Statement of Research Interest

Dongwoo Kim

As a machine learning researcher, I study statistical models to uncover hidden structures of data. Today, a vast amount of information is digitized and stored into various forms. Despite the collaborative effort to organize the information into structuralized forms, a large body of the stored information still remains in unstructured forms such as text and images. I structuralize the unstructured data by inferring hidden structures which cannot be directly observed from the data. The hidden structures of data correspond to abstract concepts, like themes, behavioral or mental states, or explicit substructures such as a common pattern occurred frequently across the data. These structures are used to understand the data themselves as well as improve many other further tasks. In my research approach, I aim to infer the hidden structure of data by 1) extracting a common sub-structures that have been frequently appeared throughout the data 2) defining a latent variable model based on a plausible generative process of the data.

My next research goal is to model the hidden structure of the complex social process by using data which contains various types of human activities. With the emergence of social media and life-logging platforms, we are facing a unique opportunity to observe and quantify the complex social processes at a scale much larger than ever before. Analyzing and understanding this data will change our understanding of the complex social processes that underlie society. More importantly, various machine learning technologies are shaping human life in various direction. For example, recommendation or algorithmic curation systems become an essential part of our daily life. Therefore, understanding how ML technology will change the opinion or behaviour of society will be a critical mission for every machine learning researcher. As a next step, I am planning to uncover the complex relationship between machine learning algorithms and human behaviour patterns beneath the social process.

1 Prediction via Motifs

From DNA sequences to molecular structure, predicting missing or unknown components of structured data has been an important challenge in machine learning. Various approaches have been attempted to address this problem. One notable direction is to model the hidden state of such structured data. For example, the hidden Markov model assumes an underlying state transition to model a sequence, and the recurrent neural network further improves the performance of such approach by placing the hidden state into more high dimensional space.

During the last couple of years, I and my colleagues tried to see the structured prediction problem from a different perspective in the context of sequence prediction problem. We focused on the fact that it is quite common to observe subsequences in a sequence that occurred frequently across the entire dataset. For instance, given a sequence of musical symbols, one can easily find the repeating musical symbols, which sometimes decides the success of the music. Based on this observation, the sequence prediction can be divided into two sub-problems; 1) finding frequent subsequence (motif) and 2) predicting future sequence based on a set of identified motifs. We proposed an approach based on a memoisation to tackle the sequence prediction problem [15]. The proposed approach memorises all possible subsequences from a previous sequence instead of selecting a few representative sub-sequences. From a sequence of observations, we design an algorithm that can memorise the importance of each sub-sequence. It is inevitable that such algorithm suffers from space and computational complexities. The increasing complexities of the proposed approach have been mitigated by an elegant data structure called a suffix tree. Given that most of existing machine learning algorithms are defined with a rigorous mathematical mechanism instead of having a practical computational design, the proposed approach combines both computer science perspective as well as mathematical perspective. Our follow-up study showed an improved efficiency of the memoisation approach when it is combined with a powerful computational framework, a neural network [14].

2 Generative Models with Nonparametric Bayesian

Alongside the prediction of structured dataset, I aim to study the nature of an unstructured data, text corpus. Specifically, during my Ph.D. studies, I aim to build probabilistic topic models based on the Bayesian nonparametric theory. The Bayesian topic models have emerged as an important tool to uncover the underlying semantic patterns, called topics, from a set of unstructured documents without any supervision. One important direction of the topic model involves the Bayesian nonparametric theory. With the model built upon the nonparametric method, the number of parameters of the model grows as more data is observed. As a result, this approach is well suited for the streaming and/or large-scale data. Despite the advantages of the nonparametrics, there has been not much work on the nonparametric topic models because constructing nonparametirc models typically entails defining a complex model construction and solving intractable posterior probability. I tried to reduce the gap between the parametric and nonparametric topic models by developing nonparametric topic models in two different ways: directly extending the existing parametric topic models and building new model construction methods based on the nonparametric theory.

On the parametric side, incorporating metadata of documents into the existing topic models has been widely studied to capture the different aspect of topics. For example, with the time index of documents, one may infer how topics can change over time [5, 4]; with the citation network among documents, one may infer how the topic effects on the citation behavior [13]; and with the authors of documents, one may infer the topic usage of each author [9]. I addressed the problem of constructing a nonparametric topic models with various types of metadata, and proposed several models to infer the latent topics which reflect the underlying structure correlated with the metadata. A selection of my research is described below.

Dirichlet Process with Mixed Random Measure (DP-MRM) A topic is a multinomial distribution over the vocabulary, and typically, it is represented by a set of high probability words. How can we interpret these probability distributions? One possible solution is to tag a topic with metadata such as categories and tages. Ramage et al. proposed labeled-LDA as a tagging process of latent topics. In the labeled-LDA, each word of a document is assigned to one of the document's labels. As a result, each inferred topic is tagged by one of the labels. This mapping improves the interpretability of topics. At the same time, the one-to-one relationship between topics and labels overly restricts the flexibility of the model because the topics of a document are restricted by the given labels of the document.

