disentanglement. In this paper, we propose a controllable variational autoencoder framework for collaborative filtering. Specifically, we adopt a modified Proportional-Integral-Derivative (PID) control

to the β-VAE objective to automatically tune the hyperparameter β using the output of Kullback-Leibler divergence as feedback. We further introduce item embeddings to guide the system to learn representation related to the real-world concepts using a factorized Gaussian distribution. Experimental results show that our model can get a crucial improvement over state-of-the-art baselines. We further evaluate our model's effectiveness to control the trade-off between reconstruction error and disentanglement quality in the recommendation.

3. DFCN: An Effective Feature Interactions Learning Model for Recommender Systems

wei yang, and tianyu hu

Abs. Data features in real industrial recommendation scenarios are diverse, high-dimensional and sparse. Effective feature crossing can improve the performance of recommendation, which is of great significance. Manual feature engineering is no longer applicable due to its high cost and low efficiency. Factorization machines introduce the second-order feature interactions to enhance learning ability. Deep neural networks (DNNs) have good nonlinear combination ability and can learn high-order feature interactions. However, DNNs implicitly learn feature interactions at the bit-wise level is not always effective. In this paper, we propose a novel factorization cross network (FCN), which is based on factorization to learn explicit feature crossing through neural network. FCN can learn low- and high-order feature interactions at the vector-wise level with linear time complexity. We introduce deep residual network (DRN) to learn implicit feature interactions. We further use learnable parameters to combine FCN and DRN, and name the new model as deep factorization cross network (DFCN). DFCN can automatically learn low- and high-order explicit and implicit feature interaction information. We have carried out comprehensive experiments on three real-world datasets. Experimental results demonstrate the effectiveness of DFCN, which performs best compared with other competitive models.

4. Tell me Where to Go Next: Improving POI Recommendation via Conversation

Changheng Li, Yongjing Hao, Pengpeng Zhao, Fuzhen Zhuang, Yanchi Liu, and Victor S. Sheng

Abs. Next Point-of-Interest (POI) recommendation estimates user preference on POIs according to past check-in history, suffering from the intrinsic limitation of obtaining dynamic user preferences. Conversational Recommendation System (CRS), which can collect dynamic user preferences through conversation, brings a solution to the above limitation. However, none of the existing CRS methods consider the spatio-temporal factors in the action selection phase, which are essential for POI conversational recommendation. In this paper, we propose a new Spatio-Temporal Conversational Recommendation System (STCRS) to fuse the spatio-temporal information and dialogue for next POI recommendation. Specifically, STCRS first learns the spatio-temporal information in the user's check-in history. Then reinforcement learning is used to decide which action (asking for an attribute or recommending POIs) to take at the next turn to achieve successful POI recommendation within as few turns as possible. Finally, our extensive experiments on two real-world datasets demonstrate significant improvements over the state-of-the-art methods.

5. MISS: A Multi-user Identification Network for Shared-account Session-aware Recommendation

Xinyu Wen, Zhaohui Peng, Shanshan Huang, Senzhang Wang, and Philip S Yu

Abs. The user's interactions with the system within a given time frame are organized into a session. The task of session-aware recommendation aims to predict the next interaction based on user’s historical sessions and current session. Though existing methods have achieved promising results, they still have drawbacks in some aspects. First, most existing deep learning methods model a session as a sequence, but neglect the complex transition relationships between items. Second, a single account is usually regarded as a single user by default, where the scenario of multiple users sharing the same account is ignored. To this end, we propose a Multi-user Identification network named MISS for the Shared-account Session-aware recommendation problem. MISS consists of two core components: one is the Dwell Graph Neural Network (DGNN), which incorporates item dwell time into the gated graph neural network to capture user interest drift across sessions. The other is a Multi-user Identification (MI) module, which draws on the attention mechanism to distinguish behaviors of different users under the same account. To verify the effectiveness of MISS, we construct two data sets with shared account

characteristics from real-world smart TV watching logs. Extensive experiments conducted on the two data sets demonstrate that MISS evidently outperforms the state-of-the-art recommendation methods.

Research Session 19: Data Mining 2
Session Chair: Fan Zhang (Guangzhou University)

1. A Local Similarity-Preserving Framework for Nonlinear Dimensionality Reduction with Neural Networks
Xiang Wang, Xiaoyong Li, Junxing Zhu, Xu Zichen, Kaijun Ren, Weiming Zhang, Xinwang Liu, and Kui Yu
Abs. Real-world data usually have high dimensionality and it is important to mitigate the curse of dimensionality for data analysis, communication, visualization, and storage of high-dimensional data. Most of existing methods for local dimensionality reduction obtain an embedding with the eigenvalue or singular value decomposition, where the computational complexities are very high for a large amount of data. Here we propose a novel local nonlinear approach named Vec2vec for general purpose dimensionality reduction, which generalizes recent advancements in embedding representation learning of words to dimensionality reduction of matrices. It obtains the nonlinear embedding using a neural network with only one hidden layer to reduce the computational complexity. To train the neural network, we build the neighborhood similarity graph of a matrix and define the context of data points by exploiting the random walk properties. Experiments demenstrate that Vec2vec is more efficient than the state-of-the-art local dimensionality reduction methods in a large number of high-dimensional data. Extensive experiments of data classification and clustering on eight real datasets show that Vec2vec is better than several classical dimensionality reduction methods in the statistical hypothesis test, and it is competitive with recent developed state-of-the-art UMAP.

2. AE-UPCP: Seeking Potential Membership Users by Audience Expansion Combining User Preference with Consumption Pattern
Xiaokang Xu, Zhaohui Peng, Senzhang Wang, Shanshan Huang, Philip S Yu, Zhenyun Hao, Jian Wang, and Xue Wang

Abs. Many online video websites or platforms provide membership service, such as YouTube, Netflix and iQIYI. Identifying potential membership users and giving timely marketing activities can promote membership conversion and improve website revenue. Audience expansion is a viable way, where existing membership users are treated as seed users, and users similar to seed users are expanded as potential memberships. However, existing methods have limitations in measuring user similarity only according to user preference, and do not take into account consumption pattern which refers to aspects that users focus on when purchasing membership service. So we propose an Audience Expansion method combining User Preference and Consumption Pattern (AE-UPCP) for seeking potential membership users. An autoencoder is designed to extract user personalized preference and CNN is used to learn consumption pattern. We utilize attention mechanism and propose a fusing unit to combine user preference with consumption pattern to calculate user similarity realizing audience expansion of membership users. We conduct extensive study on real datasets demonstrating the advantages of our proposed model.

3. Self Separation and Misseparation Impact Minimization for Open-Set Domain Adaptation

Yuntao Du, Yikang Cao, Yumeng Zhou, Yinghao Chen, Ruiting Zhang, and Chongjun Wang

Abs. Domain adaptation (DA) has achieved great success in the past few years. Most of the existing DA algorithms assume the label space in the source domain and the target domain are exactly the same. However, when applied to wild applications, such a strict assumption is difficult to satisfy. In this paper we focus on Open Set Domain adaptation (OSDA), a more realistic setting, where the target domain data contains unknown classes which do not exist in the source domain. We concluded two main challenges in OSDA: (i) Separation: Accuratly separating the target domain into a known domain and an unknown domain. (ii) Distribution Matching: deploying appropriate domain adaptation between the source domain and the target known domain. However, existing separation methods highly rely on the similarity of the source domain and the target domain and have ignored the fact that the distribution information of the target domain could help up with better separation. In this paper, we propose Self Separation and Misseparation Impact Minimization, a

semi-supervised-learning-like algorithm which explores the distribution information of the target domain to improve separation accuracy. Further, we also take into account the possible misseparated samples (the unknown-class samples which are wrongly separated as known-class) in the distribution matching step. By maximizing the discrepancy between the target known domain and the target unknown domain, we could further reduce the impact of misseparation in distribution matching. Experiments on several benchmark datasets show our algorithm outperforms state-of-the-art methods.

