PROCEEDINGS

The 14th HKBU-CSD Postgraduate Research Symposium

PG Day 2011

Department of Computer Science
Hong Kong Baptist University
September 1, 2011

The 14th HKBU-CSD Postgraduate Research Symposium (PG Day) Program

Sept. 1th 2011, Thursday		
Time	**Sessions**	
09:00-09:20	**On-site Registration (Room T716)**	
09:20-09:30	***Welcome:*** **Prof. P. C Yuen** ***(Head of Department of Computer Science, HKBU)***	
09:30-11:00	**Session I (Chair: Mr. Li You, T716)**	
	●Discovering Effects of Neighborhood Socioeconomics Status, Service Characteristics on Wait Time in Health Care Ms. Li Tao ●Characterizing the Robustness of Distribution Networks: A Case Study of the U.S. Natural Gas Pipeline Network Mr. Benyun Shi ●Modeling Incentive Strategies for Vaccination Campaigns against Epidemic Spreading Mr. Shang Xia	
11:00-11:15	**Tea Break**	
11:15-12:15	**Session II (Chair: Miss. Li Yuanxi, T716)**	
	●Data Clustering Based on a Unified Similarity Metric of Categorical and Numerical Attributes Ms. Hong Jia ●Conditional Random Fields for Emotion-based Music Recommendation Using Acoustic Features and User Preferences Mr. Jie Deng	
12:15-14:00	**Noon Break**	
14:00-15:30	**Session III_A (Chair: Miss. Li Yuanxi, RRS905)**	**Session IV_A (Chair: Mr. Liu Kai, T909)**
	●Learning Multi-Boosted HMMs for Password Protected Lip Motion Based Speaker Verification Mr. Xin Liu ●Predictors of Cyberbullying among University Students Ms. Wong Yee Man ●Learning RST-invariant Sparse Representation for Pose Editing Mr. Yongquan Lai	●Side Effect Estimation: A Filter Approach to the View Update Problem Mr. Yun Peng ●Gene Based Compression with Heterogeneous Implementation Mr. Zhao Kaiyong ●Speeding up Scoring Module of Mass Spectrometry Based Protein Identification by GPUs Mr. You Li
15:30-15:45	**Tea Break**	
15:45-17:15	**Session III_B (Chair: Mr. Peng Yun, RRS905)**	**Session IV_B (Chair: Mr. Liu Kai, T909)**
	●A Hybrid WLAN Location Estimation System with Center of Gravity Algorithm Selector Mr. Cheng Quan Jia ●PCMLogging: Reducing Transactional Logging Overhead with PCM Mr. Shen Gao	●Stochastic Network Motif Detection in Social Media Mr. Kai Liu ●CYC Based Query Expansion Framework for Effective Image Retrieval Ms. Yuanxi Li ●On Linear Dependency Modeling for Feature Fusion Mr. Ma Jinhua
17:30-18:00	**Best Paper & Best Presentation Awards Announcement (Room RRS905)**	
Closing		

14th PG Day, Student Presentation List

30minute for both presentation and Q&A

				JL	YYT	Clement	NGJ	PCY	WC	YMC	XJ	Chen Li	BC	CHU	CHL	KW	FJ	TAM
Sept. (Thursday)																		
09:30am-12:15am T716																		
Ms. Li Tao	陶麗	PhD		JL*		Clement◊				YMC◊		Chen Li#						
Mr. Benyun Shi	史本云	PhD		JL*		Clement◊			WC#	YMC◊								
Mr. Shang Xia	夏尚	PhD		JL*		Clement◊			WC#	YMC◊								
Ms. Hong Jia	賈紅	Mphil							WC◊	YMC*				CHU#	CHL◊			
Mr. Jie Deng	鄧杰	Mphil				Clement*						Chen Li#		CHU◊	CHL◊			
14:00pm-16:45pm RRS905																		
Mr. Xin Liu	柳欣	PhD								YMC*		Chen Li◊		CHU#	CHL◊			
Ms. Wong Yee Mam	王綺敏	Mphil				Clement◊							BC#		CHL◊			
Mr. Yongquan Lai	賴勇銓	MPhil						PCY*		YMC#					CHL◊			
Mr. Cheng Quan Jia	鄭銓甲	MPhil					JNG *				XJ◊			CHU◊			FJ #	
Mr. Shen Gao	高屾	MPhil					JNG◊				XJ*		BC#	CHU◊				
14:00pm-17:15pm T909																		
Mr. Yun Peng	彭雲	PhD		JL◊							XJ#		BC*					TAM◊
Mr. Zhao Kaiyong	赵开勇	PhD																
Mr. You Li	李由	PhD		JL#			JNG◊				XJ◊			CHU*				
Mr. Kai Liu	劉凱	MPhil		JL#					WC*	YMC◊					CHL◊			
Ms. Yuanxi Li	李原曦	MPhil		JL◊		Clement*			WC◊			Chen Li#						
Mr. Ma Jinhua	马锦华	PhD			YYT#			PCY*		YMC◊					CHL◊			

