Predicting taxi demand hotspots using automated Internet Search Queries

Disruptions due to special events are a well-known challenge in transport operations, since the transport system is typically designed for habitual demand. Part of the problem relates to the difficulty in collecting comprehensive and reliable information early enough to prepare mitigation measures. A tool that automatically scans the internet for events and predicts their impact would strongly support transport management in many cities in the world. This study addresses the challenges related to retrieving and analyzing web documents about real world events, and using them for demand explanation (if related to a past event) and prediction (if a future one). Transport demand is predicted with a supervised topic modeling algorithm by utilizing information about social events retrieved using various strategies, which made use of search aggregation, natural language processing, and query expansion. It was found that a two-step process produced the highest accuracy for transport demand prediction, where different (but related) queries are used to retrieve an initial set of documents, and then, based on these documents, a final query is constructed that obtains the set of predictive documents. These are then used to model the most discriminating topics related to the transport demand. A framework was proposed that sequentially handles all stages of data gathering, enrichment, and prediction with the intention of generating automated search queries.