LILIAN COELHO DE FREITAS
(ORGANIZADORA)

Collection:

APPLIED COMPUTER ENGINEERING

Atena Editora
Ano 2022

LILIAN COELHO DE FREITAS
(ORGANIZADORA)

Collection:

APPLIED COMPUTER ENGINEERING

Atena Editora
Ano 2022

Editora chefe
Profª Drª Antonella Carvalho de Oliveira
Editora executiva
Natalia Oliveira
Assistente editorial
Flávia Roberta Barão
Bibliotecária
Janaina Ramos
Projeto gráfico
Camila Alves de Cremo
Daphynny Pamplona
Gabriel Motomu Teshima
Luiza Alves Batista
Natália Sandrini de Azevedo
Imagens da capa
iStock
Edição de arte
Luiza Alves Batista

Atena Editora
Ano 2022

Atena
Editora
Ano 2022

Collection: applied computer engineering

Diagramação: Camila Alves de Cremo
Correção: Yaiddy Paola Martinez
Indexação: Amanda Kelly da Costa Veiga
Revisão: Os autores
Organizadora: Lilian Coelho de Freitas

Dados Internacionais de Catalogação na Publicação (CIP)

C697 Collection: applied computer engineering / Organizadora Lilian Coelho de Freitas. – Ponta Grossa - PR: Atena, 2022.

Formato: PDF
Requisitos de sistema: Adobe Acrobat Reader
Modo de acesso: World Wide Web
Inclui bibliografia
ISBN 978-65-5983-859-2
DOI: https://doi.org/10.22533/at.ed.592222801

1. Computer engineering. I. Freitas, Lilian Coelho de (Organizadora). II. Título.

CDD 621.39

Elaborado por Bibliotecária Janaina Ramos – CRB-8/9166

Atena Editora
Ponta Grossa – Paraná – Brasil
Telefone: +55 (42) 3323-5493
www.atenaeditora.com.br
contato@atenaeditora.com.br

Atena Editora
Ano 2022

DECLARAÇÃO DOS AUTORES

Os autores desta obra: 1. Atestam não possuir qualquer interesse comercial que constitua um conflito de interesses em relação ao artigo científico publicado; 2. Declaram que participaram ativamente da construção dos respectivos manuscritos, preferencialmente na: a) Concepção do estudo, e/ou aquisição de dados, e/ou análise e interpretação de dados; b) Elaboração do artigo ou revisão com vistas a tornar o material intelectualmente relevante; c) Aprovação final do manuscrito para submissão.; 3. Certificam que os artigos científicos publicados estão completamente isentos de dados e/ou resultados fraudulentos; 4. Confirmam a citação e a referência correta de todos os dados e de interpretações de dados de outras pesquisas; 5. Reconhecem terem informado todas as fontes de financiamento recebidas para a consecução da pesquisa; 6. Autorizam a edição da obra, que incluem os registros de ficha catalográfica, ISBN, DOI e demais indexadores, projeto visual e criação de capa, diagramação de miolo, assim como lançamento e divulgação da mesma conforme critérios da Atena Editora.

DECLARAÇÃO DA EDITORA

A Atena Editora declara, para os devidos fins de direito, que: 1. A presente publicação constitui apenas transferência temporária dos direitos autorais, direito sobre a publicação, inclusive não constitui responsabilidade solidária na criação dos manuscritos publicados, nos termos previstos na Lei sobre direitos autorais (Lei 9610/98), no art. 184 do Código Penal e no art. 927 do Código Civil; 2. Autoriza e incentiva os autores a assinarem contratos com repositórios institucionais, com fins exclusivos de divulgação da obra, desde que com o devido reconhecimento de autoria e edição e sem qualquer finalidade comercial; 3. Todos os e-book são *open access, desta forma* não os comercializa em seu site, sites parceiros, plataformas de *e-commerce,* ou qualquer outro meio virtual ou físico, portanto, está isenta de repasses de direitos autorais aos autores; 4. Todos os membros do conselho editorial são doutores e vinculados a instituições de ensino superior públicas, conforme recomendação da CAPES para obtenção do Qualis livro; 5. Não cede, comercializa ou autoriza a utilização dos nomes e e-mails dos autores, bem como nenhum outro dado dos mesmos, para qualquer finalidade que não o escopo da divulgação desta obra.

Atena
Editora
Ano 2022

APRESENTAÇÃO

Atena Editora is honored to present the e-book entitled "*Collection: Applied Computer Engineering*". This volume presents 17 chapters about applications of computer engineering in industrial automation, robotics, data science, information security, neuromarketing, speech development in children, among others.

We want to take this moment to thank all of our authors for entrusting us with their discoveries. We are also grateful to the reviewers and readers who have contributed to the success of our books.

Enjoy your reading.

Lilian Coelho de Freitas

SUMÁRIO

CAPÍTULO 2

ANÁLISE DE AGRUPAMENTO PARA APRIMORAR A EXTRAÇÃO AUTOMÁTICA DE DEMONSTRATIVOS FINANCEIROS COM ESTUDO DE ESCALABILIDADE

Data de aceite: 10/01/2022
Data de submissão: 19/11/2021

Igor Raphael Magollo
Universidade Federal de São Carlos
São Carlos - SP
http://lattes.cnpq.br/5954320813112873

Gabriel Olivato
Universidade Federal de São Carlos
São Carlos - SP
http://lattes.cnpq.br/1002422343204177

Victor Vieira Ferraz
Serasa Experian
São Carlos - SP
https://www.linkedin.com/in/victorvieiraferraz/

Murilo Coelho Naldi
Universidade Federal de São Carlos
São Carlos - SP
http://lattes.cnpq.br/0573662728816861

RESUMO: A análise das demonstrações financeiras é parte fundamental do processo de atribuição do risco de crédito, produzindo documentos que são fontes valiosas de informação sobre o património económico e financeiro das empresas. Grandes volumes desse tipo de documento exigem extração automática de dados e os localizadores conduzem as ferramentas para essa tarefa. Porém, por falta de regulamentação, não existe um *layout* padronizado para esses documentos, o que origina uma variedade de estruturas documentais. Essa variedade onera as ferramentas de extração de recursos, reduzindo seu desempenho. A análise de agrupamento supera essa sobrecarga ao encontrar os melhores grupos de documentos, permitindo o desenvolvimento de localizadores ajustados para cada grupo com base em suas características principais. Extensão de um trabalho anterior, este trabalho mostra que aplicar técnicas de agrupamento de última geração, RNG-HDBSCAN*, FOSC e MustaCHE, sobre documentos de demonstrações financeiras para avaliar seus grupos e estruturas principais, separar anomalias e analisar suas características principais, permite que os especialistas definam localizadores adequados para cada grupo, aumentando o desempenho das ferramentas de extração de dados. No entanto, com uma grande quantidade de documentos para serem agrupados, métodos sequenciais e centralizados tornam-se incapazes de executar essa tarefa em tempo hábil. Adicionalmente, o presente trabalho estuda maneiras de adaptação dessa solução para modelos escaláveis.

