E-CLEAN: A DATA CLEANING FRAMEWORK FOR PATIENT DATA

Hasimah Hj Mohamed¹, Tee Leong Kheng², Chee Collin³, Ong Siong Lee⁴

School of Computer Sciences
Universiti Sains Malaysia
11800 USM, Penang, Malaysia

E-mail: ¹hasimah@cs.usm.my, ²teeleongkheng@yahoo.com,
³chee_collin@hotmail.com, ⁴calvin_ong66@hotmail.com

Abstract

We need to prepare quality data by pre-processing the raw data. Data cleaning, also called data cleansing or scrubbing, deals with detecting and removing errors and inconsistencies from data in order to improve the quality of data. Data cleaning system are needed to support any changes in the structure, representation or content of data. There are three parts in the cleaning process, i.e. extract the invalid value, matching attributes with valid values and data cleaning algorithm. Our system uses the extract, transform and load model as the system main process model to serve as a guideline for the implementation of the system. Besides that, parsing techniques is also use for the identification of dirty data. The method that we choose for matching attributes is regular expression. Among those data cleaning algorithms, k-Nearest Neighbor algorithm is selected for the data cleaning part of this project because it is simple to understand and easy to implement.

Keywords: Component, Data cleaning, Regular expression, k-Nearest Neighbor