Research Session 20: Graph Data 3
Session Chair: Yixiang Fang (The Chinese University of Hong Kong, Shenzhen)

1. Sequence Embedding for Zero or Low Resource Knowledge Graph Completion
Zhijuan Du

Abs. Knowledge graph completion (KGC) has been proposed to improve KGs by filling in missing links. Previous KGC approaches require a large number of training instances (entity and relation) and hold a closed-world assumption. The real case is that very few instances are available and KG evolve quickly with new entities and relations being added by the minute. The newly added cases are zero resource in training. In this work, we propose a **S**equence **E**mbedding **w**ith **A**dversarial learning approach (SEwA) for zero or low resource KGC. It transform the KGC into a sequence prediction problem by making full use of inherently link structure of knowledge graph and resource-easy-to-transfer feature of adversarial contextual embedding. Specifically, the triples ($< h, r, t >$)and higher-order triples ($< h, p, t >$) containing the paths ($p = r_1 \rightarrow \cdots \rightarrow r_n$) are represented as word sequences and are encoded by pre-training model with multi head self-attention. The path is obtained by a non-parametric learning based on the one-class classification of the relation trees. The zero and low resources issues are further optimizes by adversarial learning. At last, our SEwA is evaluated by low resource datasets and open world datasets.

2. HMNet: Hybrid Matching Network for Few-shot Link Prediction

Shan Xiao, Lei Duan, Guicai Xie, Renhao Li, Zihao Chen, Geng Deng, and Jyrki Nummenmaa

Abs. Knowledge graphs (KGs) are widely used in many real-world applications, such as information retrieval, question answering system, and personal recommendation. However, most KGs are suffering from the incompleteness problem. To deal with the task of link prediction, previous knowledge graph embedding methods require numerous reference instances for each relation. It is worth noting that most relations in KGs have only a few reference instances available. Existing works for few-shot link prediction evaluate the authenticity of triplets from a single relation perspective. In this paper, we propose Hybrid Matching Network (HMNet) for few shot link prediction, evaluating triplets from entity and relation two perspectives. At the entity-aware matching network, HMNet uses attentive inductive embedding layer to aggregate entity features and relation-aware topology, and then provides entity-aware score to implement first perspective evaluation. At the relation-aware matching network, HMNet integrates feature attention mechanism to implement relation perspective evaluation. Experiments on two public datasets indicate that HMNet achieves promising performance in few-shot link prediction.

3. OntoCSM: Ontology-Aware Characteristic Set Merging for RDF Type Discovery

Pengkai Liu, Shunting Cai, Baozhu Liu, and Xin Wang

Abs. With the growing popularity and application of knowledge-based artificial intelligence, the scale of knowledge graph data is dramatically increasing. The RDF, as one of the mainstream models of knowledge graphs, is widely used to describe the characteristics of Web resources due to its simplicity and flexibility. However, RDF datasets are usually incomplete (without rdf:type information) and noisy, which hinders downstream tasks. RDF entities can be characterized by their characteristic sets that is the sets of predicates of the RDF entities. Since untyped entities can be assigned to closest types by merging characteristic sets, optimally merging characteristic sets has become a crucial issue. In this paper, aiming at the Optimal Characteristic Set Merge Problem (OCSMP), we propose an Ontology-Aware Characteristic Set Merging algorithm, called OntoCSM, which extracts an ontology structure using RDF characteristic sets and guides

the merging process by optimizing the objective function. Extensive experiments on various datasets show that the efficiency of OntoCSM is generally higher than that of the state-of-the-art algorithms and can be improved by orders of magnitude in the best case. The accuracy and scalability of our method have been verified, which shows that OntoCSM can reach competitive results to the existing algorithms while being ontology-aware.

4. EDKT: An Extensible Deep Knowledge Tracing Model for Multiple Learning Factors

Liangliang He, Jintao Tang, Xiao Li, and Ting Wang

Abs. Knowledge Tracing (KT) refers to the problem of predicting learners' future potential performance given their past learning history in e-learning systems. In order to better trace the learners' knowledge, KT tasks have become increasingly complicated recently, and various factors related to learning (such as skill, exercise, hint, etc.) have been incorporated into the modeling of knowledge state of the learner, which renders it inadequate for the traditional KT definition to formalize these tasks. Therefore, this paper first gives a more general formal definition of KT tasks, and then proposes an Extensible Deep Knowledge Tracing model for multiple learning factors based on this general definition, named EDKT. EDKT can integrate various different learning factors by extending or ablating factors in two plug-ins on the basis of minor modifications. To demonstrate the effectiveness of the proposed model, we conduct extensive experiments on three real-world benchmark datasets, and the results show that EDKT comprehensively outperforms the state-of-the-art KT models on predicting future learner responses.

5. Fine-grained Entity Typing via Label Noise Reduction and Data Augmentation

Haoyang Li, Xueling Lin, and Lei Chen

Abs. Fine-grained entity typing aims to assign one or more types for entity mentions in the corpus. Recently, distant supervision has been utilized to generate training data. However, it has two drawbacks. First, the same labels are assigned to every entity mention in a context-agnostic manner, which introduces label noise. Some approaches alleviate this issue by hand-crafted features. However, they require efforts from experts. Second, the entity mentions out of

Knowledge Base (KB) are ignored and hence cannot be added to the training data, which decreases the size of the training data. Furthermore, the existing entity typing systems neglect the types of other entity mentions in the same context which provide evidence to infer the types of the target entity mentions. In this paper, we first propose graph-based and sampling-based approaches, to reduce the label noise generated by the distant supervision, and then augment the training data by finding potential entity mentions in the corpus and inferring their types. Moreover, we propose a hierarchical neural network, which involves the types of other mentions in the context and satisfies the type consistency, to predict the types. Experiments on two datasets show that our system outperforms state-of-the-art entity typing systems.

Research Session 21: Spatial/Temporal Data 3
Session Chair: Jianqiu Xu(Nanjing University of Aeronautics and Astronautics)

1. Missing POI Check-in Identification Using Generative Adversarial Networks
Meihui Shi, Derong Shen, Yue Kou, Tiezheng Nie, and Ge Yu
Abs. The missing point-of-interest (POI) check-ins in real-life mobility data prevent advanced analysis of users' preferences and mobile patterns. Existing approaches for missing POI check-in identification mainly focus on modelling spatio-temporal dependencies and memorising transition patterns through users' check-in sequences. However, these methods cannot ensure that the generated missing records obey the same distribution as the observed check-ins. To this end, we propose a novel Bi-G2AN model, which fuses the merits of generative adversarial network (GAN) and bi-directional gated recurrent unit (GRU), to identify the missing POI check-ins. Specifically, we develop a GAN-based method to mimic the overall distribution of a given check-in dataset, and it is further utilized to generate more reasonable missing POI check-ins. In order to capture bi-directional dependencies and historical impact, a modified bi-directional GRU is utilized in GAN. Moreover, both spatio-temporal influence and local motion information are employed to learn users' dynamic preferences. Finally, experiments conducted on three real datasets demonstrate the competitiveness of the Bi-G2AN model, outperforming state-of-the-art approaches.

2. Efficiently Discovering Regions of Interest with User-Defined Score Function

Qiyu Liu, Libin Zheng, Xiang Lian, and Lei Chen

Abs. Region of Interest (ROI) queries are of great importance in many location based services. However, the previous studies on ROI queries usually adopt either a simple spatial data model or a non-flexible enough query geometry, e.g., fixed-size rectangle. In this paper, to fix these drawbacks, we propose a new ROI search operator called *Radius Bounded ROI* (RBR) query. An RBR query retrieves a subset of spatial objects satisfying colocation constraints and maximizing a user-configurable score function at the same time. We formally prove that answering an RBR query is 3SUM-hard, which implies that it is unlikely to find a sub-quadratic solution. To answer the RBR queries efficiently, we propose three algorithms, PairEnum, BaseRotation and OptRotation based on novel geometric findings. In addition, the query processing technique we proposed can be easily extended to other related problems like top-k ROI search. To demonstrate both efficiency and effectiveness of our proposed algorithms, we conduct extensive experimental studies on both real-world datasets and synthetic benchmarks, and the results show that OptRotation, our most efficient algorithm, achieves more than 10^3x efficiency improvement on both real and synthetic datasets compared with the baseline algorithm.

3. An Attention-Based Bi-GRU for Route Planning and Order Dispatch of Bus-Booking Platform

Yucen Gao, Yuanning Gao, Yuhao Li, Xiaofeng Gao, Xiang Li, and Guihai Chen

Abs. To cope with the high needs from passengers, especially for airports at night, we plan to develop a novel bus-booking platform, which can dispatch several passenger orders to one bus together. In this paper, we first give the formal definition of the Order Dispatch and Route Planning (ODRP) problem for the new bus-booking platform, and prove the ODRP problem is NP-hard. We then propose a new method based on attention mechanism and Bi-directional Gated Recurrent Unit (Bi-GRU) to realize the tasks of order dispatch and route planning simultaneously. To the best of our knowledge, this is the first method that uses main ideas of attention mechanism and Bi-GRU in order dispatch and route planning issues related to urban bus system. It can achieve the goal of

increasing passenger number and reducing platform costs. Through experiments based on real-world data, we prove the effectiveness of the proposed method.

