Research Statement

Dongwoo Kim

As a computational social science and machine learning researcher, I conduct research to uncover hidden structure of complex social processes. The social process spans from a personal communication to a decision process of political leaders. The computational social science aims to understand and reason about complex social process through right combination of computational tools, appropriate mathematical models, and deep understanding of social phenomena. So far, much of work in computational social science has focused on modelling network structure and its temporal dynamics. However, from official government documents to transcription of verbal communications, many social processes are digitised and stored into textual form. Therefore, understanding textual information is another key component to analyse complex social processes. In particular, I'm interested in modelling textual information to infer hidden structure of a corpus that corresponds to abstract concepts, like themes, behavioural or mental states.

My Ph.D studies addressed a range of problems related to build a probabilistic framework for analysing and understanding textual content. Specifically, my research aimed to build probabilistic topic models based on the Bayesian nonparametric theory. The Bayesian topic models have emerged as an important tool to infer groups of semantically related words, known as topics, from a set of unstructured documents without any supervision. The resultant topics are extremely useful for characterising semantic content as well as a variety of other tasks such as identifying communication patterns, tracking topics across different languages, and identifying thematic communities. One important direction of the topic model involves the Bayesian nonparametric theory. With 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 and allows to construct flexible models. Despite the advantages of the nonparametric, there has not been much work on the nonparametric topic models because constructing nonparametric models typically entails defining a complex model construction and solving intractable posterior probability. My recent work tried to reduce the gap between the parametric and nonparametric topic models by developing nonparametric topic models and applying these models to understand social processes.

Although my recent work has been more focused on modelling textual contents. My long-term research goal is to model the hidden structure of the complex social process by using a corpus that 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. Analysing and understanding this data will change our understanding of the complex human behaviour that underlie society. As a next step, I am planning to use what I have learned through my past research in topic modelling to develop new models to analyse the human behaviour patterns beneath the social process.

1 Modelling Textual Content

Bayesian nonparametric topic models, rooted in the hierarchical Dirichlet processes, generalise the parametric models by placing an infinite dimensional prior over the topic space. The introduction of Bayesian nonparametrics greatly increases the flexibility of the modelling assumptions. For example, we developed a hierarchical topic model [5] that infers an appropriate depth and width of topic hierarchy that cannot be modelled in a parametric way. Despite its potential capabilities, the nonparametric method has not been not widely applied to topic modelling because the model construction is complex, and posterior inference is intractable for many interesting models.

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; with the citation network among documents, one may infer how the topic effects on the citation behaviour; and with the authors of documents, one may infer the topic usage of each author. I addressed the problem of constructing a nonparametric topic models with various types of metadata, and proposed several models to infer the latent topics that reflect the underlying structure correlated with the metadata. A selection of my research is described below.

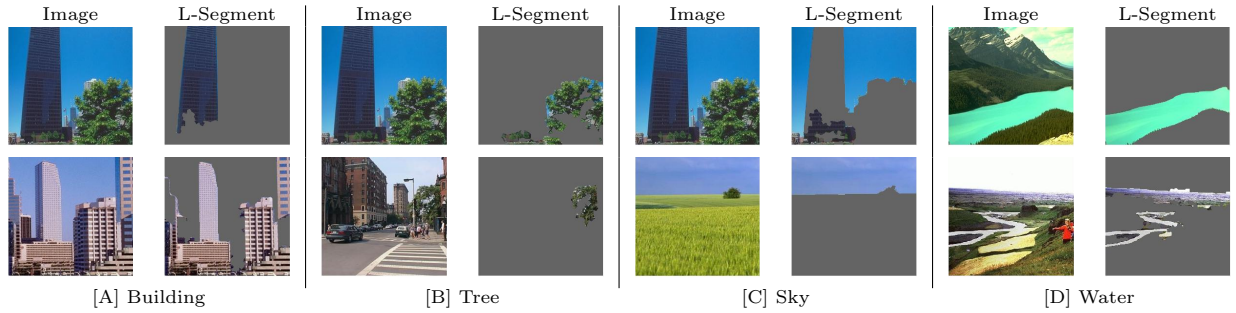

Figure 1: Labeled Segments (L-Segment) from posterior samples of DP-MRM. From left to right, the segments are extracted from *building*, *water*, and *tree* labels. With the list of objects without their location information, DP-MRM segments the images and links the segmented parts to the corresponding objects, simultaneously.

