Fast Short Text Stream Clustering for Finding Duplicate Question in Stack Overflow

This thesis focuses on the fast clustering of short texts along with an application on finding duplicate questions in Stack Overflow. In the first part of this thesis, we discuss static and dynamic clustering methods for short text corpora. In the second part, we discuss how we can apply static and dynamic clustering of short texts to find duplicate questions in Stack Overflow.

Short text clustering is an important but challenging task. Due to technological advances, short texts are generated in large volumes from different sources, such as micro-blogging, question-answering, and social news aggregation websites. Organizing these texts (e.g., grouping them by topic) is an essential step towards discovering trends (e.g., political, economic) in conversations and other data mining tasks, such as data summarization, frequent pattern analysis, and searching for and filtering information. Clustering the texts into groups of similar texts is the foundation for many of these organizational strategies. What makes clustering (grouping) of short texts challenging is the much lower accuracy in identifying the topic of each text from the few words it contains. To organize the texts properly, we developed novel methods for accurately and efficiently clustering collections of short texts; and efficiently updating the clustering, if the text collection changes over time (e.g., new tweets arrive or new questions posted on a question-answering site).

Among several question-answering sites, Stack Overflow is a popular question-answering site where questions are continuously being posted on various programming problems. Despite detailed guidelines to prevent posting duplicate questions (i.e., questions that have already been answered), duplicate questions are frequently being posted. To handle this problem, Stack Overflow employs users with high reputations to detect duplicate questions, which is a labor-intensive job and may lead to some duplicate questions remaining undetected. An automatic duplicate detection system can alleviate this problem by recommending possible duplicates of a question. Motivated by this we cluster Stack Overflow questions (using our static and dynamic clustering method) to recommend potential duplicates of a newly posted question using the clusters of questions. One of the significant challenges in clustering Stack Overflow questions is that the questions are continuously being posted on different topics over time. Therefore the number of questions to be clustered and the number of clusters to be produced are unknown. This is because we use our dynamic short text clustering method in conjunction with static clustering to cluster Stack Overflow questions to handle the evolving characteristics of clusters.

Experimental studies demonstrate that both our static and dynamic clustering methods of short texts outperform the existing state-of-the-art static and dynamic clustering methods in terms of clustering quality and running time performance on several datasets built on Stack Overflow questions. In addition, using our proposed dynamic clustering method in conjunction with static clustering, we demonstrate that we can find more duplicate questions than an existing duplicate question finding system by searching a limited set of questions.