Research Statement

LIM Ee-Peng
School of Information Systems, Singapore Management University
Tel: (65) 6828-0781; Email: eplim@smu.edu.sg
28 (Day) 12 (Month) 2021 (Year)

Background
My research interests cover data mining, machine learning, and information retrieval. In my research, I would like to develop novel methods combining the above research topics with known theories from social and behaviour science to model, analyse and improve our understanding of people, their languages and communities, knowledge about real world entities and their relationships, and insights about business, social and urban activities and processes using massive amount of user generated structured and unstructured data. With these methods, I am also interested to find new IT-enabled solutions to address business, social and urban problems at scale, as well to improve user experience in these IT solutions through personalization.

As part of my research, I always like to venture beyond standard datasets and problems, as computer science research is most valuable to new unchartered application domains. It is also these new domains where innovative datasets and multidisciplinary research problems can be studied. For example, in the area of human capital management and talent analytics, my research covers the development of machine learning models for analysing job postings data, job seeker behaviour, course selection and performance. The insights gained can benefit individuals, employers, educators and policy makers. In the area of social media analytics, my research focuses on learning Singlish word phrases that allow us to accurately analyse content generated by Singapore users. In the area of health informatics, I collaborate with other SCIS colleagues to develop the state-of-art food image segmentation techniques along with a benchmark food image segmentation dataset.

I have been the lead faculty and Director of the Living Analytics Research Centre (LARC) hosted by the School of Computing and Information Systems since 2011. In 2021, I successfully led a faculty team to secure a one-year research grant from the Singapore’s National Research Foundation to support the transition from LARC to Social Urban Observatory (SUO), a new research centre that facilitates data-driven and AI enabled social and urban research. As part of the transition, a five-year SUO research centre grant proposal has been prepared and submitted.

In the following, I summarize my recent works in a few research areas.

Research Areas

In social media analytics, I cover three research topics, namely:

• Contextual Path Retrieval
  Contextual path retrieval (CPR) refers to the task of finding path(s) between a pair of user-given entities in a knowledge graph to explain the connection between the two
entities as they appear together in a textual description. CPR can be applied to question answering and recommendation applications. For this novel retrieval task, we propose the *Embedding-based Contextual Path Retrieval (ECPR)* framework. ECPR is based on a three-component structure which includes a context encoder and path encoder that encode query context and path respectively, and a path ranker that assigns a ranking score to each candidate path to determine the one which should be the contextual path. For context encoding, we propose two novel context encoding methods, i.e., context-fused entity embeddings and contextualized embeddings. For path encoding, we propose PathVAE, an inductive embedding approach to generate path representations. Finally, we explore two path ranking approaches. In our evaluation, we construct a synthetic dataset from Wikipedia, and two real datasets of Wikinews articles constructed through crowdsourcing. Our experiments show that methods based on ECPR framework outperform baseline methods, and that our two proposed context encoders yield significantly better performance than baselines. We also analyze a few case studies to show the distinct features of ECPR based methods. This work has led to an ACM TOIS publication.

**References:**

- **Singlish Learning by Crowdsourcing:**
  Singlish is an English Creole language used by Singaporeans in non-formal settings including social and online media. In this research, we explore the combination of machine learning and crowdsourcing to recognize Singlish words and determine their word senses. We identify three research subtasks, namely: Singlish sentence classification, Singlish word phrase recognition, and Singlish annotation crowdsourcing.

  For Singlish sentence classification, we first train a SinglishBERT models with original BERT fine-tuned with Singlish sentences. With SinglishBERT embedding representation of a sentence, we then use a classification model to predict the Singlish/non-Singlish label of the sentence. Our proposed Singlish sentence classification model achieves 0.944 F1 score in 5-fold cross validation performed on more than 49000 labelled sentences. This classification model has been developed into an API and integrated with a Singlish checker web demo. Both the API and demo have been licensed to DSO National Labs.

