LEVEL 0 SUMMARY TEMPLATE

Instruction

This summary will be shared with L1, L2 and L3. Keep in mind that these levels do not have a full understanding of the subject. Try to write something easy to understand but not simplistic. Your summary should explain the main contribution of the paper with your own words. Furthermore, you can use simple examples, if necessary, to better explain the main ideas. Your grade will take into account the quality of your summary, the formal English language in which it has been written, and whether it helps the levels above in their own work.

Name of student: Sanaa Dahour

Name of your Level 1: L0

Source (e.g. scholars.google.com): Google scholars

Paper title: Data Mining: Past, Present and Future

Keywords specific to the paper:

Summary of the main contributions:

(Use text paragraphs, tables and if necessary, figures):

- AI model used (e.g. Neural network, etc.)
- Introduce the AI models
- How do they contribute the idea proposed by the paper?

Supported by a software application? (If yes, provide more details) NO.

The Data mining can be defined as a group of techniques in order to extract or discover the “hidden information” from data, focusing the importance of understanding "hidden" and "information". The word hidden in this definition is important; SQL style querying, however sophisticated, is not data mining. This technique has evolved remarkably since its origins in the late 80s, and gaining recognition by the early 90s as a sub-process of Knowledge Discovery in Databases (KDD). The definition that is most commonly used of KDD is that attributed to Fayyad et al. “The nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data” (Fayyad et al. 1996). Another sub-process in KDD includes data preparation, such as warehousing, data cleaning and pre-processing, and analysis and visualization of results. Although KDD and data mining are often used in practice, as mentioned previously, data mining is technically a specific sub-process within the larger KDD process. Initially, data mining techniques focused mainly on analyzing tabular data, with a strong emphasis on efficiency due to limited processing power. Tabular data is structured data organized into rows and columns, similar to a spreadsheet.

However, as the computing power increased, the focus was targeted on processing larger data sets. There are now several established data mining techniques for tabular data analysis, making it easier for many commercial companies and researchers to perform data mining on standard desktop computers using softwares (SPSS Clementine for ex). The amount of electronic data that’s collected by enterprises and institutions continues to grow heavily, which demands effective mechanisms for mining increasingly large datasets. So we can say that another focus of the data mining is the application of these techniques to non-standard data types, such as images, documents, videos, and network data. The popularity of data mining surged in the 1990s. The rise of its popularity was fueled by technological advancements, enabling the processing of large datasets on desktop machines. Commercial enterprises started maintaining data in computer-readable formats primarily for operational purposes, with data mining as a secondary consideration. The 1990s also saw the use of customer loyalty cards enter, which allowed the businesses to record purchases in order to mine customer behavior patterns. In recent years, there has been a growing emphasis on mining non-standard data types, reflecting the ongoing evolution and expansion of data mining applications.

DATA MINING MECHANISM

There is a distinction between data mining and machine learning. Data mining, is focused on data in all formats, is viewed as an application domain. On the other hand, traditional machine learning is centered on mechanisms for computer learning. So, the difference between data mining and machine learning is that one is seen as a technology, while the other is seen as an application. Data mining techniques are grouped into pattern extraction identification, data clustering or classification categorization. Additionally, current data mining literature incorporates methods borrowed from fields like statistics and mathematics, such as linear regression and principal component analysis (PCA).

PATTERN

Data mining has historically been centered around descovering patterns within data, in relation with customer purchasing behaviors to trends in temporal or graph data. These patterns represent frequently occurring combinations of entities, events, or objects.

CLUSTERING

According to the document, clustering is the process of dividing data into groups of categories, useful for consumer targeting for example. The effectiveness of a cluster is based on inter-cluster cohesion (that is to say how similar is the data), and inter-cluster separation (how distinct it is). We have to know that there's no better or worst algorithm for clustering.

CLASSIFICATION

This is the process of creating classifiers in order to categorize unseen data. This classification can be binary (two options), multi-class (more than two options), or multi-labeled (assigning to one or more classes). Their evaluation is based on the accuracy, sensitivity and specificity.

APPLICATIONS

TEXT MINING

Text mining results from the evolution of tabular data mining, which is focused on analyzing large collections of text such as news articles or web pages. Applications include creating classifiers to categorize or group documents, extracting opinions from questionnaire-style data, and summarizing text. The main challenge of text mining has to see with the effective representation of text data. The common method is the bag-of-words representation, where documents are described by keywords. Keyword selection can be done by experts or extracted using other data mining or natural language processing (NLP) techniques.

IMAGE MINING

Image mining, like text mining, involves the representation of digital image data in order to apply data mining techniques. The success of image segmentation is variable and remains a topic of studies. Challenges remain in image analysis, including the ability to distinguish a cat from a dog for example. Medical image mining has seen some success, for example classifying retinal imaging data and MRI scans.

GRAPH MINING

Graph and tree mining are extensions used in order to find frequent patterns in data, focusing on identifying common sub-graphs. Practitioners suggest that almost anything can be represented as a graph, from documents to chemical compounds. Tree mining is often simpler due to tree properties like no cycles.