Table of abstracts

Section I:

Discovering Effects of Neighborhood Socioeconomics Status, Service Characteristics on Wait Time in Health Care

Tao Li

Abstracts: The socioeconomic status (i.e., age, recent immigrant, income) of patients has long been considered as important factors affecting the mortality and morbidity of cardiac diseases, behaviors for pursuing cardiac care services, and thus exerts effects on the regional wait time variations. In another part, prior research shows that the characteristics of existing services (i.e., the service accessibility, the service capacity, and the service outcome) also influence patients' service selection behaviors, therefore affect the wait time indirectly. However, few studies have comprehensively studied how the neighborhood socioeconomic status and service characteristics affect the arrival and wait time variations in the context of cardiac surgery, which is the purpose of this study. Specifically, we hypothesized that (i) the neighborhood socioeconomic status has directly and indirectly associated with the wait time variations in cardiac care, along with (ii) the indirectly impact of service characteristics on wait time via arrival. Based on the secondary got from Statistic Canada, Cardiac Care Network, and Institute for Clinical Evaluative Sciences of Ontario in years of 2005 and 2006, with methods of Partial Least Squares-based Structural Equation Modeling and regression, we found that (i) the socioeconomic status, age and recent immigrant have significant positive impact on arrival, while they have negative impact on wait time; and income has positive effect both on arrival and wait time; (ii) service accessibility and service outcome have significant negative effects on arrival, while there exists significant positive impact from service capacity on arrival; (iii) the variations between age groups, ethnic groups in recent immigrant, income groups and arrival are differently.

Characterizing the Robustness of Distribution Networks: A Case Study of the U.S. Natural Gas Pipeline Network

Shi Benyun

Abstracts: Robustness is one of the fundamental characteristics for a distribution network, which represents its ability of reliably delivering a commodity from suppliers to consumers through intermediaries in face of uncertainties, such as natural disasters, accidents, and terrorist attacks. Due to the persistent recurrence of certain devastating events, it is essential to study mitigation strategies for alleviating the impact of disruption. In this study, we do investigation on mitigation strategies from three aspects, they are: (i) identifying vital nodes for disruption mitigation, (ii) studying the impact of network structure, and (iii) quantifying robustness in terms of commodity flows. First, by doing failure spreading analysis on the U.S. natural gas transmission network, an impact network can be formed to represent to what extent a specific node may be affected by disruptions on other nodes. The vital nodes can then be identified by a hierarchical structure. Official reports verify that the vital nodes we find play important roles during pipeline disruptions. Second, network structure such as edge-node ratio, trophic level is explored to be relevant to network robustness, which may help to guide the construction of transmission pipelines at a system level. Finally, we present a metric, i.e., network entropy, to measure the robustness of commodity flows based on the theory of Ruelle-Bowens random walk. The failure spreading analysis on randomized transmission networks shows the positive correlation between network entropy and the robustness of commodity flows. By doing so, we can control commodity flows in advance so as to mitigate the impacts of various uncertainties.