I directly extended the labeled-LDA to DP-MRM where the number of topics per label is automatically inferred by the nonparametric method [6]. As a result, the model infers an appropriate number of topics per label. For example, the model infers more topics for the label 'sports' than the more specific and narrow label 'soccer'. I further enhance the model by relaxing exchangeability assumption of words to model the local dependencies between words. The final model is applied to multi-labeled images for image segmentation and object labeling problems by modeling both local dependencies between pixels (words) and the global dependencies between labels across the different images (documents). Given a set of images and a list of objects for each image, DP-MRM automatically segments the images and links the segmented parts to the corresponding objects without pixel level supervision.

Hierarchical Dirichlet Scaling Process (HDSP) DP-MRM successfully applied to labeled documents and images, but sometimes, the model is not appropriate for the different types of side information. For example, if authors of a document is used as labels, then the model allocates a unique set of topics per author, which can not model the shared interest across the authors. I relaxed the assumption behind DP-MRM and developed HDSP[9] which allocates the latent correlation between topics and labels and then infers the correlation from a corpus. To incorporate the correlation into the topic model framework, I proposed novel construction method for HDSP. Within HDSP, the first level random measure is constructed through the widely used method, stick breaking process, and the second level random measures are constructed through the normalized gamma process. This construction escapes from the widely used construction method, two levels of stick breaking processes, and yields a highly controllable and flexible model.

HDSP models topics correlated with numerical labels as well as categorical labels. The model was successfully applied to the academic corpus, where the authors are used as labels, and the product review corpus, where the numerical ratings and category of a product are used as labels. Given a product category of a review, the model shows an improved performance on classifying the rating of review.

3 How ML effects Humanity

Machine learning algorithms and systems become an essential part of our daily life. Especially, from Facebook to Netflix, algorithmic curation or recommendation dramatically changed our information consumption patterns. Researchers and practitioners have been concerned about potential problems of such systems, especially, when the systems are designed to bias the preference of users toward specific topics or attitude. In other words, a user may potentially be exposed to information which reinforces the original attitude of the user without having a chance to be exposed to divergent information. The effect so-called 'filter bubble' has its name because the system filters out information that is decided irrelevant to the preference of the user. It is, however, extremely important for recommender or curation systems offering various options to users. For instance, in the context of political news recommendation and curation, providing diverse opinions on a given subject can encourage users having a balanced view on that subject. It will eventually reduce the social conflict between people with different perspectives. From the perspective of information providers such as moviemaker, diversifying user preference will eventually beneficial to content providers as well as content generators, otherwise, the market will be dominated by a small number of major content providers.

While there has been an ongoing discussion on how these systems can potentially bias users toward a specific topic and attitude, the existence of contradictory examples shows that we do not fully understand what is happening with these systems. For example, in [8], the authors found evidence that the movie recommendation system can actually lessen the effect of the filter bubble; the users accepting recommended items consume diverse items than the others. On the other hand, in [7], the authors found an evidence of the filter bubble in Youtube, where users who access an extreme right-wing video are highly likely to be recommended further extreme right content, which then potentially leads to immersion within an ideological bubble.

So far, many solutions that potentially mitigate the filter bubble problem have been focused on how to design systems that can provide diverse information in the context of human-computer interaction. For instance, [2] propose a system which shows different political opinions about a certain topic side-by-side to allow users to be exposed to diverse opinions. However, these interaction design perspective might be just a temporary expedient if the underlying algorithms are still biased toward certain perspectives. For the next couple of years, I aim to provide a theoretical framework of filter-bubble to understand and overcome related-problems. Specifically, I aim to suggest solutions for the following three problems.

Analysis of existing systems As evident from the contradictory results of the filter bubble, we do not thoroughly understand the machine learning systems currently being used. Although there have been several studies which suggest potential solutions of the filter bubble, no one shows clear understanding of existing systems on 1) how to define the filter bubble with respect to a given system 2) how diverse items the system can recommend, and 3) how the level of recommendation diversity is related to the performance of the system. Different systems use different representations of items to be recommended. For instance, one of the most popular recommender system, collaborative filtering based on matrix factorisation, uses latent or low-rank vectors to represent users and items. Given such a system, how one can determine whether there is a filter bubble or not. If recommended items are similar in the vector space, then can we say the system suffers from the filter bubble? Only when this diversity metric is defined, we can measure how diverse items are recommended and how this diversity affects an overall performance of a system. My research aims to analyse the above three criteria with respect to three major classes of recommendation algorithms: 1) low-rank factorisation [10], 2) content-based recommendation [1], 3) contextual bandit approach [3]. The analysis will help researchers have a better understanding of existing systems and will be served as a baseline result of following approaches.

