Graduate Statement of Purpose

I will finish my graduate study with a high GPA in Syracuse University and decide to pursue PhD degree in computer and information science. The reason why I make this choice is that I am very interested in data mining and large scale information networks after taking courses CIS 787 Analytical Data Mining taught by professor Reza Zafarani and CIS 675 Design and Analysis of Algorithms taught by professor Sucheta Soundarajan. I desire to extend my knowledge and do some research in the field of large scale information network.

My undergraduate study in Guangdong University of Foreign Studies helped me laid firm foundation for advanced studies in computer science, especially in Natural Language Processing and Data Mining. Natural Language Processing (NLP) is commonly used for text analysis and mining. With the knowledge of NLP, I developed a text search engine finding top-K information related to key words. From the course of data mining, I knew how to collect and preprocess data from raw data and how to use classification, association rule and clustering to mine useful information from data sets. Also, I did a research related to data analysis with my professor Jianxi Liu. And one of my research, Analysis of the Influence of Lane Occupancy on Road Capacity, won the Second Prize in China Undergraduate Mathematical Contest in Modeling. In this research, I applied one-sample t test and computed the solution using Markov transition matrix. At the same time, I wrote a script to process road video data and implemented stochastic simulation and simulation verification in MATLAB. The courses and research experience cultivated my interest in data analysis and laid a solid foundation for my future study.

My graduate study in Syracuse University not only provides me with stronger programming background, but also helps me develop valuable research skills in data analysis and algorithm. During the graduate study, I did a research topic named Exploration and Innovation of Data Processing by MapReduce and used C/C++ to implement the topic into a distribution system. The distribution system with a new scheduling method for workers - master architecture is based on MapReduce structure. The new scheduling method is named mailbox. Master and each worker have its own mailbox. Workers have two functions including map and reduce. The system can be used to find top K frequencies key words from a long paper. Firstly, a master separates a paper into several paragraphs and put paragraphs into its mailbox. Meanwhile, master's mailbox would distribute the data (strings stream) to each worker's mailbox based on a value of a hash function. When a worker receives all data from its mailbox, it would separate the strings into different words and sort (using an algorithm like quicksort) the words based on the alphabet list. After finishing its task, a worker sends a result back to the master's mailbox. Getting a result from a worker, a master uses a hash function to distribute the map results to specified workers working as reducers. The function of reducers mainly counted how many times each word appeared. Finally, a master would use an algorithm like Merge Sort to merge all results from reducers and output the top K frequencies key words. By using knowledge from algorithms and operating system in the research, I realized and questioned myself if I can combine the knowledge from algorithm, data mining and operating system to process large scale or big data information network efficiently based on a distribution system.

Besides improving my programming skill, I also strength my data analysis skill. I learnt lots of useful methods beside the classification, clustering and association rule. One of the useful methods is named bloom filter, a space-efficient probabilistic data structure used to filter email spam. When a stream element containing key X arrives, given a hash function hi(X) and a bucket, if X hashes to a bucket set to 1 for every hash function, we can declare that X is in the S without email spam. Otherwise, we discard the element X. The other useful method is used to solve a problem concerning on finding similar item. If we want to find the similarity of two large documents, firstly, we can use Shingling to get the set of strings of length K that appears in the document. Then we can adopt Min Hashing to find the signatures. The signatures are short integer vectors that represent the sets and reflect their similarity. Finally, we employ Locality Sensitive Hashing to get candidate pairs which we need to use to test for similarity. The data analysis skills I learnt from classes or by myself give me more options and ideas to put forward a new research topic or solve a research problem.

The undergraduate and graduate study experiences have tempered me as a self-reliant person who is always ready to approach problems independently with confidence; cultivated my interest in data analysis and large scale information network; and helped me develop research skills that are essential for success in computer and information science PhD program. I believe my academic background has provided me with a firm foundation for my future success in this academic discipline, and I sincerely wish to be accepted into the PhD program so I can equip myself with crucial skills and knowledge. My aim is to become an expert in the field of large scale information network in the data era. And I hope someday I can use my professional knowledge to provide or capture valuable information from large scale information network efficiently for people.