Modeling Incentive Strategies for Vaccination Campaigns against Epidemic Spreading

Xia Shang

Abstracts: Individual's acceptance of vaccination has critical important impacts on the control of infectious virus spreading. Incentives for vaccination can influence individual's decision making by reducing the payoff of vaccination. In order to examine the efficacy and effectiveness of using incentives to increase the vaccine acceptability, we develop an epidemiological game-theoretic model to characterize individuals' behavior of vaccination decision making. The model is calibrated with epidemiological data collected from 2009 H1N1 epidemic in Hong Kong. The results suggest that incentives could increase the vaccination rates in the target groups and reduce the overall attack rate. In addition, we compare the cost-effectiveness of three incentive strategies targeting different population groups with respect to their social and demographical variations. The simulation results show that incentives targeting transmissible groups would be more effective than targeting infection vulnerable groups and the universal population.

Section II:

Data Clustering Based on a Unified Similarity Metric of Categorical and Numerical Attributes

Jia Hong

Abstracts: Most of the existing clustering approaches concentrate on purely numerical or categorical data only, but not the both. In general, it is a nontrivial task to perform clustering on mixed data composed of numerical and categorical attributes because there exists an awkward gap between the similarity metrics for categorical and numerical data. This paper therefore presents a unified metric for data clustering, in which the attributes are in either one of the three types: numerical, categorical, and their both. We firstly present a general clustering framework based on the concept of object-cluster similarity. Then, a unified metric of object-cluster similarity is presented. Finally, an iterative clustering algorithm is developed, which is directly applicable to the three data types stated above without any adjustment. Experimental results show the efficacy of the proposed approach.

Conditional Random Fields for Emotion-based Music Recommendation Using Acoustic Features and User Preferences

Deng Jie

Abstracts: Emotion-based music recommendation is to recommend music based on the listener&apos;s emotion status in realtime. In this paper, according to the research on psychology and musicology, we propose a Time-Emotion model to represent the change of emotion over time by utilizing Gaussian distribution to simulate emotion decay process. Thus we use Gaussian function to represent time-emotion influence. Through our previous work on representation of music emotion by three dimensional resonance-arousal-valence space, each music belongs to one emotion state. In order to obtain the listener&apos;s current emotion status through his listening history, we apply Conditional Random Field approach to predict his emotion state which owns the largest probability. After we preliminary predicted the listener&apos;s emotion in realtime, emotion-based music recommendation can be constructed by ranking emotional similarity. In order to improve the recommendation performance and flexibility, we adopt reinforcement learning to interact with listeners to obtain emotional satisfactory scores. The initial experiment results show that this time-emotion model and applied emotion prediction approach is able to obtain good prediction accuracy.

Section III_A

Learning Multi-Boosted HMMs for Password Protected Lip Motion Based Speaker Verification

Liu Xin

Abstracts: In this paper, a novel approach, namely Multi- Boosted HMMs is learned for solving password protected lip motion based speaker verification problem. First, we propose a simple but effective lip motion segmentation algorithm to segment the password sequence into a small set of discrete subunits. Then, for each subunit, we integrate HMMs with boosting learning framework associated with the Random Subspace Method (RSM) to discriminatively model the segmental sequence of input subunit, in which the data sharing scheme is adopted to solve the small training sample size problem, such that a more precise decision boundary is formulated for subunit verification. The utilization of RSM sampling the extracted feature set aims to not only avoid the occurrences of overfitting problem when the size of training set is relatively small compared to the high dimensionality of the feature vector but also can effectively reduce the computation time because the high dimensionality always faces the problem of curse of dimensionality. Finally, the whole utterance whether belongs to the target speaker or not is determined via the verification results of all the subunits learned from multiple boosted HMMs. Our experimental results show that the proposed approach yields an improved speaker verification performance in comparison with the state-of-the-art approaches.

Predictors of Cyberbullying among university students

Yee Man Wong

Abstracts: Cyberbullying, a type of bullying behavior, harasses and makes torment to an individual. This paper aims to understand the predictive factors of cyberbullying by focusing on perpetrator perspectives. Social Cognitive Theory serves as guiding framework. Data are collected from 288 university students in Hong Kong. The result shows Social Influence, Internet Self-efficacy, Instrumental Motivations and Cyber Victimziation are the important predictors of Cyberbullying. The paper can serve as a framework of understanding the behavior of cyberbullying. Besides, it can also give some directions and insights to researchers and practitioners to tackle this issue and formulate the prevention strategies.

