A survey for the essential Data Mining technique: Clustering

Isabel B. Amaro[1], **Mirella M. Moro**[2], **Clodoveu Davis**[3]

[1]Department of Computer Science – Federal University of Minas Gerais (UFMG)
Belo Horizonte, Brazil

{isabel.amaro,mirella,clodoveu}@dcc.ufmg.br

Abstract. *Clustering is a Data Mining technique capable of group data with some non trivial similarity. Because of the advance of technology and data generation, clustering algorithms had to be developed and improved to process and extract useful information from the current large amount of data. Therefore, many areas that use Data Mining are constantly publishing articles with new clustering techniques and algorithms. This survey will study some state-of-the-art clustering algorithms, summarize, organize and present the temporal analysis in order to support the studies of its readers.*

1. Introduction

The desire of time and results optimization have been crontributed for the advance of technology in current years. Custom and fast services based on users' caracteristics and informations cause great interest for companies and users. Nevertheless, the search for this custom, pratical and optimized world requires generating large quantity of data.

Data Mining identifies relevant and non trivial patterns in a raw data. To do so, Data Mining uses clustering techiniques, which groups data with some similarity. It makes Data Mining center of most computing researches. Clustering algorithms have been developed and published in articles every year in the past currently years. Because computing is a very dynamic area, it makes the state-of-the-art of clustering hard to understand. Our goal is analyse clustering papers from different areas to help people interested on informations about the state-of-the-art os clustering.

2. Different clustering case of use

The efficient detection of desease-genes is fundamental for prevention and treatment of actual medicine cases. This was one of motivations in (Deseases, Text Mining), which used co-ocurrence clusterings to extract genetic associations in a medical database. Identifying entities in the medical database, (Deseases, Text Mining) could save time and facilitate the early detection of deseases as cancer.

Em (clustering text data streams), foi utilizada a clusterizaÃ§Ã£o de dados textuais utilizando abordagem de arvore, de modo a agrupar dados textuais de forma continua.

Em (Verb Clustering for Brazilian Portuguese), foi buscado o aperfeiÃ§oamento de tÃ©cnicas de processamento de linguagem natural utilizando dois algoritmos de lcustering: spectral clustering e data-cluster-data. O que impulsionou esse projeto foi o fato de que tais metodos apenas foram aplicados em contextos da lingua inglesa, e nunca em portugues.

A agricultura tambem pode ser beneficiada pelas pesquisas relacionadas a clustering, como no caso de [**?**], em que foram utilizados algoritmos de clustering hierarquicos

para determinar quais areas precisam de adubo ou defensivos, de modo a evitar gastos desnecessarios.

3. Data Mining

4. Clustering

5. Algorithms

5.1. K-Means

5.2. Hierarchical clustering

5.3. DBSCAN

6. Temporal analysis

7. Conclusion