1.1 Nonparametric topic model with metadata

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 tags. 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 labelling problems by modelling 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. Figure 1 shows how the model segments and labels images.

1.2 Modelling correlation among topics and metadata

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 are 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[8] 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 normalised 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

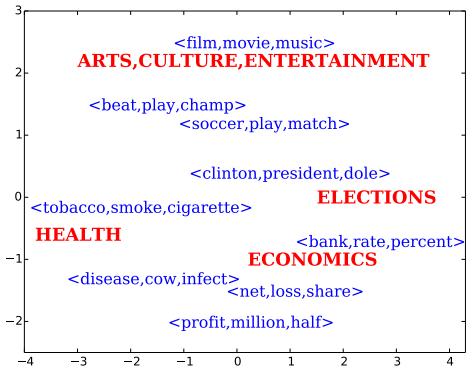

Figure 2: Relative locations of categories (capital letters in red) and latent topics (small letters in blue) inferred by HDSP from the Reuters corpus.

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 predicting the rating of review. The relationship between inferred topics and labels of a news corpus are visualised in Figure 2. HDSP and DP-MRM have been published at ICML 2014 and 2012.

1.3 Temporal dynamics of textual content

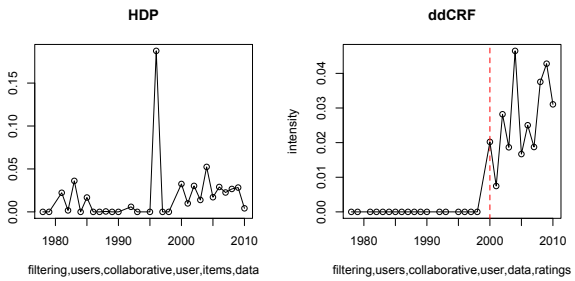

Figure 3: The emergences of 'collaborative filtering' topic inferred from academic corpora by HDP and ddCRF.

A corpus over a long term period may expose the topical characteristics correlated with the written time of each document. With the nonparametric approach, I model the emergence and disappearance of topics over time which cannot be modelled by the parametric approach because, in the parametric models, the number of topics should be fixed a priori. With the preliminary study to capture the emergency and disappearance of topics [4], I develop the distance-dependent Chinese restaurant franchise (ddCRF) model where topics emerge and decay over time [3]. The ddCRF relaxes the exchangeability between documents and allows the topics in nearby documents are more likely to be similar. With a corpus collected over several years, the model captures the emergence of topics over time.

1.4 Conditional generative process of a corpus

The models above, HDSP, DP-MRM, and ddCRF, generate the topic proportions of a document as a dependent variable of observable meta information. This modelling approach differs from the traditional definition of a generative process where the every observable variables are generated from a latent variable or parameter. For example, the supervised-LDA and max-margin LDA propose generative processes where the observable labels are generated from a topic proportion of a document. However, I believe that a more natural model of the human writing process is to decide what to write about (e.g., categories) before writing the content of a document. I proposed nonparametric models based on this perspective on the generative process and showed improved performances on real world datasets.

2 Modelling Human Behaviour from Textual Content

My long-term research aims to infer the hidden structure of complex social behaviour by modelling a corpus that contains various types of social processes and activities. My main research theme has been focused on finding the hidden structure. As the corpus is written/transcribed by many persons, we can infer various types of hidden patterns that reflects the latent characteristics and behavioural patterns of these persons. Moreover, the emergence of social media and online communities give researchers a unique opportunity to observe and quantify complex user behaviour that has not been observed so far.