PALAVRAS-CHAVE: Ciência de dados, agrupamento, extração de características.

CLUSTERING ANALYSIS FOR IMPROVING AUTOMATIC DATA EXTRACTION FROM FINANCIAL STATEMENTS WITH SCALABILITY RESEARCH

ABSTRACT: The financial statement analysis is a fundamental part of the credit risk attribution process, producing documents that are valuable sources of information about companies' economic and financial wealth. Large volumes

of that type of document demand automatic data extraction, and locators drive the tools for that task. However, due to the lack of regulation, there is not a standard layout for such documents, which originates from a variety of document structures. Such variety burdens the feature extraction tools, reducing their performance. This work is an extension of previous work, where clustering analysis overcomes such burden by finding the best document clusters, allowing the development of fine-tuned locators for each cluster based on their main characteristics. We adopted state-of-the-art clustering techniques, RNG-HDBSCAN*, FOSC and MustaCHE, over financial statements documents to assess their clusters and main structures, separate outliers, and analyze their main features, allowing the specialists to define proper locators for each cluster, increasing the performance of the data extraction tools. Nevertheless, with a large number of documents, sequential or centralized clustering may not run in a timely manner. Additionally, this work studies ways to adapt the proposed solution to scalable models.
KEYWORDS: Data science, clustering, feature extraction.

1 | INTRODUÇÃO

Desde a década passada, segundo a Lei das Sociedades por Ações (11.638/07 e 11.941/09), as empresas devem elaborar suas demonstrações contábeis conforme as normas contábeis brasileiras de elaboração convergentes aos padrões internacionais (BRASIL 2009; 2007; 1976). Tais normas determinam as peças contábeis, assim como orientam o processo de classificação e ordenação de contas. Porém, as normas não definem uma estrutura (*layout*) de relatório a ser adotado pelas empresas (COMITÊ DE PRONUNCIAMENTOS CONTÁBEIS 2011b; BANCO CENTRAL DO BRASIL 2010), permitindo que contadores e desenvolvedores de sistemas contábeis estruturam as demonstrações conforme suas necessidades.

Fonte valiosa de informações sobre a situação econômico-financeira das empresas, a análise das demonstrações contábeis é parte fundamental dos processos de atribuição do risco de crédito (ASSAF NETO 2020). A fim de agilizar tal análise, foram propostas soluções tecnológicas para extração automática de dados, através de reconhecimento óptico de caracteres (OCR) e localizadores de chaves e valores no texto, como descrições de contas e saldos. Entretanto, tais soluções são eficientes quando os documentos não apresentam alta variabilidade estrutural, possibilitando aplicar o localizador mais apropriado. Em particular, devido à falta de informação estrutural precisa, a Serasa Experian adotou um localizador "genérico" sobre 85% do volume de documentos, o que limitou a eficiência da extração. O ideal seria possuir localizadores adequados para cada um dos *layouts* dos documentos. Entretanto, não se sabe a priori quantos *layouts* encontram-se no conjunto de documentos, nem mesmo se existem documentos com *layouts* semelhantes. Tal informação ajudaria muito no desenvolvimento de localizadores apropriados para cada conjunto de documentos com *layouts* parecidos. O agrupamento de dados consiste em dividir os dados de forma que elementos correlacionados estejam em um mesmo grupo. Adicionalmente é possível

detectar elementos que não se encaixam em nenhum grupo, chamados de externos (em inglês *outliers*). O agrupamento de documentos com *layouts* correlacionados permite uma análise dos grupos obtidos, de forma que seja possível gerar localizadores específicos para suas características.

Este trabalho consiste em uma extensão do trabalho apresentado em FERRAZ et al. 2020, onde o principal objetivo consiste no agrupamento de demonstrações contábeis para fins da extração automática de descrição de contas e saldos contábeis. Baseando-se na premissa de que documentos semelhantes possuem estruturas com características muito próximas e, por isso, viabilizam a melhoria do processo de detecção e extração dos dados. Para isso, usamos um conjunto de técnicas estado-da-arte para análise de grupos, até então um trabalho aplicado inédito. RNG-HDBSCAN* (ARAUJO NETO et al. 2019), capaz de obter inúmeras hierarquias de HDBSCAN* em uma única execução, o arcabouço FOSC (CAMPELLO et al. 2013) e a ferramenta de visualização MustaCHE (ARAUJO NETO et al. 2018), que permitem a obtenção e análise de partições multiníveis para cada hierarquia obtida, resultando em uma análise robusta da estrutura dos dados e independente de parâmetros comuns para outros algoritmos, como a pré-definição do número de grupos, possíveis inicializações ou mesmo um valor mínimo de densidade. Por meio da análise apresentada nesse trabalho (FERRAZ et al. 2020), é possível selecionar os grupos de maior interesse entre a alta variedade de modelos de documentos contábeis e implementar localizadores específicos para extração de informação precisa de cada grupo, aumentando a qualidade e eficiência do processo de extração automática. Adicionalmente, a técnica aplicada permite a detecção de documentos externos (*outliers*).

Com o aumento do volume de dados gerados a cada dia, a grande quantidade de dados para análise rapidamente supera as limitações de uma única máquina, e modelos de programação que consigam trabalhar com grandes volumes de dados em tempo hábil, tolerantes a falhas e que permitam escalabilidade de processamento adicionando poder computacional sob demanda se tornam necessários, como o MapReduce (DEAN et al. 2004). Em (SANTOS et al. 2021) é apresentado o algoritmo MR-HDBSCAN*, uma adaptação aproximada para o algoritmo HDBSCAN* para o modelo MapReduce. O principal problema solucionado por esse algoritmo foi a construção da Árvore Geradora Mínima (*MST* no inglês) de maneira escalável e diferentemente do RNG-HDBSCAN*, o MR-HDBSCAN* computa uma única hierarquia. Adicionalmente, estendemos nosso trabalho anterior (FERRAZ et al. 2020) no presente trabalho, com o objetivo de estudar a escalabilidade da construção da *MST* no modelo de programação MapReduce.

2 | TRABALHOS RELACIONADOS

O aumento do volume de dados gerados levou à necessidade de classificação, organização e extração de conhecimento dos mais diversos tipos de documentos.

(MOURA 2004) analisou empiricamente uma série de ferramentas de mineração para melhorar a automação do processo de identificação, seleção e classificação de conteúdo relevante para a Agência de Informação Embrapa. Em sua proposta estabeleceu-se o uso de ferramentas para obtenção de grupos que auxiliem na classificação do conteúdo, posteriormente validados por "uma parceria com um especialista do domínio escolhido" (MOURA 2004).