4. Top-k Closest Pair Queries over Spatial Knowledge Graph
Fangwei Wu, Xike Xie, and Jieming Shi
Abs. Recently, RDF data has been enriched with spatial semantics enabling spatial keyword search. Research spatial keyword search over spatial RDF data focus on finding the spatial entities rooted at subtrees which cover given query keywords. In this work, we study how relevant spatial entity pairs can be efficiently retrieved, where the relevance is determined according to both spatial distances and textual similarities. The retrieved top-k closest pairs are ranked and then returned to users for the interests of business intelligence and recommendation. We propose a branch-and-bound framework associated with effective lower and upper bound pruning techniques and early stopping conditions for efficiently retrieving relevant top-k closet pairs. The results demonstrate the high efficiency of our proposal compared to baseline solutions.

5. HIFI: Anomaly Detection for Multivariate Time Series with High-order Feature Interactions
Liwei Deng, Xuanhao Chen, Yan Zhao, and Kai Zheng
Abs. Monitoring complex systems results in massive multivariate time series data, and anomaly detection of these data is very important to maintain the normal operation of the systems. Despite the recent emergence of a large number of anomaly detection algorithms for multivariate time series, most of them ignore the correlation modeling among multivariate, which can often lead to poor anomaly detection results. In this work, we propose a novel anomaly detection model for multivariate time series with HIgh-order Feature Interactions (HIFI). More specifically, HIFI builds multivariate feature interaction graph automatically and uses the graph convolutional neural network to achieve high-order feature interactions, in which the long-term temporal dependencies are modeled by attention mechanisms and a variational encoding technique is utilized to improve the model performance and robustness. Extensive experiments on three publicly available datasets demonstrate the superiority of our framework compared with state-of-the-art approaches.

Research Session 22: Recommendation 4
Session Chair: Jianzhong Qi (The University of Melbourne)

1. VizGRank: A Context-Aware Visualization Recommendation Method Based on Inherent Relations Between Visualizations
Qianfeng Gao, Zhenying He, Yinan Jing, Kai Zhang, and X. Sean Wang
Abs. Visualization recommendation systems measure the importance of visualizations to make suggestions. While considering each visualization individually may be enough to gauge its importance in specific scenarios, it ignores the relations between visualizations under a visual analysis context. This paper is to study a strategy via a more general method called VizGRank which models the relations between visualizations as a graph, then calculates the importance of visualizations by adopting a graph-based algorithm. In this model, the relations derived from the visual encoding of the visualizations and the underlying data schema are used for recommendation. Due to the lack of public benchmarks, the effectiveness of the model is evaluated on the synthetic results from an existing public benchmark IDEBench as a workaround. However, since the existing benchmark is specific and synthetic and does not reflect the realistic scenarios of visualization recommendation completely, a new benchmark for visualization recommendation is designed and constructed by collecting real public datasets. Extensive experiments on both the public benchmark and the new benchmark demonstrate that the VizGRank can better capture the relative importance of visualization and outperforms the existing state-of-the-art method.

2. Deep User Representation Construction Model for Collaborative Filtering
Daomin Ji, Zhenglong Xiang, and Yuanxiang Li
Abs. Model-based collaborative filtering (CF) methods can be divided into user-item methods and item-item methods. In most cases, both of them can be seen as modeling the user-item interaction and the only difference between them is that they adopt different ways to build user representations. User-item methods obtain user representations by directly assigning each user a real-valued vector and do not consider users' historical item information. However, users' historical item information can reflect users' preferences to some extent and can alleviate the problem of data sparsity. Ignoring this information may lead to incomplete construction of user representations and vulnerability to data

sparsity. Although existing item-item methods address this problem by using the users' historical items to build the user representations, they always use the same vector to represent the same historical item for different users, which may limit the expressiveness and further improvement of the models. In this paper, we propose Deep User Representation Construction Model (DURCM) to construct user presentations in a more effective and robust way. Specially, different from existing item-item methods that directly use historical item vectors to build user representations, we first adopt a conversion module to convert a user's historical item vectors into personalized item vectors, which enables that even the same item to have different expressions for different users. Second, we design a special attention module to automatically assign weights to these personalized item vectors when constructing the users' final representations. We conduct comprehensive experiments on four real-world datasets and the results verify the effectiveness of our proposed methods.

3. DiCGAN: A Dilated Convolutional Generative Adversarial Network for Recommender Systems

Zhiqiang Guo, Chaoyang Wang, Jianjun Li, Guohui Li, and Peng Pan

Abs. Generative Adversarial Network (GAN) has recently been introduced into the domain of recommendation due to its ability of learning the distribution of users' preferences. However, most existing GAN-based recommendation methods only exploit the user-item interactions, while ignoring to leverage the information between user's interacted items. On the other hand, Convolutional Neural Network (CNN) has shown its power in learning high-order correlations. In this paper, combining with the strengths of both GAN and CNN, we propose a Dilated Convolutional Generative Adversarial Network (DiCGAN) for recommendation, in which we first embed the interacted items of per user into an image in a latent space, and then use several dilated convolutional filters and a vertical convolutional filter to capture the high-order correlations among the interacted items. Moreover, an attention module is employed before convolution to generate attention maps for adaptive feature refinement. Experiments on several public datasets verify the superiority of DiCGAN over several baselines in terms of top-N recommendation. Further more, our experimental results show that when the dataset is more large and sparse, the performance gain of DiCGAN is also more significant, demonstrating the effectiveness of the CNN

component in extracting high-order correlations from interacted data for better performance.

4. RE-KGR: Relation-Enhanced Knowledge Graph Reasoning for Recommendation

Ming He, Hanyu Zhang, and Han Wen

Abs. A knowledge graph (KG) has been widely adopted to improve recommendation performance. The multi-hop user-item connections in a KG can provide reasons for recommending an item to a user. However, existing methods do not effectively leverage the relations of entities and interpretable paths in a KG. To address this limitation, in this paper, we propose a novel recommendation framework called relation-enhanced knowledge graph reasoning for recommendation (RE-KGR) that combines recommendation and explainability by reasoning user-item interaction paths (UIIPs). First, instead of applying an alignment algorithm for preprocessing, RE-KGR directly learns the semantic representation of entities from structured knowledge by stacking relation-based convolutional layers to take full advantage of the KG. Moreover, RE-KGR infers user preferences by calculating the sum of all UIIPs between users and items. Finally, RE-KGR selects several UIIPs with the highest probabilities as possible reasons for the recommendations. Extensive experiments on three real-world datasets demonstrate that our proposed method significantly outperforms several state-of-the-art baselines and achieves superior performance and explainability.

5. LGCCF: A Linear Graph Convolutional Collaborative Filtering with Social Influence

Ming He, Han Wen, and Hanyu Zhang

Abs. Collaborative filtering (CF) is the dominant technique in personalized recommendation. It models user-item interactions to select the relevant items for a user, and it is widely applied in real recommender systems. Recently, graph convolutional network (GCN) has been incorporated into CF, and it achieves better performance in many recommendation scenarios. However, existing works usually suffer from limited performance due to data sparsity and high computational costs in large user-item graphs. In this paper, we propose a linear graph convolutional CF (LGCCF) framework that incorporates the social

influence as side information to help improve recommendation and address the aforementioned issues. Specifically, LGCCF integrates the user-item interactions and the social influence into a unified GCN model to alleviate data sparsity. Furthermore, in the graph convolutional operations of LGCCF, we remove the nonlinear transformations and replace them with linear embedding propagations to overcome training difficulty and improve the recommendation performance. Finally, extensive experiments conducted on two real datasets show that LGCCF consistently outperforms the state-of-the-art recommendation methods.

Research Session 23: Text and Unstructured Data 4
Session Chair: Hong-Han Shuai (National Yang Ming Chiao Tung University)

1. MACROBERT: Maximizing Certified Region of BERT to Adversarial Word Substitutions
Fali Wang, Zheng Lin, Zhengxiao Liu, Mingyu Zheng, Lei Wang, and Daren Zha
Abs. Deep neural networks are deemed to be powerful but vulnerable, because they will be easily fooled by carefully-crafted adversarial examples. Therefore, it is of great importance to develop models with certified robustness, which can provably guarantee that the prediction will not be easily misled by any possible attack. Recently, although a certified method based on randomized smoothing is proposed, it does not take the maximized certified region into account, so we develop an approach to train models with maximized certified regions via replacing the base classifier with the soft smoothed classifier which is differentiable during propagation.

