INSTITUTO SUPERIOR TÉCNICO

Curso em Engenharia Electrotécnica e de Computadores

ALUNO

Número: 75136

Nome: Diogo Alexandre Oliveira Silva

ORIENTAÇÃO

Nome: Ana Luísa Nobre Fred

Categoria:

Afiliação:

Nome: Helena Isabel Aidos Lopes

Categoria:

Afiliação:

DISSERTAÇÃO

Título: Efficient Evidence Accumulation Clustering for large datasets/big data

Data de provas: 04/12/2015 **Idioma: Inglês**

PALAVRAS CHAVES

Métodos de agrupamento, EAC, K-Means, GPGPU, Matrizes esparsas, Single-Link

KEYWORDS

Clustering methods, EAC, K-Means, GPGPU, Sparse matrices, Single-Link

RESUMO (250 palavras)

Avanços na tecnologia permitem a recolha e armazenamento de quantidades e variedades de dados sem precedente.
A maior parte destes dados são armazenados eletrónicamente e existe interesse em realizar análise automática dos mesmos.
As técnicas de clustering estão entre as mais populares para essa tarefa porque não assumem nada sobre a estrutura dos dados a priori.
Muitas técnicas existem, mas, típicamente, não têm um bom desempenho em todos os conjuntos de dados devido às especifidades de cada um.
Técnicas de ensemble clustering tentam responder a esse desafio ao combinar outros algoritmos.
Esta dissertação foca-se numa em particular, o Evidence Accumulation Clustering (EAC).
O EAC é uma algorithm robusto que tem demonstrado bons desempenhos na literatura numa variedade de conjuntos de dados.

**No entanto, esta robustez vem com um maior custo computacional associado.
A sua aplicação não só é mais lenta como está restrita a conjuntos de dados pequenos.
Assim, o objetivo desta dissertação é escalar o EAC, possibilitando a sua a aplicação a conjuntos de dados grandes, com tecnologia disponivel numa típica estação de trabalho.
Com isto em mente, várias abordagens foram exploradas: acelerar processamento com outros algoritmos (quantum clustering), através de processamento paralelo (em GPU), escalar com algoritmos de memória externa (disco rígido) e explorando a natureza esparsa do EAC.
Além disto, foi desenvolvido um método eficiente para construir matrizes esparsas específico ao EAC.
A solução proposta é aplicável a conjuntos de dados grandes e é entre 6 a 200 vezes mais rápida que a original para conjuntos pequenos.**

ABSTRACT (250 words)

**Advances in technology allow for the collection and storage of an unprecedent amount and variety of data.
Most of this data is stored electronically and there is an interest in automated analysis for generation of knowledge and new insights.
Since the structure of the data is unknown, clustering techniques become particularly interesting for knowledge discovery and data mining, as they make as few assumptions on the data as possible.
A vast body of work on these algorithms exist, yet, typically, no single algorithm is able to respond to the specificities of all data.
Ensemble clustering algorithm address this problem by combining other algorithms.
Evidence Accumulation Clustering (EAC) is a robust ensemble algorithm that has shown good results and is the focus of this dissertation.
However, this robustness comes with higher computational cost.
Its application is slower and restricted to smaller datasets.
Thus, the objective of this dissertation is to scale EAC, allowing its applicability to big datasets, with technology available at a typical workstation.
Accordingly, several approaches were explored: speed-up with other algorithms (\emph{quantum clustering}) or parallel computing (with GPU) and reducing space complexity by using external memory (hard drive) algorithms and exploiting the sparse nature of EAC.
A relevant contribution is a novel method to build a sparse matrix specialized in EAC.
Results show that the proposed solution is applicable to large datasets and presents speed-ups between 6 and 200 over the original implementation on different phases of EAC for small datasets.**

JÚRI

Presidente:

Nome: João Fernando Cardoso Silva Sequeira

Categoria: Professor Auxiliar

Afiliação: Departamento de Engenharia Electrotécnica e de Computadores (DEEC)

Vogais:

Nome: (Orientador) Ana Luísa Nobre Fred

Categoria: Professor Associado

Afiliação: Departamento de Bioengenharia (DBE)

Nome: Ana Paula da Silva Jorge

Categoria: Tenente-Coronel

Afiliação: Academia da Força Aérea

Nome: José Manuel dos Santos Vicêncio

Categoria: Brigadeiro-General

Afiliação: Director de Sistemas de Informação da Força Aérea

Nome: Pedro Filipe Zeferino Tomás

Categoria: Professor Auxiliar

Afiliação: Departamento de Engenharia Electrotécnica e de Computadores (DEEC)