(MADEIRA 2015) relatou resultados favoráveis ao aplicar k-médias (JAIN 2010) para identificar possíveis empresas alvo de fiscalização tributária no município do Rio de Janeiro, através de dados extraídos das Notas Fiscais de Serviços Eletrônicas (NFS-e). Mais recentemente (SNOW 2018), dados extraídos de formulários enviados por empresas inglesas ao Companies House, órgão oficial de registro das empresas britânicas, foram agrupados e analisados para gerar informação relevante para o processo de definição do código SIC (Standard Industrial Classification), sistema de classificação das empresas por ramo de atividade. O modelo conseguiu identificar erros de classificação e novas tendências de atividades empresariais através da vetorização do conteúdo textual referente à descrição da atividade informada pelas próprias empresas, adicionada da incorporação bidimensional dos dados para visualização intuitiva e agrupamento hierárquico dos dados baseado em densidade, com uso do HDBSCAN*.

3 I METODOLOGIA

3.1 Caracterização e Processamento de Demonstrativos Contábeis

Dentre os principais demonstrativos obrigatórios que compõem um relatório financeiro, estão o balanço patrimonial e a demonstração do resultado do exercício. Essas peças contábeis provêm as informações necessárias para a maioria das análises financeiras (ASSAF NETO 2020) e, portanto, são o principal alvo do processo de extração automática. Com base em suas posições e composições, foram definidas 12 características dos relatórios financeiros que remetem à sua estrutura, identificadas nos documentos financeiros com uso de aplicação de extração automática.

Para efetuar a extração foi implementado um localizador que, após o processo de OCR sobre os textos originais, processou um conjunto de regras de decisão, elaboradas em conjunto com os especialistas em risco de crédito, para definir os valores dessas características. O localizador utiliza objetos que a aplicação oferece sobre o documento (páginas, linhas de texto, palavras etc.), cada qual com uma série de métodos e propriedades como, por exemplo, no caso de uma palavra, é possível saber: a distância das margens e do topo em pixels, o número da página, se ela é uma palavra-chave, dentre outras possibilidades. Neste trabalho, as características com seus respectivos pesos são apresentadas na Tabela I. O peso determina a relevância de uma característica para a

análise dos agrupamentos, segundo os especialistas em risco de crédito da Serasa Experian.

De acordo com (ASSAF NETO 2020), o balanço patrimonial é o "elemento de partida indispensável para o conhecimento da situação econômica e financeira de uma empresa" e seu conceito ("balanço") tem origem no equilíbrio entre as partes que o compõem: ativo, passivo e patrimônio líquido[1]. A forte relação conceitual entre ativo e passivo reflete em contas com descrições muito semelhantes, frequentemente observadas nos documentos financeiros. Essa relação dificulta a parametrização do método de extração, uma vez que as orientações são passadas através de palavras-chave, coordenadas cardeais e colaterais referentes aos dados. Por exemplo, uma conta pode apresentar o termo "circulante" ao seu norte (acima no texto da mesma página), que nomeia tanto o grupo de ativos, quanto de passivos de curto prazo. Nessas situações, é necessário identificar o grupo correto para que a conta receba a classificação apropriada. Posto isso, definiu-se a característica 1, a mais relevante a ser identificada no documento financeiro: a posição do ativo em relação ao passivo, com três situações: ativo à esquerda e passivo à direita, lado a lado na mesma página; ativo acima do passivo, na mesma página; e ativo e passivo em páginas diferentes.

Nº.	Peso	Tipo	Descrição
1	16%	Categórica	Posição do ativo em relação ao passivo
2	7%	Booleana	Indica se o grupo de ativos de longo prazo e permanentes existe
3	9%	Booleana	Indica se o grupo de passivos de longo prazo existe
4	1%	Booleana	Indica se a demonstração do resultado do exercício existe
5	9%	Categórica	Indica o tipo: balanço ou balancete
6	12%	Booleana	Indica se os saldos das contas estão desdobrados
7	10%	Percentual	Proporção de números em relação ao total de termos
8	12%	Inteiro	Quantidade de colunas com saldos e/ou movimentação de contas
9	8%	Inteiro	Quantidade de páginas do ativo
10	8%	Inteiro	Quantidade de páginas do passivo
11	8%	Inteiro	Quantidade de páginas do documento
12	0%	Booleana	Indica se o modelo já é padronizado SPED

Tabela 1. Características estruturais dos relatórios financeiros.

Conforme definido pelo (COMITÊ DE PRONUNCIAMENTOS CONTÁBEIS 2011b) a empresa deve apresentar ativos e passivos circulantes e não circulantes como grupos separados no balanço patrimonial. Isso reflete na ocorrência de contas com descrições idênticas e classificações diferentes, comumente observadas nas demonstrações contábeis, conflitando a parametrização através de palavras-chave e coordenadas. Por exemplo, um passivo pode apresentar uma conta com a descrição "contas a pagar" tanto no grupo circulante, quanto no não circulante. Esse fator é definido nas características 2 e 3.

1 O passivo e o patrimônio líquido são comumente apresentados em conjunto e denominados simplesmente como passivo.

A demonstração do resultado do exercício (DRE) contém o cálculo do resultado (lucro ou prejuízo) auferido pela empresa em um período. Ele é composto pelas contas de receitas, despesas e custos em uma sequência esquematizada (ASSAF NETO 2020). A DRE apresenta conteúdo bem distinto do restante de um documento, o que minimiza conflitos na extração automática. Uma parcela pequena de relatórios financeiros são incompletos, contendo somente o balanço patrimonial. Sendo assim, a característica 4 indica a presença da DRE no documento.

O (COMITÊ DE PRONUNCIAMENTOS CONTÁBEIS 2011a) estabelece o conteúdo mínimo das demonstrações contábeis intermediárias, ou seja, aquelas que são elaboradas no decorrer do exercício e, portanto, se referem à parte do ano fiscal. Essas demonstrações, denominadas balancetes, são elaboradas para fins gerenciais e de acompanhamento do desempenho da empresa e, por isso, costumam apresentar alto grau de detalhamento dos dados. Isso aumenta significativamente a complexidade do processo de extração e, portanto, balanços e balancetes são indicados pela característica 5.

Relatórios financeiros mais complexos, que demandam maior quantidade de regras específicas na parametrização do localizador de características, geralmente apresentam:

1. Demonstrações contábeis comparativas entre períodos distintos;
2. Demonstração da movimentação dos saldos (saldo inicial, créditos, débitos e saldo final);
3. Desdobramento dos saldos das contas;
4. Alta proporção de termos numéricos em relação ao total de termos na página inicial do balanço;
5. Elevada quantidade de páginas como um todo, e/ou especificamente nos grupos ativo e/ou passivo.

