Máster de Formación Permanente
en BIG DATA
e Inteligencia Artificial

Universitat d'Alacant
Universidad de Alicante

Module 5
Analítica de datos y extracción de conocimiento mediante técnicas de IA (Data Analytics)

Task 1: Multidimensional Conceptual Design

Description:

The use case refers to a technological company which gathers data related to its orders and shipments. The company wishes to better understand the evolution of its sales and returns globally. In a more concrete way, the company is absolutely worried with improving its order's preparation time. To this aim, the company is gathering data related to the exact date when (i) the order is received, and (ii) the shipment is ready and placed in the shipment company. Apart from this critical issue, the company would like to better understand its clients, the main ordered products as well as the main returned products.

The collected data is distributed into three main datasets:

1.- Orders.

Basically, an order represents products ordered by clients. The key information gathered about the order of a product is the number of units (quantity), the sales value (sales), discount, profit, and the shipping cost of an order. The type of orders and the priority are features chosen by clients (also provided in the data set).

The main information provided for customers are their names, customer number, type of customer, zip code, city, state, country, region and market to which they belong to. This latter feature has a lot of potential as it will allow us to analyze potential markets among other factors, as well as assess which are our best clients or identify potential clients.

The main information gathered from products are the name of the product and their different categories depending on their types.

Finally, we are also interested in knowing which products have been returned or not for any reason.

2.- Returns

It contains information on which product from which region has been returned.

3.- Markets

This data set provides further information about which country belongs to which Region and to which Market. In this particular case study, this information has been provided in another different data file.

Try to help your company improve its situation by creating a data analytical repository (data warehouse) that contains the information that will allow us to contain the main data that facilitates its ulterior analytical queries. In order to create this data warehouse, this first task is to specify a multidimensional conceptual schema.

Submit a pdf file that contains a conceptual multidimensional (MD) schema following the main notation described in the theoretical lessons to describe the main issues of this conceptual MD schema. You should include (i) **a screenshot of the conceptual MD schema accomplished by any graphical editor (e.g. visio), MS Word, or MS Powerpoint or any other tool**, and (ii) **a description of its main elements.**

Remember that the file submitted must **include the full name(s) of the student(s) involved**.

To do:

Follow the main steps described in the lessons to define a proper conceptual MD schema, including the following elements:

1.- The main analytical process
2.- The main fact or facts
3.- Dimensions
4.- Measures
5.- Classification hierarchies for each dimension

We should provide the conceptual MD schema by following the Directed Acyclic Graphs (DAGs) structures followed in the lessons. Then, textual information should also be provided to specify the main attributes within the facts, dimensions and classification hierarchies.

In order to do so, we should analyze the data files and define all the above-required elements. Any further analysis on the data such as if there are missing values, if they are critical, or if there is anomalous data, should be accomplished by students on hand. Actually, in the further Task 4, we will learn how to accomplish a data profiling in a more automatic way.

Generally speaking, and as a matter of reminding the main goal of a conceptual MD schema, the data analysist should be able to accomplish an easy and in depth analysis about issues such as:

- How much is the average shipping time of packages according to their category?
- How much is the average profit that the company obtains per package category?
- How much are the total losses for each package category?
- How much profit do the 10 best clients generate compared to the rest?
- From which regions are the main orders and returns?
- What are the main types of products ordered?
- What is the main difference between ordered and shipment dates and for which regions?
- What type of products are the most ordered ones?
- How much is the total profit per market?

Please, remember that this MD schema will be later implemented in Task 2, and its ulterior data visualization will be accomplished in further modules (e.g. data visualization).