  For Singlish word phrase recognition, we focus on determining Singlish word phrases in Singlish sentences. To achieve this objective, we first aggregate the annotations contributed by different crowdsourcing workers into “ground truth” annotated Singlish word phrases. The Singlish sentences with these “ground truth” annotated Singlish word phrases are then used to train our proposed Singlish phrase recognition model which leverages on BERT embedding generation and a Conditional Random Field (CRF) to determine the start and end of Singlish word phrases. Our experiments have shown that our proposed model achieves about 70% accuracy. This is the first time a phrase recognition model is developed for Singlish word phrases.

  Singlish OK Boh is a Singlish annotation crowdsourcing system developed to demonstrate that Singlish annotation tasks can be designed as crowdsourcing tasks
which can be easily performed by non-expert workers, and that these tasks can be integrated with the Singlish sentence classification and Singlish word phrase recognition. Our research proposes two different crowdsourcing tasks, namely, (i) Singlish word phrase annotation, and (ii) Singlish replacement word annotation. The former requires workers to mark-up Singlish word phrases in Singlish sentences. The latter requires workers to suggest word replacements for Singlish word phrases in Singlish sentences. Correspondingly, the appropriate task user interfaces have been created and tested on non-expert users. Singlish OK Boh has successfully been deployed and consistently collected 1500 task results every week. The Singlish OK Boh technology has been licensed to DSO National Labs.

- **Learning Semantics from Heterogeneous Networks with Multimodal and Positional Attributes**

In this research, we focus on learning the representations of multi-type entities in a heterogeneous network with multimodal (e.g. textual, visual) and positional (e.g. spatial location, sequence order and hierarchy level) attributes. An example of this heterogeneous network is the hierarchy of user interfaces (UI) objects, namely, applications, screens, view class, and other types of design objects. Such a network not only represents how users understand the visual layout of UIs, but also influences how users would interact with applications through these UIs. To model the UI semantics well for different UI annotation, search, and evaluation tasks, we propose the **Heterogeneous Attention-based Multimodal Positional (HAMP)** graph neural network model. HAMP combines graph neural networks with the scaled dot-product attention used in transformers to learn the embeddings of heterogeneous nodes and associated multimodal and positional attributes in a unified manner. HAMP is evaluated with classification and regression tasks conducted on three distinct real-world datasets. Our experiments demonstrate that HAMP significantly out-performs other state-of-the-art models on such tasks. This work has been published at ACM IUI 2022.

**References:**
Gary Ang, Ee-Peng Lim. Learning User Interface Semantics from Heterogeneous Networks with Multimodal and Positional Attributes. ACM Annual Conference on Intelligent User Interfaces (IUI 2022), Helsinki Finland, March 2022.

In **human capital management and talent analytics**, I focus on the following two research problems, namely:

- **Student Resilience and Academic Performance Analytics**

Analysing student resilience is important as research has shown that resilience is related to students’ academic performance and their persistence through academic setbacks. While questionnaires can be conducted to assess student resilience directly, they suffer from human recall errors and deliberate suppression of true responses. In this paper, we propose **ACREA (ACademic RESilience Analytics)** framework which adopts a data driven approach to analyse student resilient behavior with the use of student-course data. ACREA defines academic setbacks experienced by students and measures how well students overcome such setbacks using a quasi-experimental design. By applying ACREA on a real world student-course dataset, we analyse different types of effects on future term and course performance due to earlier setbacks. We found that setbacks in early academic term significantly affect more
subsequent academic results. We also analyse the multiplier and redemption effects due to the resilience-driven behavior. The insights from the analysis contribute to a better understanding of student resilience using their performance after some academic setbacks. When the recovery of post-setback academic performance is not satisfactory, one can consider introducing new measures to strengthen student resilience. Students may also benefit from the findings when they can be better guided to recover from academic setbacks. This work involves a collaboration with Asst Prof Aldy Gunawan (SCIS) and has been published at IEEE TALE 2021.