Tais características foram definidas de 6 a 11. Em particular, as características 6, 7 e 8 consideraram a probabilidade de alta correlação e tendência a médias superiores para o conjunto de documentos do tipo balancete. A característica 12 identificou quais documentos já apresentavam o único modelo padronizado de relatório financeiro considerado até então: o modelo SPED (Sistema Público de Escrituração Digital), com o único objetivo de excluir esses casos da amostra (uma vez que são padronizados e a extração automática é facilitada). Adicionalmente também foram excluídos da amostra os vetores de características duplicados e documentos para os quais não foram extraídas todas as características. Por fim, o conjunto final amostrado possui 1.492 vetores de características de documentos financeiros.

3.2 HDBSCAN*

O algoritmo HDBSCAN* (CAMPELLO et al. 2013) possui diversas vantagens comparado a outros algoritmos particionais e hierárquicos tradicionais. Combina os

aspectos de agrupamento baseado em densidade e hierárquico, onde suas hierarquias são construídas pelas densidades dos grupos, dos quais podem ser extraídos os mais proeminentes. Seus resultados podem ser visualizados através de um dendrograma, uma árvore de agrupamento simplificada e outras técnicas de visualização que não necessitam de nenhum parâmetro crítico como entrada. O algoritmo HDBSCAN* recebe apenas um parâmetro de entrada, m_{pts}, que pode ser entendido como um fator de suavização de densidade não paramétrico realizado pelo algoritmo. HDBSCAN* também é flexível na análise de seus resultados, deixando o usuário analisar diretamente a hierarquia e a árvore de agrupamentos diretamente ou até realizar cortes na árvore para obter um particionamento dos dados equivalente ao DBSCAN.

Dado um conjunto de dados $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ com $\mathbf{n}$ objetos e um valor de suavização $\mathbf{m}_{pts}$, as seguintes definições são utilizadas pelo HDBSCAN*:

- *Core Distance*: distância de núcleo de um objeto $\mathbf{x}_p \in \mathbf{X}$ para $\mathbf{m}_{pts}$, $d_{core}(\mathbf{x}_p)$ é a distância de $\mathbf{x}_p$ até o seu m_{pts}-ésimo vizinho mais próximo (incluindo $\mathbf{x}_p$), ou seja, o raio mínimo em que x_p é considerado um objeto denso (*core object*).
- *Mutual Reachability Distance:* a distância de alcance mútuo entre dois objetos $\mathbf{x}_p$, $\mathbf{x}_q \in \mathbf{X}$ para $\mathbf{m}_{pts}$ é definida como $d_{mreach}(\mathbf{x}_p,\mathbf{x}_q) = \max(d_{core}(\mathbf{x}_p), d_{core}(\mathbf{x}_q), d(\mathbf{x}_p,\mathbf{x}_q))$ e pode ser interpretada como um raio mínimo tal que ambos os objetos são densos e estão dentro da vizinhança um do outro.
- *Mutual Reachability Graph:* o grafo de alcance mútuo de um conjunto de dados $\mathbf{X}$ para m_{pts} é um grafo completo e ponderado, $\mathbf{G}_{mreach}$, em que os objetos são os vértices e o peso de cada aresta entre cada par de vértices é dados por sua *mutual reachability distance*.

Os passos principais do HDBSCAN* estão descritos no Algoritmo 1 (CAMPELLO et al. 2013). Com seu resultado é possível realizar análise de agrupamentos, detecção de ruído e visualização dos dados.

Algorithm 1: HDBSCAN*

Data: $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n\}$, m_{pts}
Result: Hierarquia HDBSCAN*

(1) Dado um conjunto de dados $\mathbf{X}$, calcular as *core distances* de todos os seus objetos.
(2) Calcular a árvore de abrangência mínima *MST*, sobre o *mutual reachability graph*, $G_{m_{reach}}$.
(3) Estender a *MST* com laços a cada vértice com peso iguais ao seu *core distance*, obtendo MST_{ext}.
(4) Extrair a hierarquia HDBSCAN* como um dendrograma da MST_{ext}.
  (a) Todos os objetos são definidos com o mesmo rótulo (a raíz da árvore de grupos).
  (b) Iterativamente remover arestas da MST_{ext} em ordem decrescente de pesos.
    i. Arestas com o mesmo peso são removidas simultaneamente.
    ii. Após a remoção de uma aresta, os rótulo do agrupamento são designados aos dois componentes conexos que contém um vértice da aresta removida. Um novo rótulo é adicionado se o componente tem pelo menos uma aresta restante,. Caso contrário, os objetos no componente são rotulados como externos (*outliers*).

Algoritmo 1. HDBSCAN*.

O HDBSCAN* necessita do número de objetos mínimo, chamado de m_{pts}, para uma região ser considerada "densa", e sua mais recente versão, o RNG-HDBSCAN* (ARAUJO NETO et al. 2019), calcula de forma eficiente todas as hierarquias para diferentes valores em um intervalo [*min,max*] de m_{pts} definido. Outra vantagem da utilização do HDBSCAN* no contexto empresarial é ser relacional, ou seja, o algoritmo necessita apenas das relações (similaridades) entre os objetos a serem agrupados, mas não dos objetos propriamente ditos. Portanto, algoritmos relacionais podem garantir a confidencialidade dos dados. Adicionalmente, é possível analisar visualmente todas as hierarquias geradas por meio do MustaCHE (ARAUJO NETO et al. 2018), uma poderosa ferramenta de visualização que permite a análise de múltiplas hierarquias de agrupamentos baseados em densidade gerados com o HDBSCAN*. Através de gráficos interativos, ele permite a comparação simultânea dos resultados para diversos valores de m_{pts}. O MustaCHE utiliza o índice de acordo hierárquico (HAI) (JOHNSON et al. 2013) para relacionar os agrupamentos de acordo com suas semelhanças, podendo fundir os agrupamentos obtidos utilizando o próprio HDBSCAN*, o que dá origem a um meta-agrupamento.

3.3 MR-HDBSCAN*

Em (SANTOS et al. 2021), é abordado o problema da construção da *MST* em um ambiente escalável. Ao trabalhar com grandes conjuntos de dados (Big Data), métodos escaláveis são essenciais, pois é possível utilizar o poder computacional de *N* máquinas ao invés de apenas uma. Em Teoria dos Grafos, a *MST* é uma árvore que conecta todos os vértices de um grafo com custo mínimo. A *MST,* assim como o grafo dos *k* vizinhos mais próximos (*k-NNG* no inglês), é um grafo de proximidade, ou seja, carrega consigo um resumo de informações de proximidade entre os vértices (SHIMOMURA et al. 2021). Alguns métodos de agrupamento utilizam informações de proximidade e constroem a *MST* como parte de seu processo, como é o caso do HDBSCAN*. Nesse contexto, cada objeto no conjunto de dados é representado por um vértice, e as relações entre cada par de objetos são representadas por arestas cujos custos são medidas de dissimilaridade e similaridade, frequentemente interpretadas como a distância entre os objetos, que precisam ser calculadas. Dessa forma, a complexidade computacional é proporcional ao quadrado da quantidade de pontos no conjunto de dados, e em grandes conjuntos de dados, nem mesmo sistemas escaláveis são capazes de lidar em tempo hábil. Portanto, soluções aproximadas, cujo objetivo é particionar o espaço obtendo resumos de proximidade para reduzir a quantidade de cálculos de distâncias são comumente abordadas para soluções escaláveis, e esse é o caso do MR-HDBSCAN*.

