Họ và tên: Dương Đức Duy

Mã sv: 20020287

**BTVN tuần 1: 10 thuật ngữ quan trọng về KPDL**

1. Data-driven

The term data-driven describes a business state where data is used to power decision-making and other related activities efficiently, in real-time.

*(“What Is Data-Driven”* - Dataversity)

1. Data-warehouse

A data warehouse is a type of data management system that is designed to enable and support business intelligence (BI) activities, especially analytics. Data warehouses are solely intended to perform queries and analysis and often contain large amounts of historical data. The data within a data warehouse is usually derived from a wide range of sources such as application log files and transaction applications.

(“*What is a data warehouse*” - Oracle)

Kho dữ liệu là tập hợp các dữ liệu định hướng theo chủ đề, được tích hợp lại, có tính phiên bản theo thời gian và kiên định được dùng để hỗ trợ việc tạo quyết định quản lý

(“*Giáo trình khai phá dữ liệu 2013*”)

1. Database

A database is a collection of related data. However, the common use of the term database is usually more restricted. A database has the following implicit properties:

■ A database represents some aspect of the real world, sometimes called the

miniworld or the universe of discourse (UoD). Changes to the miniworld

are reflected in the database.

■ A database is a logically coherent collection of data with some inherent

meaning. A random assortment of data cannot correctly be referred to as a

database.

■ A database is designed, built, and populated with data for a specific purpose.

It has an intended group of users and some preconceived applications in

which these users are interested.

(“*Fundamentals of database systems, 7th edition*” -Ramez A. Elmasri, Shamkant Navathe )

Một cơ sở dữ liệu là một tập hợp các dữ liệu có liên quan với nhau, được lưu

trữ trên máy tính, có nhiều người sử dụng và được tổ chức theo một mô hình

(“*Giáo trình nhập môn hệ cơ sở dữ liệu*” – Nguyễn Tuệ)

1. Feature extraction and portability:

The raw data is often in a form that is not suitable for processing. Examples include raw logs, documents, semistructured data, and possibly other forms of heterogeneous data. In such cases, it may be desirable to derive meaningful features from the data. Generally, features with good semantic interpretability are more desirable because they simplify the ability of the analyst to understand intermediate results. Furthermore, they are usually better tied to the goals of the data mining application at hand. In some cases where the data is obtained from multiple sources, it needs to be integrated into a single database for processing. In addition, some algorithms may work only with a specific data type, whereas the data may contain heterogeneous types. In such cases, data type portability becomes important where attributes of one type are transformed to another. This results in a more homogeneous data set that can be processed by existing algorithms.

(“*Data Mining The Textbook*” – Charu C.Aggarwal)

1. Data cleaning:

In the data cleaning phase, missing, erroneous, and inconsistent entries are removed from the data. In addition, some missing entries may also be estimated by a process known as imputation.

(“*Data Mining The Textbook*” – Charu C.Aggarwal)

Kỹ thuật này tìm cách tính toán các giá trị còn thiếu, loại bỏ và làm mịn các giá trị nhiễu trong quá trình xác định đặc trưng, cũng như chỉnh sửa sự nhất quán của dữ liệu

(“*Giáo trình Khai phá dữ liệu 2013*”)

1. Data reduction, selection, and transformation:

In this phase, the size of the data is reduced through data subset selection, feature subset selection, or data transformation. The gains obtained in this phase are twofold. First, when the size of the data is reduced, the algorithms are generally more efficient. Second, if irrelevant features or irrelevant records are removed, the quality of the data mining process is improved. The first goal is achieved by generic sampling and dimensionality reduction techniques. To achieve the second goal, a highly problem-specific approach must be used for feature selection. For example, a feature selection approach that works well for clustering may not work well for classification.

(“*Data Mining The Textbook*” – Charu C.Aggarwal)

1. Classification:

Classification is a data-mining technique that assigns categories to a collection of data to aid in more accurate predictions and analysis.

(“*The Use of Classification in Data Mining*” – Lifewire)

Phân lớp (Classification) thực hiện việc xây dựng (mô tả) các mô hình (hàm) dự báo nhằm mô tả hoặc phát hiện các lớp hoặc khái niệm cho các dự báo tiếp theo.

(“*Giáo trình Khai phá dữ liệu 2013*”)

1. Clustering:

Clustering is the grouping of specific objects based on their characteristics and their similarities. As for data mining, this methodology divides the data that is best suited to the desired analysis using a special join algorithm.

(“*What is Clustering in Data Mining?*” – EDUCBA)

Phân cụm (Clustering) thực hiện việc nhóm dữ liệu thành các “cụm” (có thể coi là các lớp mới) để có thể phát hiện được các mẫu phân bố dữ liệu trong miền ứng dụng. Phân cụm là một bài toán mô tả hướng tới việc nhận biết một tập hữu hạn các cụm hoặc các lớp để mô tả dữ liệu.

(“*Giáo trình Khai phá dữ liệu 2013*”)

1. Regression:

Regression is a statistical method used in finance, investing, and other disciplines that attempts to determine the strength and character of the relationship between one dependent variable (usually denoted by Y) and a series of other variables (known as independent variables)

(“*Regression Definition*” – Brian Beers)

Hồi quy (regression) là một bài toán điển hình trong phân tích thống kê và dự báo, trong đó tiến hành việc dự đoán các giá trị của một hoặc một số biến phụ thuộc vào giá trị của một tập hợp các biến độc lập.

(“*Giáo trình Khai phá dữ liệu 2013*”)

10. Machine learning:

Machine learning is a branch of artificial intelligence (AI) and computer science which focuses on the use of data and algorithms to imitate the way that humans learn, gradually improving its accuracy.

(“*Machine learning*” – IBM)

Machine learning (ML) hay máy học là một nhánh của trí tuệ nhân tạo (AI), nó là một lĩnh vực nghiên cứu cho phép máy tính có khả năng cải thiện chính bản thân chúng dựa trên dữ liệu mẫu (training data) hoặc dựa vào kinh nghiệm (những gì đã được học). Machine learning có thể tự dự đoán hoặc đưa ra quyết định mà không cần được lập trình cụ thể.

(“*Machine learning là gì?*” – TopDev)