2. A Diversity-Enhanced and Constraint-Relaxed Augmentation for Low-Resource Classification
Liu Guang, Huang Hailong, Mao Yuzhao, Gao Weiguo, Li Xuan, and Jianping Shen
Abs. Previous studies on Data Augmentation (DA) mostly use a fine-tuned Language Model (LM) to strengthen the constraints but ignore the fact that the potential of diversity could improve the effectiveness of generated data. To address this dilemma, we propose a Diversity-Enhanced and Constraints-Relaxed Augmentation (DECRA) that has two essential components on top of a transformer-based backbone model, including a k-β augmentation

and a masked language model loss. Extensive experiments demonstrate that our DECRA outperforms state-of-the-art approaches by 3.8% in the overall score.

3. Neural Demographic Prediction in Social Media with Deep Multi-View Multi-Task Learning

Yantong Lai, Yijun Su, Cong Xue, and Daren Zha

Abs. Utilizing the demographic information of social media users is very essential for personalized online services. However, it is difficult to collect such information in most realistic scenarios. Luckily, the reviews posted by users can provide rich clues for inferring their demographics, since users with different demographics such as gender and age usually have differences in their contents and expressing styles. In this paper, we propose a neural approach for demographic prediction based on user reviews. The core of our approach is a deep multi-view multi-task learning model. Our model first learns context representations from reviews using a context encoder, which takes semantics and syntactics into consideration. Meanwhile, we learn sentiment and topic representations from selected sentiment and topic words using a word encoder separately, which consists of a convolutional neural network to capture the local contexts of reviews in word-level. Then, we learn a unified user representation from context, sentiment and topic representations and apply multi-task learning for inferring user's gender and age simultaneously. Experimental results on three real-world datasets validate the effectiveness of our approach. To facilitate future research, we release the codes and datasets at https://github.com/icmpnorequest/DASFAA2021_DMVMT.

4. An Interactive NL2SQL Approach With Reuse Strategy

Xiaxia Wang, Sai Wu, Lidan Shou, and Ke Chen

Abs. This paper studies a recently proposed task that maps contextual natural language questions to SQL queries in a multi-turn interaction. Instead of synthesizing an SQL query in an end-to-end way, we propose a new model which first generates an SQL grammar tree, called Tree-SQL, as the intermediate representation, and then infers an SQL query from the Tree-SQL with domain knowledge. For semantic dependency among context-dependent questions, we propose a reuse strategy that assigns a probability for each sub-tree of historical TreeSQLs. On the challenging contextual Text-to-SQL

benchmark SParC with the ‘value selection’ task which includes values in queries, our approach achieves SOTA accuracy of 48.5% in question execution accuracy and 21.6% in interaction execution accuracy. In addition, we experimentally demonstrate the significant improvements on the reuse strategy.

Research Session 24: Spatial/Temporal Data 4
Session Chair: R. Uday Kiran (The University of Aizu)

1. Incentive-aware Task Location in Spatial Crowdsourcing
Fei Zhu, Shushu Liu, Junhua Fang, and An Liu
Abs. With the popularity of wireless network and mobile devices, spatial crowdsourcing has gained much attention from both academia and industry. One of the critical components in spatial crowdsourcing is task-worker matching, where workers are assigned to tasks to meet some pre-defined objectives. Previous works generally assume that the locations of tasks are known in advance. However, this does not always hold, since in many real life applications, where to put tasks is not specific and needs to be determined on the fly. In this paper, we propose Incentive-aware Task Location (ITL), a novel problem in spatial crowdsourcing. Given a location-unspecific task with a fixed budget, the ITL problem seeks multiple locations to place the task and allocates the given budget to each location, such that the number of workers who are willing to participate the task is maximized. We prove that the ITL problem is NP-hard and propose three heuristic methods to solve it, including even clustering, uneven clustering and greedy location methods. Through extensive experiments on a real dataset, we demonstrate the efficiency and effectiveness of the proposed methods.

2. Efficient Trajectory Contact Query Processing
Pingfu Chao, Dan He, Lei Li, Mengxuan Zhang, and Xiaofang Zhou
Abs. During an infectious disease outbreak, the contact tracing is regarded as the most crucial and effective way of disease control. As the users' trajectories are widely obtainable due to the ubiquity of positioning devices, the contact tracing can be achieved by examining trajectories of confirmed patients to identify other trajectories that are contacted either directly or indirectly. In this paper, we propose a generalised Trajectory Contact Search (TCS) query, which models the

contact tracing problem as well as other similar trajectory-based problems. In addition, we answer the query by proposing an iterative algorithm that finds contacted trajectories progressively along the transmission chains, and we further optimise each iteration in terms of time and space efficiency by proposing a hop scanning algorithm and a grid-based time interval tree. Extensive experiments on large-scale real-world data demonstrate the effectiveness of our proposed solutions over baseline algorithms.

3. STMG: Spatial-Temporal Mobility Graph for Location Prediction
Xuan Pan, Xiangrui Cai, Jiangwei Zhang, Yanlong Wen, Ying Zhang, and Xiaojie Yuan
Abs. Location-Based Social Networks (LBSNs) data reflects a large amount of user mobility patterns. So it is possible to infer users' unvisited Points of Interest (POIs) through the users' check-in records in LBSNs. Existing location prediction approaches typically regard user check-ins as sequences, while they ignore the spatial and temporal correlations between non-adjacent records. Moreover, the serialized form is insufficient to analog user complex POI moving behaviors. In this paper, we model user check-in records as a graph, named Spatial-Temporal Mobility Graph (STMG), where the nodes and edges fuse the spatial-temporal information in absolute and relative aspect respectively. Based on STMG, we propose a location prediction model named Spatial-temporal Enhanced Graph Neural Network (SEGN). In SEGN, the STMG nodes are encoded as the embeddings with specific time and location semantics. Last but not the least, we introduce three kinds of matrices, which completely depict the user moving behaviors among POIs, as well as the relative relationships of time and location on STMG edges. Extensive experiments on three real-world LBSNs datasets demonstrate that with specific time information, SEGN outperforms seven state-of-the-art approaches on four metrics.

4. Shadow: Answering Why-not Questions On Top-k Spatial Keyword Queries Over Moving Objects
Wang Zhang, Yanhong Li, Lihchyun Shu, Changyin Luo, and Jianjun Li
Abs. The popularity of mobile terminals has generated massive moving objects with spatio-textual characteristics. A Top-k spatial keyword query over moving objects (Top-k SKM query) returns the Top-k objects, moving or static, based on

a ranking function that considers spatial distance and textual similarity between the query and objects. To the best of our knowledge, there hasn't been any research into the why-not questions on Top-k SKM queries. Aiming at this kind of why-not questions, a two-level index called Shadow and a three-phase query refinement approach based on Shadow are proposed. The first phase is to generate some promising refined queries with different query requirements and filter those unpromising refined queries before executing any promising refined queries. The second phase is to reduce the irrelevant search space in the level 1 of Shadow as much as possible based on the spatial filtering technique, so as to obtain the promising static objects, and to capture promising moving objects in the level 2 of Shadow as fast as possible based on the probability filtering technique. The third phase is to determine which promising refined query will be returned to the user. Finally, a series of experiments are conducted on three datasets to verify the feasibility of our method.

Research Session 25: Recommendation 5
Session Chair: Yang Chen(Fudan University)

1. Sirius: Sequential Recommendation with Feature Augmented Graph Neural Networks
Xinzhou Dong, Beihong Jin, Wei Zhuo, Beibei Li, and Taofeng Xue
Abs. Many practical recommender systems recommend personalized items for different users by mining user-item interaction sequences. The interaction sequences, as a whole, imply the manifold collaborative relations among users and items. Further, from the view of users, the item orders and time intervals between interactions could expose the evolution of user interests, and from the view of items, attributes of the items on interaction sequences may reveal the variation of item popularity. However, most of the existing recommendation models ignore those valuable information, and cannot fully explore the intrinsic implication of interaction sequences. In the paper, we propose a method named Sirius, which develops GNNs(Graph Neural Networks) to model the collaborative relations and capture the dynamics of time and attribute features in sequences. We give the workflow of the Sirius method, and describe the implementations about graph construction, item embedding generation, sequence embedding generation and next-item prediction. Finally, we give an

example of Sirius recommendations, which visually shows the impact of feature information on the recommendation results. At present, Sirius has been adopted by MX Player, one of India's largest streaming platforms, recommending movies for thousands of users.