Para a construção da *MST* em MapReduce, Santos et al. (2021) apresenta um novo particionamento dos dados para distribuição de processamento. O particionamento, denominado *Recursive Sampling*, amostra o conjunto de dados e agrupa as amostras utilizando o HDBSCAN*. Com o agrupamento das amostras, ele obtém informação

espacial resumida acerca dos grupos mais proeminentes. Então, ele classifica o restante do conjunto de dados utilizando esse modelo e distribui as partições para as unidades de processamento de acordo com a classificação. Cada unidade de processamento verifica se a partição cabe em sua memória principal; caso a partição não caiba, o processo se repete de maneira recursiva até que cada partição possa ser processada em memória principal. Nessa etapa, para cada partição, o algoritmo transforma os dados para a G_{mreach}, e computa as *MST*s locais. Posteriormente, o método conecta as *MST*s locais em uma única *MST* através da ordem da recursão.

3.4 Construção de uma *MST* Aproximada Utilizando um *k-NNG*

Em (DONG et al. 2011), é apresentado um método heurístico iterativo para construção de uma aproximação de um *k-NNG*. O algoritmo, denominado NNDescent, inicializa com um grafo de *k* vizinhos aleatórios, e a cada iteração, busca na vizinhança dos vizinhos de cada vértice por candidatos a vizinhos mais próximos e assim otimizar a vizinhança inicial. O método se mostrou eficiente, com uma complexidade empírica de $\Theta(\mathbf{n}^{1,18})$ e revocação superior a 0,9 na maior parte dos experimentos. Posteriormente, Warashina et al. (2014) apresentou uma adaptação desse método para o MapReduce, cuja complexidade empírica foi de $\Theta(\mathbf{n}^{1,35})$.

Apesar da *MST* não estar inteiramente contida no *k-NNG,* ele também é um grafo de proximidade, pois resume informações estruturais com relação à proximidade dos vértices (SHIMOMURA et al. 2021). Além disso, o *k-NNG* com $k = m_{pts} - 1$ é o suficiente para computar todas as *Core Distances* e transformar as arestas obtidas para o espaço da *Mutual Reachability Distance.* Por fim, o NNDescent calcula várias arestas das quais não estão entre os *k* mais próximos durante seu processo de otimização, arestas essas, denominadas aqui como arestas residuais, são cálculos necessários e inevitáveis. Dado isso, apresentamos um método escalável para construir uma aproximação da *MST* de maneira escalável utilizando o *k-NNG* utilizando como hipótese que as arestas residuais geradas no processo de otimização do NNDescent são arestas candidatas que faltam no *k-NNG* para completar a *MST.*

O método proposto consiste em construir o *k-NNG* utilizando o NNDescent escalável armazenando separadamente as arestas residuais. Em seguida, aplicar o algoritmo de Boruvka para construir a *MSF (*Floresta Geradora Mínima do inglês) a partir do grafo constituído pela união entre o *k-NNG* e as arestas residuais. O algoritmo de Boruvka foi escolhido por ser conhecido como um método naturalmente paralelizável e facilmente escalável (BADER; CONG, 2006). Em seguida, caso a *MSF* não for uma *MST,* conectar as árvores através de amostragem de arestas entre os componentes conexos do grafo.

Os conjuntos de dados utilizados nos experimentos foram gerados a partir de bolhas gaussianas isotrópicas. No total, cinco conjuntos de dados foram construídos com três bolhas variando a dimensão e a quantidade de pontos entre 2 e 32,e entre 2000 e 8000,

respectivamente, como mostrado na Tabela 2.

ID	Tipo de dado	Dimensão	Nº de pontos	Nº de bolhas
D2N2	Bolhas Gaussianas Isotrópicas	2	2000	3
D2N4	Bolhas Gaussianas Isotrópicas	2	4000	3
D2N8	Bolhas Gaussianas Isotrópicas	2	8000	3
D8N2	Bolhas Gaussianas Isotrópicas	8	2000	3
D32N2	Bolhas Gaussianas Isotrópicas	32	2000	3

Tabela 2. Descrição dos conjuntos de dados gerados para testes.

Os experimentos foram realizados em um computador com um processador AMD Ryzen™ 5 1600 com 12 *threads*, 16 GB de memória RAM rodando o sistema operacional Ubuntu 20.04.2 LTS e a plataforma Apache Spark 3.0.0.

4 | ANÁLISE DOS RESULTADOS

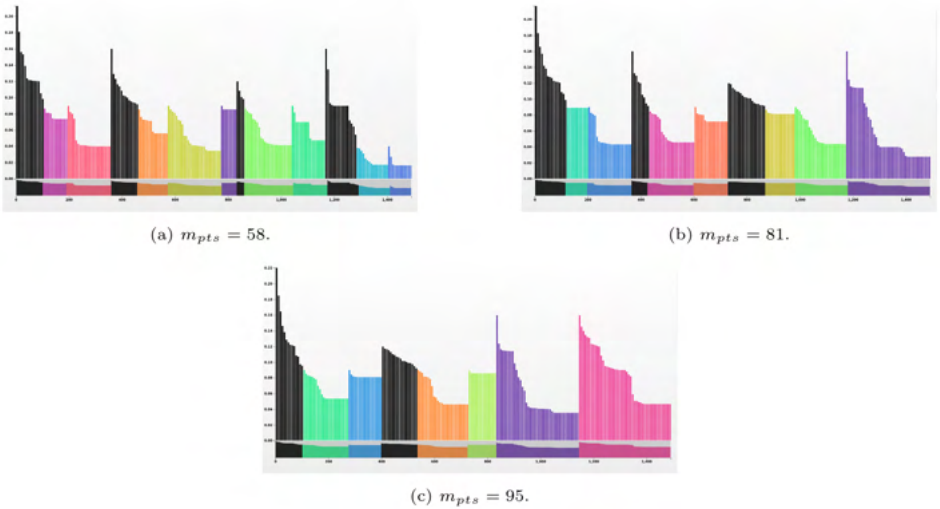
4.1 Agrupamento dos Documentos Contábeis

A execução do RNG-HDBSCAN*, seguida da análise dos resultados por meio do MustaCHE, três partições geradas com diferentes valores de m_{pts} foram escolhidas porque possuíam menos de 10 grupos e maior homogeneidade dos valores de alcance mútuo entre documentos do mesmo grupo. A Tabela 3 resume os resultados e a Figura 1 mostra os gráficos de alcance mútuo para valores de m_{pts}.

m_{pts}	58	81	95
Quantidade de grupos	9	7	6
Percentual de externos	23	21	16

Tabela 3. Quantidade de grupos e percentual de externos *outliers* por m_{pts}.