References:

- Analyzing Workers’ Browsing Behavior & Annotation Quality In Crowdsourcing Tasks
  In this work, we investigate the connection between browsing behavior and task quality of crowdsourcing workers performing annotation tasks that require information judgements. Such information judgements are often required to derive ground truth answers to information retrieval queries. Instead of aggregating annotation results to obtain the ground truth answer for an information retrieval query so as to evaluate a worker's annotation task result against the ground truth answer, we explore the use of workers' browsing behavior to directly determine the quality of their annotations. We hypothesize user attention to be the main factor contributing to annotation quality. To predict annotation quality at the task level, we model two aspects of user attention known as general and semantic user attentions. Each aspect of user attention can be modeled using different types of browsing behavior features. For annotation tasks involving information judgement, we propose the semantic user attention to capture workers’ understanding of task content using task-semantics specific behavior features. We train several prediction models using behavior features corresponding to different aspects of user attention and conduct experiments on a set of annotation tasks performed by 51 Amazon Mechanical Turk workers. We show that the prediction model using both general and semantic attention features can achieve the best performance with nearly 75% accuracy. This result has made the prediction model viable to be incorporated into the crowdsourcing interface to enhance the qualification of workers, selection of annotation results for deriving ground truth labels, and annotation behavior interventions (e.g., suggesting workers to take a break, reminding workers to stay focused, etc.). We also show that 70% prediction accuracy of worker annotations can be achieved as early as the first 12 seconds or 4/6 of the whole browsing behavior trajectory. A paper covering this work has been submitted to a conference. This work is partially supported by the Lee Kong Chian Professorship.

In health analytics, my research has recently focused on the following key research topic.

- Deep Learning-based Image Segmentation: Food image segmentation is a critical and indispensable task for developing health-related applications such as estimating food calories and nutrients. Existing food image segmentation models are underperforming due to two reasons: (1) there is a lack of high quality food image datasets with fine-grained ingredient labels and pixel-wise location masks --- the existing datasets either carry coarse ingredient labels or very few images; and (2) the
complex appearance of food makes it difficult to localize and recognize ingredients in food images, e.g., the ingredients may overlap one another in the same image, and the identical ingredient may appear very differently in different food images. In this work, we build a new food image dataset FoodSeg103 (and its extension FoodSeg154) containing 9,490 images. We annotate these images with 154 ingredient classes and each image has an average of 6 ingredient labels and pixel-wise masks. In addition, we propose a multi-modality pre-training approach called ReLeM that explicitly equips a segmentation model with rich and semantic food knowledge. In experiments, we use three popular semantic segmentation methods (i.e., Dilated Convolution based, Feature Pyramid based, and Vision Transformer based) as baselines, and evaluate them as well as ReLeM on our new datasets. The FoodSeg103 (and its extension FoodSeg154) and the pre-trained models using ReLeM can serve as a benchmark to facilitate future works on fine-grained food image understanding. We make all these datasets and methods publicly available. This work has been conducted in collaboration with Prof Steven Hoi and Asst Prof Sun Qianru and has been published at ACM MM2021. Since its release, the FoodSeg103 dataset has been downloaded by more than 100 researchers for a wide variety of research topics. The food image segmentation model has also been implemented as a research demo. Moving forward, our research team continues to expand the dictionary of ingredient classes and to increase the dataset size.

References:
Wu, Xiongwei and Fu, Xin and Liu, Ying and Lim, Ee-Peng and Hoi, Steven CH and Sun, Qianru. A Large-Scale Benchmark for Food Image Segmentation. ACM international conference on Multimedia (ACM MM 2021), 2021.

Future Research. I shall continue using my past research experience in data integration and data/text mining, as well as my collaboration with domain experts to identify and work on important social and urban research problems. Ideally, I would prefer to focus on both the “science” and “engineering” dimensions of computer science research in my work. With two external grant endings in the 2022, I will also spend more time to ensure that the corresponding projects will eventually all its stated objectives.

Other Selected Publications and Outputs


