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.*

Introduction

The desire of time and results optimization have been crontributed for the advance of technology in current years. Custom services which generates large quantity of data continually.

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 most of 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.

Um dos pontos de interesse principais do avanco da tecnologia e o oferecimento de servicos personalizados baseados nas informacoes adquiridas, podendo essas serem especificas de um usuario, de um texto ou de um banco de dados, por exemplo. Devido a isso, uma grande quantidade de dados vem sendo gerada nos ultimos anos para esse consumo, despertando uma area de interesse da computacao: mineracao de dados.

A mineracao de dados busca identificar padroes relevantes nao triviais em um banco de dados raw. Para isso, a mineracao de dados utiliza a tecnica de clusterizacao, a qual consiste no agrupamento de dados semelhantes.

Por causa disso e de outras coisas, a minecacao de dados vem sendo trabalhada cada vez mais em pesquisas. Novos algoritmos para clustering vem sendo desenvolvidos e publicados em artigos a cada ano, dificultando o entendimento do verdadeiro estado-da-arte, devido ao fato da computacao ser uma area muito dinamica. Nosso objetivo Ã© analizar artigos de Ã¡reas diferentes que utilizaram tecnicas de clustering diferentes para atingir seus objetivos de forma a auxiliar o leitor na busca de informacoes pelo estado da arte de algoritmos de clustering.

Different clustering case of use

A boa e rapida deteccao de genes de doenca e fundamental para prevencao e tratamento de casos na medicina atual. Esta foi uma das motivacoes de (Deseases, Text Mining), que utilizou clustering de co-ocorrencia para extrair associacoes geneticas em um banco de dados medicos por meio de identificacao de entidades de forma a economizar tempo e facilitar a deteccao precoce de doencas como o 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 lclustering: 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 (COMPARACAO DE ALGORITMOS DE CLUSTERING HIERARQUICO EM DADOS REAIS: UM ESTUDO DE CASO NA AGRICULTURA), em que foram utilizados algoritmos de clustering hierarquicos para determinar quais areas precisam de adubo ou defensivos, de modo a evitar gastos desnecessarios.

Data Mining

Clustering

Algorithms

K-Means

Hierarchical clustering

DBSCAN

Temporal analysis

Conclusion