Abstract

Project Code: TRG5780220

Project Title: Hierarchical Multi-Label Classification for Large Scale Data

Investigator: Asst. Prof. Peerapon Vateekul Ph.D.

E-mail Address: peerapon.v@chula.ac.th

Project Period: May 12, 2014 to Dec 31, 2019

Abstract:

Hierarchical Multi-Label Classification (HMC) is a variant of classification where an example can belong to two or more classes at the same time, and the classes are mutually related by generalization/specialization operations. Recently, the task has gained considerable attention due to the need to organize data with predefined class hierarchy across various domains of application, such as, web repositories, digital libraries, protein function prediction, music genre categorization, etc. In text classification, the size of data is usually very large, particularly on the web, e.g., Wikipedia and Dmoz Open Directory Project. There are not only millions of documents and words (features), but also the number of classes in the hierarchy can be thousands of topics. Unfortunately, most classification algorithms are memory resident and cannot handle very large scale of data. Moreover, analysis under this scenario often shows low prediction accuracy caused by the imbalanced training sets, where one class outnumbers the other.

In this project, we aim to propose an algorithm for HMC, especially for text corpora that are usually large and contain many challenging issues. To achieve best accuracy, there are several approaches investigated in this work: Support Vector Machine (SVM) and Deep Learning techniques. Furthermore, additional preprocessing and postprocessing are also presented in order to improve both accuracy and training time.

Keywords: Hierarchical Multi-Label Classification, Large Scale Data, Support Vector Machines, Deep Learning