Graduate Statement of Purpose

After earning my graduate degree with a high GPA from Syracuse University, I will pursue a PhD in computer and information science. I am very interested in data mining and large scale information networks since taking "Analytical Data Mining" with Professor Reza Zafarani and "Design and Analysis of Algorithms" with Professor Sucheta Soundarajan. I am eager to extend my knowledge and do some research in the field of large scale information network.

My undergraduate study at Guangdong University of Foreign Studies helped me establish a 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 my knowledge of NLP, I developed a text search engine finding top-K information related to key words. From my data mining course, I learned 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. In addition, I did research related to data analysis with my professor, Jianxi Liu. And one of my research projects, "Analysis of the Influence of Lane Occupancy on Road Capacity," won Second Prize in the China Undergraduate Mathematical Contest in Modeling. In this research project, I applied a one-sample t test and computed the solution using a 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 prepared me well for my future study.

My graduate study in Syracuse University has not only strengthened my programming background but also helped me develop valuable research skills in data analysis and algorithm. During my graduate study, I completed a research project, "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, with a master architecture based on MapReduce structure. The new scheduling method is named mailbox. Master and each worker have their own mailboxes. 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 my knowledge of algorithms and operating system in the research, I wondered whether I could combine the knowledge from algorithms, data mining and operating system to process large scale or big data information network more efficiently based on a distribution system.

In addition to improving my programming skills, I have also strengthened my data analysis skill. I have learned lots of useful methods beside for the classification, clustering and association rule. One of the more useful methods is the 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 finding a similar item. If we want to find the similarity of two large documents, we can begin by using Shingling to get the set of strings of length K that appear 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 have learned from classes or through independent study have given me more options and ideas to propose new research topics or solve new research problems.

My undergraduate and graduate study experiences have made me a self-reliant person who is always ready to approach problems independently and 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 a computer and information science Ph.D. program. I believe my academic background has provided me with a firm foundation for my future success in this academic discipline, and I sincerely hope to be accepted into the PhD program so that 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. Someday, I will use my professional knowledge to provide or capture valuable information from a large scale information network efficiently for people.