My early research addressed a range of problems relating to the analysis of textual information used in complex social processes. I developed a new approach to extract interests of Twitter users [2]. Even when users do not explicitly write about certain characteristics, our algorithm can discover them using the Twitter social network information. Although this paper was a workshop paper, it was the first paper to study Twitter lists, an important feature from mining user characteristics.

Our recent work examined how diversity seeking users (e.g. the users who always crave new stories) can benefit the online news outlet which provides the social feed to users [7]. In this work, we propose a metric to measure users' pursuit of diversity, and investigate how the metric can be used to identify influential users who span structural holes in social networks who bridge dense clusters of strong connections and attain the benefits of accessing less-redundant information that pass through them. Also, we look at how diversity-seeking users benefit social news sites by increasing its connectivity; they tend to bridge articles to prevent the information network from being fragmentised. These new approaches try to overcome the limitations

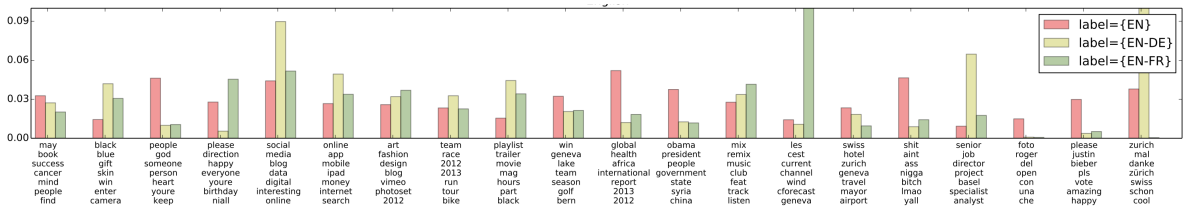

Figure 4: Conversation themes with respect to the monolingual and different types of bilinguals in Switzerland. The results are analysed by HDSP on the Tweets of the Swiss.

of traditional social science research; small scale experiments using controlled study or questionnaire-based methods.

A two-stage approach has been widely applied to analyse the user behaviour; 1) the latent topics of corpus are inferred by relatively simple models, and then 2) the inferred structure is combined with additional information, such as a social network structure, for the further analysis on user behaviour[1]. One advantage of using the Bayesian latent variable models is the model built upon explicit assumptions and prior beliefs on data, and thus, yielding mathematically rigorous and easily interpretable models. The two-stage approach excludes side information from the first stage, therefore the inferred topics or latent variables may not reflect the latent semantics that we would like to know about. One possible direction is to devise a holistic model with various types of information together in order to infer the latent semantics that correspond to the generative process defined by appropriate assumptions about user behaviour. For example, in our ongoing study, we examine the conversation patterns of bilingual and monolingual in multilingual society through HDSP. We assume that the differences in linguistic fluency may result in different conversation patterns. We answer this question by looking at Switzerland, a highly multilingual society, with a large corpus of geotagged Twitter data. Inferred topics from HDSP, as shown in Figure 4, show the topical differences among different linguistic groups.

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 and Bayesian statistics, as well as computational social science.

References

[1] Il-Chul Moon, **Dongwoo Kim**, Yohan Jo, and Alice Oh. Learning influence propagation of personal blogs with content and network analyses. In *2010 IEEE Second International Conference on Social Computing (SocialCom)*, pages 669–674. IEEE, 2010.

[2] **Dongwoo Kim**, Yohan Jo, Il-Chul Moon, and Alice Oh. Analysis of twitter lists as a potential source for discovering latent characteristics of users. In *Workshop on Microblogging at CHI 2010*, 2010.

[3] **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.

[4] **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.

[5] Joon Hee Kim, **Dongwoo Kim**, Suin Kim, and Alice Oh. Modeling topic hierarchies with the recursive chinese restaurant process. In *Proceedings of the 21th ACM Conference on Information and Knowledge Management (CIKM)*, 2012.

[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] Jooyeon Kim, Joon Hee Kim, **Dongwoo Kim**, and Alice Oh. Diversity-seeking users and their influence on social news sites. In *Workshop on Data Science for News Publishing (NewsKDD) at KDD 2014*, 2014.

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