Os gráficos de alcance (Figura 1) mostram que grupos homogêneos são mais estáveis, pois há uma baixa variação no seu nível de densidade mútua. Os “vales” são as regiões mais densas de cada grupo, e no geral, eles contêm os documentos mais semelhantes com o restante de seu grupo. Ao aumentar o parâmetro m_{pts} a quantidade de grupos diminui, assim como os externos (*outliers*), de cor preta.



(a) $m_{pts} = 58$. (b) $m_{pts} = 81$. (c) $m_{pts} = 95$.

Figura 1. Visualização dos grupos nos gráficos de alcance mútuo gerados a partir de diferentes valores de m_{pts}.

Os resultados foram analisados pelos especialistas em risco de crédito da Serasa Experian. A posição do ativo em relação ao passivo foi a característica determinante na definição dos resultados, posto que em todos os grupos formados, todos os documentos apresentaram valores iguais. Documentos com ativo e passivo em páginas diferentes foram predominantes, em média 42% dos casos agrupados para cada m_{pts}, formando grupos exclusivos. Com 95 e 81 m_{pts}, foram formados 4 grupos de documentos nessa situação, distintos entre si pela presença de contas do passivo de longo prazo e saldos desdobrados e, ainda, pela quantidade de colunas com saldos de contas e/ou pela quantidade de páginas do ativo e do passivo. Com 58 m_{pts}, foram formados 5 grupos de documentos com ativo e passivo em páginas diferentes e, além dos padrões identificados acima, notou-se que um dos grupos continha somente balanços e outro somente balancetes. Também notou-se influência dessas características na distinção dos agrupamentos de documentos com ativo à esquerda e passivo à direita, lado a lado na mesma página, e com ativo acima do passivo na mesma página.

A Tabela 4 traz um quadro com o resumo das características predominantes em cada grupo formado com 58 m_{pts}, dado o detalhamento e grau de distinção obtido entre os documentos agrupados com esse valor. Os grupos foram identificados com cores de acordo com o gráfico da figura 1 e a coluna “%” refere-se à proporção de objetos em relação à amostra. Nota-se a predominância de agrupamentos com documentos com ativo e passivo em páginas diferentes e as demais características evidenciam os padrões de distinção descritos anteriormente. A coluna 5 exibe o percentual de documentos do tipo balanço e as demais colunas trazem os valores médios de cada uma das características,

observados para o respectivo agrupamento.

Característica	%	Posição do ativo e do passivo	2	3	4	5	6	7	8	9	10	11
Azul Claro	8	lado a lado, mesma página	100%	100%	71%	100%	0%	31%	2.00	1.00	1.00	7.79
Azul	5	lado a lado, mesma página	100%	100%	82%	100%	0%	25%	1.00	1.27	1.27	5.36
Lilás	6	ativo acima, mesma página	20%	0%	70%	100%	0%	30%	1.20	1.00	1.10	8.50
Rosa	11	ativo acima, mesma página	100%	100%	90%	100%	0%	31%	2.10	1.00	1.20	9.60
Roxo	4	páginas diferentes	100%	100%	71%	0%	100%	53%	3.86	3.71	3.71	14.29
Verde musgo	14	páginas diferentes	92%	100%	67%	100%	100%	31%	1.67	2.79	2.13	11.13
Laranja	8	páginas diferentes	64%	0%	79%	100%	100%	33%	1.50	2.00	1.50	8.64
Verde Claro	12	páginas diferentes	91%	100%	91%	100%	0%	35%	2.23	1.09	1.14	21.09
Verde	9	páginas diferentes	47%	0%	100%	100%	0%	34%	1.40	1.00	1.33	6.80
Média	-	-	**82%**	**72%**	**81%**	**95%**	**33%**	**33%**	**1.83**	**1.59**	**1.50**	**10.94**

Tabela 4. Análise para 58 m_{pts}.

Além do levantamento estatístico dos vetores de características dos documentos agrupados, também foram resgatados os arquivos originais com os relatórios financeiros em uma quantidade equivalente a 12% da amostra. Esses foram examinados quanto à similaridade visual e complexidade técnica pelos especialistas em análise de risco de crédito de empresas. Para esse processo, os relatórios foram organizados conforme os resultados do agrupamento e examinados comparativamente em conjunto.

O exame dos arquivos originais também apresentou conclusões favoráveis aos resultados do agrupamento, para todos os valores de m_{pts}. Em média, 88% dos documentos de um mesmo grupo apresentaram alto grau de similaridade visual e técnica. Os especialistas apontaram poucas exceções (cerca de 12%), para as quais foram identificadas inconsistências nos valores de determinadas características, dadas situações não previstas pela árvore de decisões do *script* de localização.

Para validar os agrupamentos no processo de extração automática de dados, os especialistas selecionaram o documento mais representativo de 4 agrupamentos com características distintas. A qualidade da extração automática dos dados do ativo e do passivo desses documentos foi aferida através do processo atual de extração e, então, foram apurados os erros e dados não extraídos passíveis de tratamento, considerado o conhecimento prévio de características de acordo com o agrupamento. Tal análise permitiu estimar o potencial de melhoria da qualidade da extração automática dos dados desses documentos, demonstrado na tabela abaixo. A coluna Rep. média exibe a representatividade média do grupo no qual o documento foi classificado, considerando os 3 valores de m_{pts}.

	mpts			Rep. média	Qualidade da extração	
	58	81	95		Genérica	Com agrupamento
Clusters	Azul Claro	Roxo	Roxo	17%	69%	80%
	Rosa	Azul	Rosa	15%	67%	77%
	Verde Musgo	Verde Claro	Laranja	13%	45%	69%
	Verde Claro	Rosa	Verde	12%	43%	60%

Tabela 5. Potencial de melhoria da qualidade da extração automática.

O agrupamento demonstrou potencial de melhoria no processo de extração automática de dados, com a qualidade média variando de 56% para 71% nos documentos analisados. Considerando ainda que atualmente apenas 15% dos documentos têm *layout* conhecido (SPED), destaca-se que os agrupamentos analisados representam, em média, 57% do volume processado, o que indica potencial para aumento significativo da faixa de casos cobertos por métodos padronizados de extração automática.

4.2 Método Proposto para Escalabilidade da MST

Para medir a qualidade e comprovar a hipótese descrita anteriormente, foram construídas *MST*s utilizando as informações residuais do NNDescent (referenciadas como *MSTR*) e *MST*s sem utilizar essas informações (referenciadas como MSTN).