Learning RST-invariant Sparse Representation for Pose Editing

Ranch Y.Q. Lai

Abstracts: Inverse kinematics is a cardinal component for key-frame animation. Classical approaches that rely on numerical solutions often suffer from the under-determination problem. Existing data-driven approaches effectively address this problem by learning from motion capture data to narrow down the space of plausible solutions. However, when facing a large variety of poses, these approaches may not be applicable in real-time environment due to their high computation complexity. We propose a SParse representation Inverse Kinematics(SPIK) model for editing articulated poses. To handle the spatial transformations of pose data, we introduce a

rotation, scaling and translation (RST-)invariant framework into the model. In training stage, our model learns a sparse code and a transformation operator for each pose together with a dictionary for the whole training set. In the interactive pose synthesis stage, our model can utilize a dictionary that has tens of thousands of atoms. These atoms contain useful information learned from millions of training samples. Experiments show that our model can synthesize realistic and natural poses interactively and thus can effectively speed up the pose editing process.

Section III_B

A Hybrid WLAN Location Estimation System with Center of Gravity Algorithm Selector

Quan Jia Cheng

Abstracts: With the prevalence of mobile Wi-Fi devices and infrastructures, there are growing interests in mobile surveillance and device tracking for better location-aware services in metropolitan areas. With a good location estimation algorithm integrated into a wireless infrastructure, system administrators can closely monitor the network traffic as well as the behavior of the mobile users. The Received Signal Strength(RSS), easily available information from Access Point(AP) Sensors, has become the most popular research approach. However, in reality received signal strength is affected by factors such as obstacles and angle of arrival. There had been proposed estimation systems that adapt to the environment they are in, but such systems are drawn back by their time-intensive training and retraining process. The solution to Signal Strength-based estimation, therefore, is to devise a system that minimizes training while attaining most accuracy. This paper proposes a hybrid location estimation system whose principal estimation method is Center of Gravity(CG). The CG also serves as an algorithm selector that the system can switch to another estimation algorithm if need be. The aim of this system is to reduce the high cost of training and re-calibration but attain an accuracy comparable to the Fingerprinting approach.

PCMLogging: Reducing Transactional Logging Overhead with PCM

Shen Gao

Abstracts: Phase Changing Memory (PCM), as one of the most promising next-generation memory technologies, offers various attractive properties such as non-volatility, bit-alterability, and low idle energy consumption. Recently, PCM has received increasing attention from the database community. In this paper, we present PCMLogging, a novel logging scheme that exploits PCM devices for both update buffering and transactional logging. Specifically, in order to reduce disk IOs, PCMLogging buffers small database updates on PCM devices and further maintains implicit logs in the buffered updates in support of transaction processing. Different from the traditional approach where buffered updates and transactional logs are completely separated, they are integrated in this new scheme. In addition, PCMLogging makes checkpoint unnecessary and simplifies the recovery in that redo operations can be eliminated. Our preliminary experiments show an up to 40% improvement in disk IO performance in comparison with a basic buffering and logging approach.

Section IV_A

Side Effect Estimation: A Filtering Approach to the View Update Problem

Peng Yun

Abstracts: Updates through views have been a classical problem in databases. Unfortunately, existing literature has shown that translation of updates through views is intractable for a large class of view definitions. This paper presents a novel {\em data-oriented} approach for a practical support of view updates. In particular, we propose a summarization of the source database of views to serve as an {\em update filter}. The update filter aims to efficiently reject untranslatable view updates by {\em estimating the side effects} caused by the view updates, {\it i.e.}, avoiding a complete yet potentially costly analysis of view updates. In comparison, the majority of the conventional approaches do not specially capitalize on (source) data in translation of view updates. This paper first shows that the notion of errors needed to be revised to quantify the quality of an update filter, as the true error is defined with possibly infinitely-many insertions. We then present a novel join cardinality summary, namely {\sc JCard}, of the source database derived from an extended dangling tuple graph and an algorithm to estimate side effects caused by join queries. We present an analysis on the estimation errors of {\sc JCard} and propose algorithms to construct accurate cardinality summary. Furthermore, we extend {\sc JCard} to support projections and selections. Our experiments with TPC-H and real-life datasets show that view filters are accurate and improve overall performance of view updates.