2. Combining Meta-path Instances into Graphs for Recommendation
Mingda Qian, Bo Li, Xiaoyan Gu, Zhuo Wang, Feifei Dai, and Weiping Wang
Abs. In the recommendation area, the concept of meta-path is famous for inferring explicit and effective relationships between nodes such as users and items. To extract useful information from the instances of meta-paths, existing methods embed meta-path instances separately. However, they ignore the complicated semantics presented by multiple instances. These complicated semantics not only provide additional information but also affect the semantics of single instances. Without considering the complicated semantics, the information extracted from the instances may be incomplete and less effective. To solve the problem, we propose to learn the complicated semantics by combining meta-path instances into layer-wise graphs (instance-graphs) for recommendation. Following the idea, we develop an Instance-Graph based Recommendation method (IGR). IGR combines meta-path instances into layer-wise instance-graphs. Then, the instance-graphs are investigated layer by layer to generate effective embeddings. Finally, these embeddings are discriminatively merged into user/item embeddings to make predictions. Extensive experimental results show that IGR outperforms various state-of-the-arts recommendation methods.

3. GCAN: A Group-wise Collaborative Adversarial Networks for Item Recommendation
and Xuehan Sun
Abs. Recommendation System aims to provide personalized recommendationfor different users. Recently, Generative Adversarial Networks(GANs) based recommendation systems have attracted considerableattention. In previous research, GAN has shown potentialand flexibility to learn latent features of users' preferences. However,GANs are hard to train to converge and waste many processesof fulfilling the empty data, especially when meeting with the datasparsity problem.In this paper, we propose a new group-wise framework,

namelyGroup-wise Collaborative Adversarial Networks (GCAN) to solvethe data sparsity problem and enable GAN to converge faster. Wecombine GAN with traditional collaborative filtering methods togenerate recommendations (named as CAN), and then propose binarymasking and sample shifting to achieve GCAN. Binary maskingseparates binary user-item interaction and abstracts group-wiserelationship from these binary vectors, while sample shifting isdesigned to avoid incorrect learning process. A noise corruptionparameter is then introduced with experiments to show the robustnessof GCAN. We compare GCAN with other baseline methodson YP and SC dataset, where GCAN achieves the state-of-the-artperformances for personalized item recommendation.

Research Session 26: Graph Data 4
Session Chair: Yingxia Shao (Beijing University of Posts and Telecommunications)

1. DMSPool: Dual Multi-Scale Pooling for Graph Representation Learning
Hualei Yu, Chong Luo, Yuntao Du, Hao Cheng, Meng Cao, and Chongjun Wang
Abs. Graph neural networks (GNNs) have recently become a powerful graph representation technique for graph-related tasks in various fields. However, the existing GNN models mainly focus on generalizing convolution and pooling operations in a pre-defined unified architecture, limiting the model's ability to capture meaningful information of different nodes or local structures and deliver sub-optimal performance. Besides, the importance of subgraphs at various levels has not been well-reflected. To address the above challenges, we propose Dual Multi-Scale Pooling (DMSPool), a hierarchical graph representation model, which uses multiple architectures concurrently to integrate graph convolution and pooling modules in an end-to-end fashion. Specifically, these modules adopt multiple GNN architectures to learn node-level embeddings and nodes' importance from different aggregation iterations. Additionally, we employ attention mechanism to determine the contribution of subgraphs' representations at varying levels to graph classification and integrate them to perform the cross-scale graph level representation. Experiment results on five widely used benchmarks show that DMSPool achieves superior graph classification performance over the state-of-the-art graph representation learning methods.

2. A Parameter-free Approach for Lossless Streaming Graph Summarization

Ziyi Ma, Jianye Yang, Kenli Li, Xu Zhou, Yuling Liu, and Yikun Hu

Abs. In rapid and massive graph streams, it is often impractical to store and process the entire graph. Lossless graph summarization as a compression technique can provide a succinct graph representation without losing information. However, the problem of lossless streaming graph summarization is computationally and technically challenging. Although the state-of-the-art method performs well with respect to efficiency, its summarization quality is usually unstable and unsatisfactory. This is because it is a randomized algorithm and depends heavily on the pre-tuned parameters. In this paper, we propose a parameter-free lossless streaming graph summarization algorithm. As the graph changes over time, we incrementally maintain the summarization result, by carefully exploring the influenced subgraph, which is shown to be a bounded neighborhood of the inserted edge. To enhance the performance of our method, we further propose two optimization techniques regarding candidate supernodes refinement and destination supernode selection. The experiment results demonstrate that the proposed methods outperform the state-of-the-art by a large margin in terms of compression quality with comparable running time on the majority of datasets.

3. Expanding Semantic Knowledge for Zero-shot Graph Embedding

Zheng Wang, Ruihang Shao, Changping Wang, Changjun Hu, Chaokun Wang, and Zhiguo Gong

Abs. Zero-shot graph embedding is a major challenge for supervised graph learning. Although a recent method RECT has shown promising performance, its working mechanisms are not clear and still needs lots of training data. In this paper, we give deep insights into RECT, and address its fundamental limits. We show that its core part is a GNN prototypical model in which a class prototype is described by its mean feature vector. As such, RECT maps nodes from the raw-input feature space into an intermediate-level semantic space that connects the raw-input features to both seen and unseen classes. This mechanism makes RECT work well on both seen and unseen classes, which however also reduces the discrimination. To realize its full potentials, we propose two label expansion strategies. Specifically, besides expanding the labeled node set of seen classes,

we can also expand that of unseen classes. Experiments on real-world datasets validate the superiority of our methods.

4. A Semi-supervised Framework with Efficient Feature Extraction and Network Alignment for User Identity Linkage
Zehua Hu, Jiahai Wang, Siyuan Chen, and Xin Du
Abs. Nowadays, people tend to join multiple social networks to enjoy different kinds of services. User identity linkage across social networks is of great importance to cross-domain recommendation, network fusion, criminal behaviour detection, etc. Because of the high cost of manually labeled identity linkages, the semi-supervised methods attract more attention from researchers. Different from previous methods linking identities at the pair-wise sample level, some semi-supervised methods view all identities in a social network as a whole, and align different networks at the distribution level. Sufficient experiments show that these distribution-level methods significantly outperform the sample-level methods. However, they still face challenges in extracting features and processing sample-level information. This paper proposes a novel semi-supervised framework with efficient feature extraction and network alignment for user identity linkage. The feature extraction model learns node embeddings from the topology space and feature space simultaneously with the help of dynamic hypergraph neural network. Then, these node embeddings are fed to the network alignment model, a Wasserstein generative adversarial network with a new sampling strategy, to produce candidate identity pairs. The proposed framework is evaluated on real social network data, and the results demonstrate its superiority over the state-of-the-art methods.

Research Session 27: Big Data 2
Session Chair: Ya-Wen Teng (Academia Sinica)

1. Dirty-Data Impacts on Regression: an Experimental Evaluation
Zhixin Qi, and Hongzhi Wang
Abs. Data quality issues have attracted widespread attentions due to the negative impacts of dirty data on regression model results. The relationship between data quality and the accuracy of results could be applied on the selection of appropriate regression model with the consideration of data quality and the determination of

data share to clean. However, rare research has focused on exploring such relationship. Motivated by this, we design a generalized framework to evaluate dirty-data impacts on models. Using the framework, we conduct an experimental evaluation for the effects of missing, inconsistent, and conflicting data on regression models. Based on the experimental findings, we provide guidelines for regression model selection and data cleaning. We believe that our guidelines will be useful for regression tasks.

2. UniTest: A Univeral Testing Framework for Database Management Systems
Gengyuan Shi, Chaokun Wang, Bingyang Huang, Hao Feng, and Binbin Wang
Abs. With the continuous development of data collection, network transmission, and data storage, Big Data are now rapidly expanding in all science and engineering domains. Considering the characteristics of Big Data including quick generation, large size, and diverse data models, higher requirements are placed on the functionality and performance of database management systems. Therefore, it is essential for users to choose a stable and reliable database management system. However, finding the best way to evaluate the reliability and stability of database management systems is still a huge challenge, and it is difficult for users to design their own test cases for evaluating these systems. In order to address this problem, we carefully design a universal testing framework, called UniTest, which can perform effective functional testing and performance testing for different types of database management systems. Extensive testing experiments in multiple types of database management systems show the universality and efficiency of our framework.