Ao observar as Tabelas 6 e 7, observou-se que o tempo computacional médio do cômputo da *MST* utilizando informação residual é consideravelmente maior do que a construção sem utilizar essa informação. Isso se dá pelo aumento na quantidade de arestas no momento da execução do algoritmo de Boruvka.

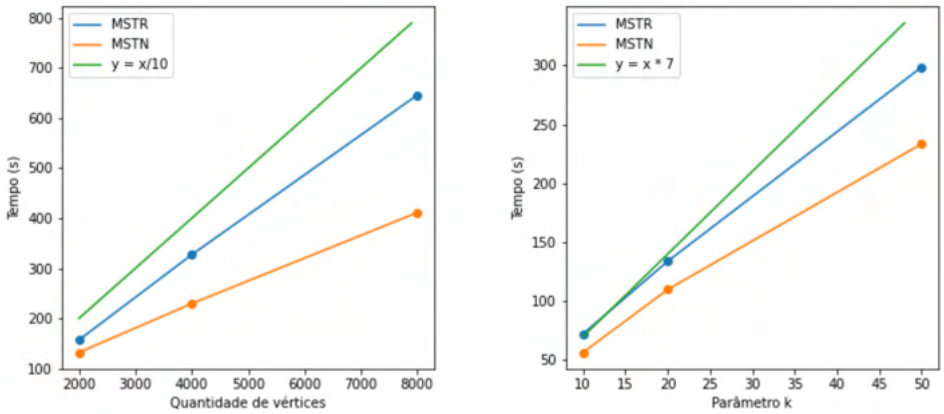
Método	Base	Tempo *NNDescent* (s)	Tempo *MST* (s)	Total (s)
MSTR	D2N2	118,69	38,18	156,87
	D2N4	239,66	87,57	327,23
	D2N8	448,79	197,31	646,10
MSTN	D2N2	126,46	5,61	132,06
	D2N4	221,12	8,46	229,57
	D2N8	401,85	9,51	411,36

Tabela 6. Tempo de execução variando o número de vértices, com $k = 25$.

Método	k	Tempo *NNDescent* (s)	Tempo *MST* (s)	Total (s)
MSTR	10	53,78	18,54	72,32
	20	101,94	31,96	133,90
	50	232,45	66,01	298,46
MSTN	10	50,93	5,66	56,60
	20	104,30	5,62	109,92
	50	226,76	6,65	233,41

Tabela 7. Tempo de execução variando o *k* na base D2N2.

Na Figura 2 foi observado que ambos os métodos obtiveram uma complexidade próxima a linear. O método que não utiliza a informação residual se mostrou mais rápido, devido a menor quantidade de arestas para computar a *MST.*



(a) Número de vértices x Tempo de execução (b) k x Tempo de execução

Figura 2. Comparação de desempenho entre os métodos.

Para auferir a qualidade do método proposto, comparou-se o custo das *MST*s construídas com o custo da *MST* exata. A comparação foi feita através da razão entre o custo da *MST* aproximada pelo custo da *MST* exata, como mostra a Tabela 8. Quanto mais próximo de 1, melhor. Além disso, foi adicionado um custo relativo de uma tentativa de aproximação através de uma escolha aleatória de arestas como referência. Notou-se que em geral a utilização da informação residual diminuiu o custo das aproximações construídas.

Base	k	Custo rel. MSTR	Custo rel. MSTN	Custo rel. Aleatório
D2N2	10	1,1307	1,1476	8,7156
	20	1,0189	1,0398	5,9857
	50	1,0048	1,0291	3,8181
D8N2	10	1,0328	1,0358	2,0495
	20	1,0036	1,0057	1,7687
	50	1,0003	1,0029	1,5354
D32N2	10	1,0177	1,0183	1,3277
	20	1,0030	1,0039	1,2457
	50	1,0001	1,0007	1,1801

Tabela 8. Custo relativo (*MST* aproximada / *MST* exata) para cada método.

Por fim, verificou-se a quantidade de arestas erradas (arestas encontradas na aproximação que não fazem parte da *MST* exata). Além disso, foi feita uma contagem das arestas obtidas por amostragem para conectar a *MSF* em uma única *MST*. Novamente foi adicionada uma aproximação aleatória como referência. Observou-se que apesar do custo da aproximação obtida com informação residual ser mais baixo, a quantidade de arestas erradas é equivalente às arestas erradas na aproximação sem utilizar a informação residual.

Ou seja, o método que utiliza as arestas residuais encontrou melhores alternativas, porém, não corretas. Além disso, observou-se também que em geral, ao utilizar a informação residual, uma *MST* foi obtida ao final do algoritmo de Boruvka, sem a necessidade de amostragem de arestas. Veja a Tabela 9.

Base	k	Arestas obtidas por amostragem			$\mid MST_{aprox.}(V) - MST_{exata}(V) \mid$		
		MSTR	MSTN	Aleatório	MSTR	MSTN	Aleatório
D2N2	10	0	2	0	74	74	1976
	20	0	2	0	7	7	1955
	50	0	2	0	3	3	1896
D8N2	10	0	2	0	281	281	1988
	20	0	2	0	33	33	1951
	50	0	2	0	2	3	1910
D32N2	10	0	2	0	569	569	1973
	20	0	2	0	115	115	1958
	50	0	2	0	3	4	1895

Tabela 9. Quantidade de arestas obtidas por amostragem e quantidade de arestas diferentes da MST exata.

5 I CONCLUSÃO

Esta pesquisa avaliou o potencial de melhoria do processo de extração automática de dados de demonstrações contábeis de empresas através da análise dos agrupamentos gerados com HDBSCAN* sobre vetores de características extraídas desses documentos. A ferramenta MustaCHE apoiou oprocesso de seleção dos 3 valores de m_{pts}, cujos resultados foram analisados quanto à significância dos agrupamentos formados e à aderência ao processo de extração automática de dados.

De acordo com os resultados apurados com a análise estatística dos vetores de características dos documentos agrupados e com os apontamentos à partir do exame dos arquivos originais realizado pelos especialistas em análise de risco de crédito, o HDBSCAN* agrupou os documentos de forma coerente, em grupos distintos entre si, com os 3 valores de m_{pts} selecionados. Dada a quantidade maior de agrupamentos formados com 58 m_{pts}, notou-se um grau superior de detalhamento e distinção entre os grupos, o que viabilizou a identificação de complexidades de extração automática diferentes e a percepção de maior aproveitamento conceitual das características e suas respectivas relevâncias.

A atribuição de pesos às características contribuiu para a formação de agrupamentos aderentes ao processo de extração automática de dados de demonstrações contábeis e, consequentemente, potencial de melhoria através da implementação de localizadores mais robustos e ajustados conforme as características estruturais mais relevantes dos documentos agrupados.