Gene Based Compression with Heterogeneous Implementation

Zhao Kaiyong

Abstracts: Data of Gene are increasing very day. BGI has a capacity of sequencing 130 peoples' DNA raw data every day, means the row data of gene information in BGI increasing more than 10T every data. Genomics Institute of sequencing output raw data per day, from 500G to the 10T per day. Mass data storage is a huge challenge. In this paper, we create a new gene based compression algorithm to storage FASTA DNA data. We want our algorithm can be used in different systems. In this paper we will give the heterogeneous implementation. Our algorithm called fazip. The fazip is better than other algorithms, like gzip, zlib, in the compression of gene based raw data.

Speeding up Scoring Module of Mass Spectrometry Based Protein Identification by GPUs

Li You

Abstracts: Database searching is a main method for protein identification in shotgun proteomics, and till now most research effort is dedicated to improve its effectiveness. However, the efficiency of database searching is facing a serious challenge, due to the ever fast increasing of protein and peptide databases resulting from genome translations, enzymatic digestions, and post-translational modifications. On the other hand, as a general-purpose and high performance

parallel hardware, Graphics Processing Units (GPUs) develop continuously and provide another promising platform for parallelizing database searching based protein identification to increase its efficiency. In this paper, we propose to systematically research on speeding up database search engines by GPUs for protein identification. Considering the scoring module is the most time-consuming part, we mainly utilize GPUs to speed it up. We choose two popular scoring method: firstly, SDP based method, which is chosen by X!Tandem, reaches a speedup of thirty to one hundred; secondly, KSDP, which is adopted by pFind, achieves a speedup of five to ten.

Section IV_B

Stochastic Network Motif Detection in Social Media

Liu Kai

Abstracts: Network motifs refer to patterns of interconnections which are found to be over-represented in real networks when compared with random ones. Such basic building blocks can well characterize the structure of complex networks. Extending them to stochastic ones allow more robust motif detection in stochastic networks. Motif analysis, though common in bioinformatics, has only recently been applied to online social media. In this paper, we propose to detect stochastic network motifs in social media with the conjecture that social interactions are of stochastic nature. In particular, we apply a stochastic motif detection algorithm based on the finite mixture model to both synthesized datasets and real on-line datasets to evaluate the effectiveness. Also, we discuss how the obtained stochastic motifs could be interpreted and compared qualitatively of some of the results obtained with some others which are recently reported in the literature.

CYC Based Query Expansion Framework for Effective Image Retrieval

Li Yuanxi

Abstracts: We study several semantic concept-based query expansion and re-ranking scheme and compare different ontology-based expansion methods in image search and retrieval. In particular, we exploit the functions of CYC Knowledge Base for concept expansion. Furthermore, we combine CYC with our image retrieval framework - Pixearch to expand the user's queries and re-rank the searching results. With the visualized baseline results and user's interactive pruning, the image retrieval precision and recall can yield significantly increase. Preliminary experiments have been able to demonstrate that the proposed retrieval mechanism has the potential to outperform unaided approaches and other query expansion methods.

On Linear Dependency Modeling for Feature Fusion

Andy J Ma

Abstracts: This paper addresses the independent assumption issue in fusion process. In the last decade, dependency modeling techniques were developed under a specific distribution of classifiers. This paper proposes a new framework to model the dependency between features without any assumption on feature/classifier distribution. In this paper, we prove that feature dependency can be modeled by a linear combination of the posterior probabilities under some mild assumptions. Based on the linear combination property, two methods, namely Linear Classifier Dependency Modeling (LCDM) and Linear Feature Dependency Modeling (LFDM), are derived and developed for dependency modeling in classifier level and feature level, respectively. The optimal models for LCDM and LFDM are learned by maximizing the margin between the genuine and imposter posterior probabilities. Both synthetic data and real datasets are used for experiments. Experimental results show that LFDM outperforms all existing combination methods.