3. Towards Generating Hi Fi Databases
Anupam Sanghi, Rajkumar S, and Jayant Haritsa
Abs. Generating synthetic databases that capture essential data characteristics of client databases is a common requirement for database vendors. We recently proposed Hydra, a workload-aware and scale-free data regenerator that provides statistical fidelity on the volumetric similarity metric. A limitation, however, is that it suffers poor accuracy on unseen queries. In this paper, we present HF-Hydra (HiFi-Hydra), which extends Hydra to provide better support to unseen queries through (a) careful choices among the candidate synthetic databases and (b)

incorporation of metadata constraints. Our experimental study validates the improved fidelity and efficiency of HF-Hydra.

4. Modelling Entity Integrity for Semi-structured Big Data
Ilya Litvinenko, Ziheng Wei, and Sebastian Link
Abs. We propose a data model for investigating constraints that enforce the entity integrity of semi-structured big data. Particular support is given for the volume, variety, and veracity dimensions of big data.

Research Session 28: Query Processing
Session Chair: Ladjel Bellatreche (ISAE-ENSMA, Poitiers, France)

1. Accurate Cardinality Estimation of Co-occurring Words Using Suffix Trees
Jens Willkomm, Martin Schäler, and Klemens Böhm
Abs. Estimating the cost of a query plan is one of the hardest problems in query optimization. This includes cardinality estimates of string search patterns, of multi-word strings like phrases or text snippets in particular. At first sight, suffix trees address this problem. To curb the memory usage of a suffix tree, one often prunes the tree to a certain depth. But this pruning method "takes away" more information from long strings than from short ones. This problem is particularly severe with sets of long strings, the setting studied here. In this article, we propose respective pruning techniques. Our approaches remove characters with low information value. The various variants determine a character's information value in different ways, e.g., by using conditional entropy with respect to previous characters in the string. Our experiments show that, in contrast to the well-known pruned suffix tree, our technique provides significantly better estimations when the tree size is reduced by 60% or less. Due to the redundancy of natural language, our pruning techniques yield hardly any error for tree-size reductions of up to 50%.

2. DBL: Efficient Reachability Queries on dynamic Graphs
Qiuyi Lyu, Yuchen Li, Bingsheng He, and Bin Gong
Abs. Reachability query is a fundamental problem on graphs,which has been extensively studied in academia and industry. Sincegraphs are subject to frequent updates in many applications, it is essential to support efficient graph updates while

offering good performancein reachability queries.Existing solutions compress the original graph with the Directed AcyclicGraph (DAG) and propose efficient query processing and index updatetechniques. However, they focus on optimizing the scenarios wherethe Strong Connected Components (SCCs) remain unchanged and haveoverlooked the prohibitively high cost of the DAG maintenance when SCCsare updated. In this paper, we propose the DBL framework, an efficientDAG-free index to support the reachability query on dynamic graphswith insertion-only updates. DBL builds on two complementary indexes:Dynamic Landmark (DL) label and Bidirectional Leaf (BL) label. Theformer leverages landmark nodes to quickly determine reachable pairswhereas the latter prunes unreachable pairs by indexing the leaf nodesin the graph. We evaluate DBL against the state-of-the-art approacheson dynamic reachability index with extensive experiments on real-worlddatasets. The results have demonstrated that our sequential implementation achieves orders of magnitude speedup in terms of index update,while still producing competitive query efficiency.

3. Towards Expectation-Maximization by SQL in RDBMS

Kangfei Zhao, Jeffrey Xu Yu, Yu Rong, Ming Liao, and Junzhou Huang

Abs. Integrating machine learning techniques into RDBMSs is an important task since many real applications require modeling (e.g., business intelligence, strategic analysis) as well as querying data in RDBMSs. Without integration, it needs to export the data from RDBMSs to build a model using specialized machine learning toolkits and frameworks, and import the model trained back to RDBMSs for further querying. Such a process is not desirable since it is time-consuming and needs to repeat when data is changed. In this paper, we provide an SQL solution that has the potential to support diff erent machine learning models in RDBMSs.We study how to support unsupervised probabilistic modeling, that has a wide range of applications in clustering, density estimation, and data summarization, and focus on Expectation-Maximization (EM) algorithms, which is a general technique for finding maximum likelihood estimators. To train a model by EM, it needs to update the model parameters by an E-step and an M-step in a while-loop iteratively until it converges to a level controled by some thresholds or repeats a certain number of iterations. To support EM in RDBMSs, we show our solutions to the matrix/vectors representations in RDBMSs, the relational algebra operations to support the linear algebra operations required by EM, parameters

update by relational algebra, and the support of a while-loop by SQL recursion. It is important to note that the SQL'99 recursion cannot be used to handle such a while-loop since the M-step is nonmonotonic. In addition, with a model trained by an EM algorithm, we further design an automatic in-database model maintenance mechanism to maintain the model when the underlying training data changes. We have conducted experimental studies and will report our findings in this paper.

4. MLSH: Mixed Hash Function Family for Approximate Nearest Neighbor Search in Multiple Fractional Metrics
Kejing Lu, and Mineichi Kudo
Abs. In recent years, the approximate nearest neighbor search in multiple l_p metrics (MMS) has gained much attention owing to its wide applications. Currently, LazyLSH, the state-of-the-art method only supports limited values of p and cannot always achieve high accuracy. In this paper, we design a mixed hash function family consisting of two types of functions generated in different metric spaces to solve MMS problems. In order to make the mixed hash function family work properly, we also design a novel searching strategy to ensure a theoretical guarantee on the query accuracy. Based on the given scenario, MLSH constructs the corresponding mixed hash function family automatically by determining the proportions of two types of hash functions. Experimental results show that MLSH can meet the different user-specified recall rates and outperforms other state-of-the-art methods on various datasets.

Demo Session 1: Demo 1

1. FedTopK:Top-K Queries Optimization over Federated RDF System
Ningchao Ge, Zheng Qin, Peng Peng, and Lei Zou
Abs. Recently, how to evaluate SPARQL queries over federated RDF systems has become a hot research topic. However, most existing studies mainly focus on implementing and optimizing the basic queries over federated SPARQL systems, and few of them discuss top-k queries. To remedy this defect, this demo designs a system named *FedTopK* that can support top-k queries over federated RDF systems. FedTopK employs a cost-based optimal query plan generation algorithm and a query plan execution optimization strategy to minimize the top-k query cost. In addition, FedTopK uses a query decomposition optimization scheme which

allow merge triple patterns with the same multi-sources into one subquery to reduce the remote access times. Experimental studies over real federated RDF datasets show that the demo is efficient.

2. Shopping around: CoSurvey helps you make a wise choice
Qinhui Chen, Liping Hua, Junjie Wei, Hui Zhao, and Gang Zhao
Abs. When shopping online, customers usually compare commodities with each other before making their purchase decision. In addition to the product price, they also concern the word-of-mouth. However, marketing strategies from various e-commerce platforms, along with the diverse online commodities, make it difficult for customers to distinguish the most cost-effective products. Present cross-platform commodity comparison applications merely focus on product prices, without jointly concerning the reviews. In this demonstration, we developed a web-based application, CoSurvey, which matches commodities from various e-commerce platforms and analyzes product comment sentiment on the base of the proposed Attention-BiLSTM-CNN Model. The model uses an attention-based Bi-LSTM network to learn sentence sequence information, uses a CNN to learn sentence structure information, and uses a multilayer perceptron (MLP) to learn meta-information. The meta-information in the comment sentiment analysis task includes comment's like number, reviewer level, additional image, deliver time, and sentence length. Besides the keyword query, CoSurvey provides customers a survey of cross-platform products price changing trends and comment sentiment evolutions. The high concurrency requirements and load balance are also concerned.