Vale ressaltar que o exame dos relatórios financeiros, realizado pelos especialistas em análise de risco de crédito, ratificou alta assertividade do localizador porscriptem

aproximadamente 88% dos casos.Além disso, identificou-se oportunidades de melhoria nesse processo de extração das características,fator fundamental para implementação do modelo de classificação dos documentos financeiros.

Adicionalmente, neste trabalho estudamos métodos para promover a escalabilidade do HDBSCAN*. Para isso, mostramos que é possível escalar as partes fundamentais do algoritmo, o gráfico de vizinhança e a MST, por meio de uma implementação aproximada. Os experimentos mostraram que algoritmo permite a manutenção de grande parte da qualidade do algoritmo original, com a vantagem de ser passível de escalabilidade.

Dentre possíveis trabalhos futuros, temos o estudo de localizadores customizados com base nas características predominantes nos documentos de cada agrupamento, com o objetivo de auferir a melhoria do processo de extração automática de dados. Além disso, sugere-se o refinamento das características relativas ao balanço patrimonial, em paralelo ao aprimoramento do localizador por *script*, além de maior exploração de características exclusivas da demonstração de resultado do exercício (DRE), a fim de avaliar o potencial de resultados de agrupamentos baseados em densidade gerados isoladamente para essas peças contábeis.

AGRADECIMENTOS

Agradecimentos especiais à FAPESP (Grant 2019/09817-6), à Serasa Experian e à UFSCar, por incentivarem, apoiarem e disponibilizarem os recursos necessários para a realização deste trabalho. Parcerias como essa destacam a importância do engajamento entre o setor privado e a academia.

REFERÊNCIAS

ARAUJO NETO, A. C. et al. **Efficient computation and visualization of multiple density-based clustering hierarchies**. IEEE Transactions on Knowledge and Data Engineering, 2019.

ARAUJO NETO, A. C. et al. **Mustache: A multiple clustering hierarchies explorer**. Proc. VLDB Endow. 11 (12): 2058–2061, Agosto, 2018.

ASSAF NETO, A. **Estrutura e análise de balanços: um enfoque econômico-financeiro**. Atlas, 2020.

BADER, D. A.; CONG, G. **Fast shared-memory algorithms for computing the minimum spanning forest of sparse graphs**. Journal of Parallel and Distributed Computing, v. 66, n. 11, p. 1366–1378, 2006. ISSN 0743-7315.

BANCO CENTRAL DO BRASIL. **Diagnóstico da convergência às Normas Internacionais: IAS 1 - Presentation of financial statements**. Banco Central do Brasil, Brasília/DF, 2010.

BRASIL. **Lei nº 6.404, de 15 de dezembro de 1976**. Diário Oficial da União, 1976.

BRASIL. **Lei nº 11.638, de 28 de dezembro de 2007**. Diário Oficial da União, 2007.

BRASIL. **Lei nº 11.941, de 27 de maio de 2009**. Diário Oficial da União, 2009.

CAMPELLO, R. J. G. B. et al. **A framework for semi-supervised and unsupervised optimal extraction of clusters from hierarchies**. Data Mining and Knowledge Discovery 27 (3): 344–371, Novembro, 2013.

CAMPELLO, R. J. G. B.; MOULAVI, D.; SANDER, J. **Density-based clustering based on hierarchical density estimates**. Lecture Notes in Computer Science, vol. 7819. Springer, pp. 160–172, 2013.

COMITÊ DE PRONUNCIAMENTOS CONTÁBEIS. **Pronunciamento Técnico CPC 21 (R1): Demonstração intermediária: Correlação às Normas Internacionais de Contabilidade - IAS 34 (IASB - BV 2011)**. Comitê de Pronunciamentos Contábeis, Brasília/DF, 2011a.

COMITÊ DE PRONUNCIAMENTOS CONTÁBEIS. **Pronunciamento Técnico CPC 26 (R1): Apresentação das demonstrações contábeis: Correlação às Normas Internacionais de Contabilidade - IAS 1 (IASB - BV 2011)**. Comitê de Pronunciamentos Contábeis, Brasília/DF, 2011b.

DEAN, J.; GHEMAWAT, S. **Mapreduce: Simplified data processing on large clusters**. In Proceedings of the 6th Conference on Symposiumon Operating Systems Design amp; Implementation - Volume 6, ser. OSDI'04. USA: USENIX Association, 2004, p. 10

DONG, W.; MOSES, C.; LI, K. **Efficient k-nearest neighbor graph construction for generic similarity measures**. In: Proceedings of the 20th International Conference on World Wide Web. New York, NY, USA: Association for Computing Machinery, 2011. (WWW '11), p. 577586. ISBN 9781450306324.

FERRAZ, V. et al. **Improving automatic data extraction from financial statements with clustering analysis**. In: SYMPOSIUM ON KNOWLEDGE DISCOVERY, MINING AND LEARNING (KDMILE), 8. , 2020, Evento Online. Anais [...]. Porto Alegre: Sociedade Brasileira de Computação, 2020. p. 1-8. ISSN 2763-8944. DOI: https://doi.org/10.5753/kdmile.2020.11952.

JAIN, A. K. **Data clustering: 50 years beyond k-means**. Pattern Recognition Letters 31 (8): 651 – 666, 2010.

JOHNSON, D. et al. **Comprehensive cross-hierarchy cluster agreement evaluation**, 2013.

MADEIRA, R. d. O. C. **Aplicação de técnicas de mineração de texto na detecção de discrepâncias em documentos fiscais**. M.S. thesis, Fundação Getúlio Vargas, Rio de Janeiro/RJ, 2015.

MOURA, M. F. **Proposta de utilização de mineração de textos para seleção, classificação e qualificação de documentos**. Tech. rep., Embrapa Informática Agropecuária. Dez., 2004.

SANTOS, J. A. d. et al. **Hierarchical density-based clustering using mapreduce**. IEEE Transactions on Big Data, vol. 7, no. 1, pp. 102–114, 2021.

SHIMOMURA, L. C. et al. **A survey on graph-based methods for similarity searches in metric spaces**. Information Systems, vol. 95, p. 101507, 2021.

SNOW, M. **Unsupervised document clustering with cluster topic identification**. Tech. rep., Office for National Statistics. Abr., 2018.

WARASHINA, T. et al. **Efficient k-nearest neighbor graph construction using mapreduce for large-scale data sets**. IEICE Transactions on Information and Systems, E97.D, n. 12,p. 3142–3154, 2014.

ÍNDICE REMISSIVO

U

V

www.atenaeditora.com.br
contato@atenaeditora.com.br
@atenaeditora
www.facebook.com/atenaeditora.com.br

Collection:

APPLIED COMPUTER ENGINEERING

Atena Editora
Ano 2022

www.atenaeditora.com.br
contato@atenaeditora.com.br
@atenaeditora
www.facebook.com/atenaeditora.com.br

Collection:

APPLIED COMPUTER ENGINEERING

Atena Editora
Ano 2022