Applying for MCDS Program

11.6.2015

Statement of Purpose

On my career plan, I want to become an outstanding data scientist to be involved in the most fascinating and challenging field of mining the potential values of high volume data. With the experience in this area, I can confidently claim that I have been equipped proper competence and skills to begin and complete the period of Master and future career.

At high-school ages, my extensive involvement in Olympiad of Informatics built up my first concepts of computer programming, algorithm and data structures. Due to excellent academic record of high school and my achievements on competitions. I was fortunately selected(why I chose this program) to become the candidate of Top - notch Program called Fundamental Mathematics and Physics of Tsinghua University (32 out of 30000 applicants), one of top universities in China. In this special class, we accepted most advanced solid mathematics and physics education, then we had chances to select the research area we really have interests in without limitation. (solid foundation for future research) study XXX.

(My research interests on machine learning germinated when XXX.) The turning-point(?) of my research interests occurs in sophomore year when I searched on Amazon to find item I want to buy, I found the items Amazon recommended to me fantastically met my taste. It was amazing because I did’t provide additional information to amazon. I was shocked, and like a kid with a new toy(?), I started investigating more related algorithms and concepts. It was my experience with Machine Learning. It has a wide range of applications in industries and academia where I saw their power.

To broaden my horizon and acquire sufficient knowledge on what the field entails, in fall semester of junior year, I engaged in State Key Laboratory of Intelligence Tech. and System as research assistant. I participated in a project on recommender system, which is the initial spark of my interests in this field. [I spent significant amount of time solidifying my foundation of Machine Learning, probability and statistics through text book, classic papers and online courses.](too long) With the survey on state-of-the-art algorithms on recommender systems, [I found most of these algorithms are based on collaborative filtering methods and matrix factorization such as Latent Factor Model, which is the most classical model in the recommender system domain. These models that evolved from LFM all focus on rated-data, learning a matrix represents(?) features of users and items. However, the relative position of recommended items are not taken into consideration. The items with higher predicted rate will be ranked higher but difference between neighbor items are often tiny. And in realistic, most user will only focus on the top-ranked items.](previous mehtods) Based on this observation, we proposed XXX. (I thought learning to rank algorithms which are utilized on information retrieval domain will gave me a good answer. But how to combine two types of methods and which kind of learning to rank algorithms should I choose? Point-wise, pair-wise or list-wise? When question was clear, there were lots of tough work left.) I have done extensive experiments (result?) and continually modified my framework. Consequently, my ensemble learning model got success on many datasets such as IMDB Movies datasets. The manuscript Ensemble learning methods in recommender system will be submitted on the 16th China Conference of Machine Learning(CCML 16) conference.

Almost one year’s research on recommender system strengthened my skills of solving unknown problems and construct a scientific methodology to handle difficulties. Of course, my knowledge of this area was enriched.

In school, I enrolled amounts of computer science courses which are related to my research. In Data Structure and Algorithm class, I have learnt many advanced structure like AVL tree, red-black tree etc, and many algorithms applied on graph. In Introduction to Machine Learning class, I got extra knowledge of machine learning theories like (VC theory and Graphic Models, and I implemented some algorithms like Hidden Markov Model，Support Vector Machine). Data Mining class told me approaches to do data engineering and how to use machine learning algorithms into realistic issues. I also enrolled in a class for graduate students called Big Data System. In this class, I got a systemic knowledge of doing data mining. I made my hands dirty on coursework and learned how to use distributed systems like Hadoop and Spark; how to solve the issues of load balancing and how to utilize NoSQL database etc. 99 points in final evaluation which is ranked 1st in all class(1/120) certificated my efforts. Beyond classes, I attended Kaggle contests which are held for data scientists particularly. In Click-Through Rate Prediction contest, I experienced whole processing of doing data mining including data cleaning, extracting features, reduce dimensions of data using PCA, finding proper algorithms to train model, model validation and giving final prediction. Fortunately, my submission was ranked top 5% in 1604 submissions from world. In the contest of Diabetic Retinopathy Detection, I used deep learning approach such as Convolutional Neural Network to make detection from thousands image of eyes. This fancy contest let me feel the power of deep learning and my submission was ranked 10% over 661 teams.

In summer of 2015, I went to National University of Singapore (NUS) as summer intern supervised by Prof. Min-Yen Kan. With the dramatically growth of amounts of users which receive online education, the issues related to MOOC has became more and more crucial. This time, I concentrated on the discussion forum and tried to map the term in discussion post with real resource like *lecture, slides, videos,* etc*.* It can significantly reduce the time cost of understanding concepts in discussion forum without extra operation.According to analysis of captured data, I found the frequency of certain term’s occurrence is high and can be divided into two main types, concrete one and implicit one. I built a system with Django framework to resolve the discussion page on *Coursera.* Common regular expression can find most of concrete type terms like *lecture 2-3* with high precision but low recall. And some terms like *algorithms* may not only have a common meaning but also represents a lecture. So I utilized Latent Dirichlet Allocation (LDA), a topic model to generate topics of documents, to figure out which is real representation of terms. Then I developed the application of this system as a Chrome extension. Finally, through mining relationship between real entities and terms, I wanted to build a knowledge base like wikipedia in MOOC platform. As the basis of further research of group, our work will be submitted on Small Interests Group of Information Retrieval (SIGIR) on January.

Back to China, I applied for a research internship position in machine learning group at Microsoft Research Asia (MSRA), after several rounds high selective review, I was granted to become a research intern there. In MSRA, my independent research focus on the embedding represent of documents. Like former successful skip-ngram and word2Vec model, I expected to find a new approach to convert document to vector space with preservation of its intrinsic properties and let machine to learn relationship of concepts in a semantical way. The work is ongoing. My internship career will last to June, 2016. It is really a long period and I definitely think I will get a substantial growth after this work.

Up to now, it has been almost 2 years since I started to touch the field of Machine Learning. It is just a short period but long enough to ensure my determination in the future career direction. MCDS in CMU is the best graduate program of Data Science over the world. (I got to know this program from a Tsinghua alumni and also a former student of MCDS, Ran Chen. He is now at *Trulia* as data scientist and highly appreciates the program.) CMU is my dream school and I believe I can get the best education of Computer Science and Data Science there. For these reasons and my passionate motivation, I am sincerely applying to the Master of Computational Data Science program. I will feel a great honor if you offer me the chance to prove my qualification and I will accept it without any hesitation.