The data collection component of research is common to all fields of study. With the introduction of web 2.0 a massive amount of information is available over the internet which gained attention from researchers. Web crawling is one of the well-accepted and widely used techniques both in academia and industry. Web crawling (a.k.a., web scraping) refers to gathering web pages, by following hyperlinks, starting from a small set of web pages. A program which performs crawling is variably known as a crawler, a spider or simply a bot. Although it seems pretty straightforward, writing an efficient web crawler is not that easy. There are a good number of challenges that should be considered specifically while implementing a large-scale crawler. In this seminar I will discuss these challenges and its potential solutions. Finally, I will demonstrate web-crawling using some exemplar scripts written in JAVA language.

Ms Nargis Pervin is a Ph.D. candidate of the Department of Information Systems in the School of Computing at National University of Singapore (NUS). Her research interests include recommender systems and social network analysis. She has published her work in journals such as ACM MONET and ACM TMIS, and conferences such as ICIS, PACIS, and MobiCASE.