3. SQL-Middleware: Enabling the Blockchain with SQL
Xing Tong, Haibo Tang, Nan Jiang, Wei Fan, Yichen Gao, Sijia Deng, Zhao Zhang, Cheqing Jin, Yingjie Yang, and Gang Qin
Abs. With the development of blockchain, blockchain has a broad prospect as a new type of data management system. However, limited to the data modeling method of blockchain, the usability of blockchain is restricted; In addition, every blockchain system has its own native but naive interfaces, when developing based on the different blockchain systems, which will leads to low development efficiency and high development costs. In this study, we construct a SQL-Middleware for blockchain system to solve these problems. The

SQL-Middleware first performs relational modeling of blockchain data, mapping the blockchain data into a relational table; On the basis of modeling the blockchain data, SQL-Middleware encapsulates a set of SQL interfaces for blockchain system, thus realizing the unification of interface access methods of different blockchain systems. At last, we implement the SQL-Middleware based on the open source blockchain system CITA. Demonstration shows that the SQL-Middleware greatly improves the data management capabilities of blockchain and simplifies the blockchain access steps.

4. NRCP-Miner: Towards the Discovery of Non-Redundant Co-location Patterns
Xuguang Bao, Jinjie Lu, Tianlong Gu, Liang Chang, and Wang Lizhen
Abs. Co-location pattern mining, which refers to discovering neighboring spatial features in geographic space, is an interesting and important task in spatial data mining. However, in practice, the usefulness of prevalent (interesting) co-location patterns generated by traditional frameworks is strongly limited by their huge amount, which may affect the user's following decisions. To address this issue, in this demonstration, we present a novel schema, named NRCP-Miner, aiming at the redundancy reduction for prevalent co-location patterns, i.e., discovering non-redundant co-location patterns by utilizing the spatial distribution information of co-location instances. NRCP-Miner can effectively remove the redundant patterns contained in prevalent co-location patterns, thus furtherly assists the user to make the following decisions. We evaluated the efficiency of NRCP-Miner compared with related state-of-the-art approaches.

Demo Session 2: Demo 2

1. ARCA: A Tool for Area Calculation based on GPS Data
SuJing Song, Jie Sun, and Jianqiu Xu
Abs. This paper develops a tool to efficiently and effectively calculate agricultural machinery's working area based on farming machinery's GPS data. The tool works as follows. Firstly, the paper pre-processes GPS data, including removing duplicate data, abnormal data, and invalid data. Data projection is implemented using Gauss-Kruger. The minimum value after projection is selected for data transformed and shifted. Secondly, this paper operates farming machinery trajectory fitting. Finally, an algorithm to form the farming machinery's area based on trajectory data

produced in the first two steps. The algorithm proposed in this paper can achieve an error rate of 0.29 percent, and it only takes 0.03 seconds to process about 60 pieces of data generated in one minute.

2. LSTM Based Sentiment Analysis for Cryptocurrency Prediction

Xin Huang, Wenbin Zhang, Yiyi Huang, Xuejiao Tang, Mingli Zhang, Jayachander Surbiryala, Vasileios Iosifidis, Zhen Liu, and Ji Zhang

Abs. Recent studies in big data analytics and natural language processing develop automatic techniques in analyzing sentiment in the social media information. In addition, the growing user base of social media and the high volume of posts also provide valuable sentiment information to predict the price fluctuation of the cryptocurrency. This research is directed to predicting the volatile price movement of cryptocurrency by analyzing the sentiment in social media and finding the correlation between them. While previous work has been developed to analyze sentiment in English social media posts, we propose a method to identify the sentiment of the Chinese social media posts from the most popular Chinese social media platform Sina-Weibo. We develop the pipeline to capture Weibo posts, describe the creation of the crypto-specific sentiment dictionary, and propose a long short-term memory (LSTM) based recurrent neural network along with the historical cryptocurrency price movement to predict the price trend for future time frames. The conducted experiments demonstrate the proposed approach outperforms the state of the art auto regressive based model by 18.5% in precision and 15.4% in recall.

3. IntRoute: An Integer Programming based Approach for Best Bus Route Discovery

Chang-Wei Sung, Xinghao Yang, Chung-Shou Liao, and Wei Liu

Abs. An efficient data-driven public transportation system can improve urban potency. In this research, we propose IntRoute, an Integer Programming (IP) based approach to optimize bus route planning. Specifically, IntRoute first contracts bus stops via clustering and then derives a new bus route via a mixed integer linear program (ILP). This two-phase strategy brings three major merits, i.e., a single bus route without any transfer, the minimal total time consuming, and an efficient optimization algorithm for large-scale problems. Experimental results show that our IntRoute reduces the traditional commuting time in Sydney from 31.53 minutes

down to 18.06 minutes on average. The full draft of this research can be found from here.

4. Loupe: A Visualization Tool for High-level Execution Plans in SystemDS
Zhizhen Xu, Zihao Chen, and Chen Xu
Abs. The declarative programming language in SystemDS simplifies users to implement machine learning algorithms. It is able to generate execution jobs on different data processing engines including MapReduce and Spark. The GUI in data processing engines typical visualizes the low-level execution process (e.g., RDD transformation in Spark). However, the low-level description in Spark GUI does not show the relationship between DML operations and RDD primitives. In this work, we propose Loupe, a tool to visualize high-level execution plans in SystemDS to ease users to understand the execution process. This paper introduces the design of the tool and demonstrates a visualization case.

Industry Session 1: Industry 1

1. LinkLouvain: Link-Aware A/B Testing and Its Application on Online Marketing Campaign (Industry Track)
Tianchi Cai, Daxi Cheng, Chen Liang, ziqi liu, Lihong Gu, Huizhi Xie, Zhiqiang Zhang, Xiaodong Zeng, and Jinjie Gu
Abs. A lot of online marketing campaigns aim to promote user interaction. The average treatment effect (ATE) of campaign strategies need to be monitored throughout the campaign. A/B testing is usually conducted for such needs, whereas the existence of user interaction can introduce interference to normal A/B testing. With the help of link prediction, we design a network A/B testing method LinkLouvain to minimize graph interference and it gives an accurate and sound estimate of the campaign's ATE. In this paper, we analyze the network A/B testing problem under a real-world online marketing campaign, describe our proposed LinkLouvain method, and evaluate it on real-world data. Our method achieves significant performance compared with others and is deployed in the online marketing campaign.

2. An Enhanced Convolutional Inference Model with Distillation for Retrieval-based QA

Shuangyong Song, Chao Wang, Xiao Pu, Zehui Wang, and Huan Chen
Abs. A common solution of automatic question-answering (QA) systems is retrieving the most similar question for a given user query from a QA knowledge base. Even though some models have got promising performance on this task, it may be hard for them to achieve a balance between accuracy and efficiency. In this paper, we propose an enhanced convolutional inference model with StructBert distillation, called StructBert-ECIM, to achieve such balance.

3. Graph Attention Networks for New Product Sales Forecasting in E-Commerce
Chuanyu Xu, Xiuchong Wang, Binbin Hu, Da Zhou, Yu Dong, Chengfu Huo, and Weijun Ren
Abs. Aiming to discover competitive new products, sales forecasting has been playing an increasingly important role in real-world E-Commerce systems. Current methods either only utilize historical sales records with time series based models, or train powerful classifiers (e.g., DNN and GBDT) with subtle feature engineering. Despite effectiveness, they have limited abilities to make prediction for new products due to the sparsity of product-related features. With the observation on real-world data, we find that some additional time series features (e.g., brand and category) implying product characteristics also play vital roles in new product sales forecasting. Hence, we organize them as a new kind of dense feature called CPV (Category-Property-Value) and propose a Time Series aware Heterogeneous Graph (TSHG) to integrate CPVs and products based time series into a unified framework for fine-grained interaction. Furthermore, we propose a novel Graph Attention Networks based new product Sales Forecasting model (GASF) that jointly exploits high-order structure and time series features derived from THSG for new product sales forecasting with graph attention networks. Moreover, a multi trend attention (MTA) mechanism is also proposed to solve temporal shifting and spatial inconsistency between the time series of products and CPVs. Extensive experiments on an industrial dataset and online system demonstrate the effectiveness of our proposed approaches.

4. Constraint-Adaptive Rule Mining in Large Databases
Meng Li, Ya-Lin Zhang, Qitao Shi, Xinxing Yang, Qing Cui, longfei li, and Jun Zhou

Abs. Decision rules are widely used due to their interpretability, efficiency, and stability in various applications, especially for financial tasks, such as fraud detection, loan assessment, and automatic service offering. In many scenarios, it is highly demanded to generate decision rules under some specific constraints. However, the performance, efficiency, and adaptivity of previous methods, which take no consideration of these constraints, is far from satisfactory in these scenarios, especially when the constraints are relatively tight. In this paper, we propose a constraint-adaptive rule mining algorithm named CARM(Constraint Adaptive Rule Mining) to deal with this problem. A novel decision tree model is designed for rule induction. To provide a practical balance between purity and constraint fitness when building the trees, an adaptive criterion is designed and applied to better meet the constraint. Besides, a rule extraction and pruning process is applied to satisfy the constraint and further alleviate the overfitting problem. In addition, to improve the coverage, an iterative covering framework is proposed in this paper. Experiments on both public and business data sets show that the proposed method is able to achieve better performance, competitive efficiency, as well as low rule complexity when comparing with other methods.

