Journal of Computer Science and Information Technology, Vol. 1, No 1, 2016, pp. 71–81

Detecting Duplicate Entry in Email Field using Alliance Rules-based Algorithm

Arif Hanafi[1*], Sulaiman Harun[2], Sofika Enggari[3], and Larissa Navia Rani[3]

[1]Facultyof Computer Systems & Software Engineering
Universiti Malaysia Pahang,26300 Kuatan,Pahang, Malaysia
[2]Asia Pacific University, Kuala Lumpur, Malaysia
[3]Universitas Putra Indonesia YPTK, Padang, Sumatera Barat, Indonesia

*sulaiman.harun@apu.edu.my

Received 02 February 2016; Revised 19 March 2016; Accepted 20 June 2016

Abstract

The way that email has extraordinary significance in present day business communication is certain. Consistently, a bulk of emails is sent from organizations to clients and suppliers, from representatives to their managers and starting with one colleague then onto the next. In this way there is vast of email in data warehouse. Data cleaning is an activity performed on the data sets of data warehouse to upgrade and keep up the quality and consistency of the data. This paper underlines the issues related with dirty data, detection of duplicatein email column. The paper identifies the strategy of data cleaning from adifferent point of view. It provides an algorithm to the discovery of error and duplicates entries in the data sets of existing data warehouse. The paper characterizes the alliance rules based on the concept of mathematical association rules to determine the duplicate entries in email column in data sets.

Keywords: Datacleaning, Algorithm, Alliance rule, Duplication.

1. Introduction

Email is one of the devices for communication through text. It is evaluated that a normal PC client gets 40 to 50 emails for each day. Numerous applications need take emails as inputs, for instance, email examination, email routing, email separating, email outline, data extraction from email, and newsgroup analysis [1]. Unfortunately, email data can be very noisy. Specifically, it may contain headers, signatures, quotations, and program codes. It also may contain extra line breaks, extra spaces, and special character tokens. It may have spaces and periods inaccurately removed and it may contain words badly cased or non-cased and words misspelled. In order to achieve high quality data mining, it is necessary to conduct data cleaning at the first step [2,3].

Data cleaning is the method of identifying and removing inaccurate records from a record set, table, or database [4]. It is mainly used in databases; the phrase indicates to identifying incomplete, incorrect, inaccurate, irrelevant, and etc. as parts of the data and then replacing, modifying, or deleting this dirty data or unclean data [5,6]. Data cleaning is also called data scrubbing; it is the method of changing or deleting data in data warehouse that is inaccurate, incomplete, inappropriately designed, or duplicated. Organization in a data environment field like insurance, retailing, banking, telecommunications, or transportation might use a data scrubbing tool to analytically study information error by implementing technique, algorithms, and certain data mining rules. Basically, a data cleaning tool includes a framework that were capable of correct a number of specific types of mistakes, such as missing values in database or finding duplicate records. Using a data cleaning technique appropriately will definitely save a database administrator a significant amount of time and can be less costly than fixing errors manually. Data cleaning technique task practice to load in missing values, unified date format, converting nominal