Theoretical framework Similar to the filter bubble which potentially has an enormous impact on society, the fairness of ML algorithms has been a controversial subject since researchers have discovered some algorithms are biased toward sensitive information such as gender and race. While the public and researchers recognise the problem of both filter bubble and fairness, the filter bubble has not been considered as important as the fairness problem since it does not harm users directly, whereas since the fairness directly affects various real-world problems such as examination of parole. As a consequence of relative indifference, researchers have developed different notions of algorithmic fairness [12] which defines the mathematical properties of ML algorithms to achieve a certain level of algorithmic fairness, whereas relatively fewer things are known for the filter bubble to analyse and understand the problem. As the algorithmic fairness has been served as a fundamental role in the fairness research, we need a notion of algorithmic diversity in the recommendation to develop further theories on the problem. Although we would end up that the filter bubble could be unsolvable as what algorithmic fairness found stating that one needs to sacrifice an accuracy for the price of fairness, any finding can be served as a baseline for the further discussion between policymakers and regulators[1]. I aim to define a notion of algorithmic diversity on which we can rigorously analyse and compare the existing systems in terms of their biasness toward a certain topic. Especially, the contextual bandit approaches would be a useful tool to derive a theoretical framework since the algorithm utilises a theoretical analysis of exploration and exploitation trade-off, in which the exploration stage can be understood as a diversity seeking approach.

Strategic recommendation Typical recommendation systems recommend multiple candidate items at a time, while most of the existing recommendation algorithm consider recommending a single item at a time. In this traditional formulation, each recommended item considered independent of each other, and consequently, they compete against each other. The assumption is not necessarily true in general, and moreover the recommended items are mutually beneficial in some cases. For example, think about recommending a list of news articles on a certain topic, where the first article matches the perspective of a given user, and the perspectives are slightly changing as the list goes down. This approach may intrigue

[1] Note that EU recently announced a new regulation in which at least 30 percent of a content provider's library are dedicated to local content. Recommending these items, however, is another matter.

users actively seeking diverse perspectives in the end. The question is then how we can curate a set of items to mitigate the filter bubble while keeping the interest of users. Recently, a structured recommendation has been proposed [11], where a set of related items is recommended as a whole such as a music playlist. Through the structured recommendation, we may curate a set of items with diverse perspectives or attitudes while not having radically different perspectives. This will mitigate the problem with selective exposure.

This research direction is inherently interdisciplinary. Social scientists identify the most vital research questions, while machine learning researchers contribute for developing novel computational tools. This approach has the potential to change our understanding of the complex social processes that underlie society. I am open to collaborate with people from various field, and I will contribute to both machine learning as well as social science.

References

[1] Michael J Pazzani and Daniel Billsus. Content-based recommendation systems. In *The adaptive web*, pages 325–341. Springer, 2007.

[2] Souneil Park, Seungwoo Kang, Sangyoung Chung, and Junehwa Song. Newscube: delivering multiple aspects of news to mitigate media bias. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 443–452. ACM, 2009.

[3] Lihong Li, Wei Chu, John Langford, and Robert E Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th international conference on World wide web*, pages 661–670. ACM, 2010.

[4] **Dongwoo Kim** and Alice Oh. Accounting for data dependencies within a hierarchical dirichlet process mixture model. In *Proceedings of the 20th ACM Conference on Information and Knowledge Management (CIKM)*, 2011.

[5] **Dongwoo Kim** and Alice Oh. Topic chains for understanding a news corpus. In *Proceedings of the 12th International Conference on Intelligent Text Processing and Computational Linguistics (CICLING)*, 2011.

[6] **Dongwoo Kim**, Suin Kim, and Alice Oh. Dirichlet process with mixed random measures: a nonparametric topic model for labeled data. In *Proceedings of the 29th International Conference on Machine Learning (ICML)*, 2012.

[7] Derek O'Callaghan, Derek Greene, Maura Conway, Joe Carthy, and Pádraig Cunningham. The extreme right filter bubble. *arXiv preprint arXiv:1308.6149*, 2013.

[8] Tien T Nguyen, Pik-Mai Hui, F Maxwell Harper, Loren Terveen, and Joseph A Konstan. Exploring the filter bubble: the effect of using recommender systems on content diversity. In *Proceedings of the 23rd international conference on World wide web*, pages 677–686. ACM, 2014.

[9] **Dongwoo Kim** and Alice Oh. Hierarchical dirichlet scaling process. In *Proceedings of the 31th International Conference on Machine Learning (ICML)*, 2014.

[10] **Dongwoo Kim**, Lexing Xie, and Cheng Soon Ong. Probabilistic knowledge graph construction: Compositional and incremental approaches. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, pages 2257–2262. ACM, 2016.

[11] Dawei Chen, **Dongwoo Kim**, Lexing Xie, Minjeong Shin, Aditya Krishna Menon, Cheng Soon Ong, Iman Avazpour, and John Grundy. Pathrec: Visual analysis of travel route recommendations. In *Proceedings of the Eleventh ACM Conference on Recommender Systems*, RecSys '17, pages 364–365, New York, NY, USA, 2017. ACM.

[12] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 797–806. ACM, 2017.

[13] Jooyeon Kim, **Dongwoo Kim**, and Alice Oh. Joint modeling of topics, citations, and topical authority in academic corpora. *Transactions of the Association for Computational Linguistics*, 5:191–204, 2017.

[14] **Dongwoo Kim** Christian Walder. Neural dynamic programming for musical self similarity. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2018.

[15] **Dongwoo Kim** and Christian Walder. Self-bounded prediction suffix tree via approximate string matching. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2018.