Industry Session 2: Industry 2

1. Familia: A Configurable Topic Modeling Framework for Industrial Text Engineering

Di Jiang, Yuanfeng Song, Rongzhong Lian, Siqi Bao, Jinhua Peng, Huang He, Hua Wu, Chen Zhang, and Lei Chen

Abs. In this paper, we propose a configurable topic modeling framework named Familia. Familia supports an important line of topic models that are widely applicable in text engineering scenarios. In order to relieve burdens of software engineers without knowledge of Bayesian networks, Familia is able to conduct automatic parameter inference for a variety of topic models. Simply through changing the data organization of Familia, software engineers are able to easily explore a broad spectrum of existing topic models or even design their own topic models, and find the one that best suits the problem at hand. With its superior extendability, Familia has a novel sampling mechanism that strikes balance between effectiveness and efficiency of parameter inference. Furthermore, Familia is essentially a big topic modeling framework that supports parallel parameter

inference and distributed parameter storage. The utilities and necessity of Familia are demonstrated in real-life industrial applications. Familia would significantly enlarge software engineers' arsenal of topic models and pave the way for utilizing highly customized topic models in real-life problems.

2. Generating Personalized Titles Incorporating Advertisement Profile (Industrial Track)

Jingbing Wang, Zhuolin Hao, Minping Zhou, Jiaze Chen, Hao Zhou, Zhenqiao Song, Jinghao Wang, Jiandong Yang, and Shiguang Ni

Abs. Advertisement (Ad) title plays a significant role in the effectiveness of online commercial advertising. However, it's difficult for most advertisers to think of attractive titles for their products. By mining keywords from current ad material, traditional retrieval methods and neural text generation models have been applied to solve this problem. However, few of them focus on personalized ad titles generation. Ad titles from different advertisers can be very diversified, and there is massive previous advertising data available, which can tell the style, content, and vocabulary of specific advertisers. Based on massive previous advertising data and current ad material, we propose an Ad-Profile-based Title Generation Network (APTGN) to automatically generate personalized titles for ads. The model utilizes massive advertising data and current ad material to construct a profile for each ad, which is further integrated into the generation model to help recognize the preferences of specific ads. Automatic evaluation metrics and online A/B testing both show that our model significantly outperforms all the baselines, increasing the adoption rate of recommendation titles by 27.22%. Through our deployed model, once an advertiser needs to customize an ad title for their products, satisfactory titles can be recommended automatically without bothering to write any words.

3. Parasitic Network: Zero-Shot Relation Extraction for Knowledge Graph Populating (Industrial Track)

Shengbin Jia, Shijia E, Ling Ding, xiaojun chen, LingLing Yao, and Yang Xiang

Abs. The relation tuple is the basic unit of the knowledge graph. Conventional relation extraction methods can only identify limited relation classes and not recognize the unseen relation types that have no pre-labeled training data. In this paper, we explore the zero-shot relation extraction to overcome the challenge. The only requisite information about an unseen type is the label name. We propose a

Parasitic Neural Network (PNN), where unseen types are parasitic on seen types to get automatic annotation and training. The model learns a mapping between the feature representations of text samples and the distributions of unseen types in a shared semantic space. Experiment results show that our model significantly outperforms others on the unseen relation extraction task and achieves effect improvement of more than 20% when there are not any manual annotations or additional resources. This model, with good performance and fast implementation, can support the industrial knowledge graph populating.

4. Transportation Recommendation with Fairness Consideration
Ding Zhou, Hao Liu, Tong Xu, Le Zhang, Rui Zha, and Hui Xiong
Abs. Recent years have witnessed the widespread use of online map services to recommend transportation routes involving multiple transport modes, such as bus, subway, and taxi. However, existing transportation recommendation services mainly focus on improving the overall user click-through rate that is dominated by mainstream user groups, and thus may result in unsatisfactory recommendations for users with diversified travel needs. In other words, different users may receive unequal services. To this end, in this paper, we first identify two types of unfairness in transportation recommendation, (i) the under-estimate unfairness which reflects lower recommendation accuracy (i.e., the quality), and (ii) the under-recommend unfairness which indicates lower recommendation volume (i.e., the quantity) for users who travel in certain regions and during certain time periods. Then, we propose the Fairness-Aware Spatiotemporal Transportation Recommendation (FASTR) framework to mitigate the transportation recommendation bias. In particular, based on a multi-task wide and deep learning model, we propose the dual-focal mechanism for under-estimate mitigation and tailor-designed spatiotemporal fairness metrics and regularizers for under-recommend mitigation. Finally, extensive experiments on two real-world datasets verify the effectiveness of our approach to handle these two types of unfairness.

PhD Consortium Session 1: PhD Consortium

1. Algorithm Fairness through Data Inclusion, Participation, and Reciprocity.
Olalekan J. Akintande

Abs. Learning algorithms have become the basis of decision making and the modern tool of assessment in all spares of human endeavours. Consequently, several competing arguments about the reliability of learning algorithm remain at AI global debate due to concerns about arguable algorithm biases such as data inclusiveness bias, homogeneity assumption in data structuring, coding bias e.t.c, resulting from human imposed bias, and variance among many others. Recent pieces of evidence (computer vision - misclassification of people of colour, face recognition, among many others) have shown that there is indeed a need for concerns. Evidence suggests that algorithm bias typically can be introduced to learning algorithm during the assemblage of a dataset; such as how the data is collected, digitized, structured, adapted, and entered into a database according to human-designed cataloguing criteria. Therefore, addressing algorithm fairness, bias and variance in artificial intelligence imply addressing the training set bias. We propose a framework of data inclusiveness, participation and reciprocity.

2. Performance Issues in Scheduling of Real-Time Transactions
Sarvesh Pandey, and Udai Shanker
Abs. The multi-site real-time transactional data-analysis based applications and the underlying research efforts to improve the performance of such applications have got renewed attention by researchers in the last four years. It reveals that the current scenario possesses numerous unanswered and truly relevant issues and challenges requiring a multi-disciplinary research approach to work on and solve the core database transaction processing related issues. Our focus is to cover most of the issues and challenges with transaction scheduling algorithms in one place to put out the current research status. At a high level, the domains covered are — real-time priority assignment heuristics, real-time concurrency control protocols, and real-time commit processing. The article indeed guides towards the immediate-future directions requiring actions/ efforts by the modern data-driven research community.

3. Semantic Integration of Heterogeneous and Complex Spreadsheet Tables
Sara Bonfitto, and Marco Mesiti
Abs. A great number of companies and institutions use spreadsheets for managing, publishing and sharing their data. Though effective, spreadsheets are mainly designed for being interpreted by humans, and the automatic extraction of their

content and interpretation is a complex task. The task becomes even harder when tables present different kinds of mistakes and their layout is complex. In this paper, we outline the approach that we wish to develop during the PhD for answering the research question "how to semi-automatically extract coherent semantic information from heterogeneous and complex spreadsheets?".

4. Abstract Model for Multi-Model Data

Pavel Čontoš

Abs. In recent years, many so-called multi-model database management systems have emerged, mainly as extensions of the existing single-model systems, regardless they used to be relational or NoSQL. These new database systems make new demands on their users. From the point of view of the conceptual and logical representation, the so far widely used approaches, especially ER and UML, prove not to be sufficient enough in many aspects due to the specific properties of multi-model data. In addition, it is also difficult to query data that is represented in various and often overlapping data models at the logical level.

5. User Preference Translation Model for Next Top-k Items Recommendation with Social Relations

Hao-Shang Ma, and Jen-Wei Huang

Abs. Recommendation systems are used to predict the interests of users through the analysis of historical preferences. Collaborative filtering-based approaches usually ignore the sequential information and sequential recommendation usually focus on the next item prediction. In this work, we would like to determine the next top-k items recommendation problem. We propose User Preference Translation Model (UPTM) with item influence embedding and social relations between users. In addition, we will also solve the cold start problem in UPTM.