JOURNAL OF COLLEGE OFEDUCATION ___ NO 6.....2010

Basic Steps to Get Data Quality for Data Mining

Dr. ZAKI .S. TOWFIK
Department of Computer Science
Collage of Science
University of AI-Mustansiriya - Iraq
Email- zeik_sead@yahoo.com
Email- zekisaeed@gmail.com

Abstract

The Data extracted from many sources will be integrated and then transform into suitable form. Tthese data may be includes many errors and noise or inconsistencies data. It is necessary to clean the data to get quality data before the data mined from errors and noise data. The cleaning is the first task before any data analysis. The resultant of cleaning analysis/model can be stamped for data quality which very impotent for data minig process because without data quality the algorithms of data nining can not work well or the result of algorithms not good.

Therefor this paper deal with Basics steps to clean data that extracted from many sources to get good quality data for data mining also reduce processing time, storage data and reducing costs and increasing profits, for this case an implementation for data selected from clinical chemical test for yarmook hospital education to detect and remove the errors or noise and or inconsistencies data.

Keywords: Knowledge Discovery, Data Preparation, Data Cleaning is Not Completed Yet,

Basics steps of Data Cleaning

خطوات أساسية لتحسين نوعية البيانات

لغرض تعدين البيانات

الخلاصة:

البيانات المستخرجة من مصادر المختلفة تتكامل وتحول إلى الصيغة المناسبة. هذه البيانات تحتوي على العديد من الأخطاء أو عدم تناسق في البيانات. لذلك من الضروري تنظيف إزالة الأخطاء من البيانات الموجودة للحصول على بيانات عالية جودة قبل إجراء عملية تعدين البيانات. تعدّ عملية التنظيف البيانات من هذه المهمة الأولى لأي تحليل البيانات وقد يترتب العملية تحليل البيانات ووضع نموذج جيد من البيانات. من دون إجراء عملية تنظيف البيانات تكون من الصعب تطبيق خوارزميات تعدين البيانات لان النتايج سوف تكون غير جيدة.

لذلك تم وضع خطوات أساسيات لتنظيف البيانات التي تستخرج من مصادر عديدة للحصول على بيانات ذات نوعية جيدة وذات قيمة عالية في عملية تعدين البيانات والتي تعمل على تخفيض الوقت اللازم عند عمليات تعدين البيانات وتخزين البيانات وتخفيض التكاليف وزيادة الأرباح نتيجة دقة البيانات الناتجة ، مع تنفيذ لبيانات لتنفيذ خطوات تم اختيار بيانات من استمارة الفحوص المختبرية للكيمياء السريرية لمستشفى اليرموك ألتعلمي لكشف الأخطاء وإزالتها أو عدم اتساق البيانات المرضى.

1. Introduction

A process used to determine inaccurate, incomplete, or unreasonable data and then improving the quality through correction of detected errors and omissions. One of important and complex steps in quality data process is the cleaning data step, also called data cleansing or scrubbing. It deals with detecting and removing errors or noise and inconsistencies from data before data reach to data mining in order to improve the quality of data. Data quality problems are present in single data collections, such as files and databases, e.g., due to misspellings during data entry, missing information or other invalid data. When multiple data sources need to be integrated. This is because the sources often contain redundant data in different representations. In order to provide access to accurate and consistent data, consolidation of different data representations and elimination of duplicate information become necessary.

Data Mining require and provide extensive support for data cleaning. They load and continuously refresh huge amounts of data from a variety of sources so the probability that some of the sources contain "dirty data" is high. Furthermore, data mining are used for decision making, so that the correctness of their data is vital to avoid wrong conclusions. For instance, duplicated or missing information will produce incorrect or misleading statistics[1].

Data is not integrated as for data mining but needs to be extracted from multiple sources, transformed and combined during query runtime. The corresponding communication and processing delays can be significant, making it difficult to achieve acceptable response times. The effort needed for data cleaning during extraction and integration will further increase response times but its mandatory to achieve useful query results[2].

Some research groups concentrate on general problems not limited but relevant to data cleaning, such as special data mining