UNIVERSIDADE DE SÃO PAULO
Instituto de Ciências Matemáticas e de Computação

Identificação de conceitos a partir de relatórios gerados pelo scriptLattes

Artur José Ferro Sampaio

Monografia - MBA em Inteligência Artificial e Big Data

ICMC USP SÃO CARLOS

Artur José Ferro Sampaio

Identificação de conceitos a partir de relatórios gerados pelo scriptLattes

Monografia apresentada ao Departamento de Ciências de Computação do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo - ICMC/USP, como parte dos requisitos para obtenção do título de Especialista em Inteligência Artificial e Big Data.

Área de concentração: Inteligência Artificial

Orientador: Prof. Dr. Diego Raphael Amancio

Versão original

São Carlos

2023

Ficha catalográfica elaborada pela Biblioteca Prof. Achille Bassi, ICMC/USP, com os dados fornecidos pelo(a) autor(a)

Sampaio, Artur José Ferro
S192 Identificação de conceitos a partir de relatórios gerados pelo scriptLattes / Artur José Ferro Sampaio ; orientador Prof. Dr. Diego Raphael Amancio. – São Carlos, 2023.
41 p. : il. (algumas color.) ; 30 cm.

Monografia (MBA em Inteligência Artificial e Big Data) – Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2023.

1. LaTeX. 2. abnTeX. 3. Classe USPSC. 4. Editoração de texto. 5. Normalização da documentação. 6. Tese. 7. Dissertação. 8. Documentos (elaboração). 9. Documentos eletrônicos. I. Amancio, Diego Raphael, orient. II. Título.

Artur José Ferro Sampaio

Identification of concepts from reports generated by scriptLattes

Monograph presented to the Departamento de Ciências de Computação do Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo - ICMC/USP, as part of the requirements for obtaining the title of Specialist in Artificial Intelligence and Big Data.

Concentration area: Artificial Intelligence

Advisor: Prof. Dr. Diego Raphael Amancio

Original version

São Carlos

2023

Este trabalho é dedicado aos funcionários do ICMC/USP, que contribuem incessantemente para o desenvolvimento das ciências e, portanto, para o desenvolvimento social do país.

AGRADECIMENTOS

Agradeço à Profa. Dra. Solange Oliveira Rezende por ser exemplo de dedicação ao desenvolvimento das ciências e eterna incentivadora da evolução pessoal.

Agradeço ao Prof. Dr. Diego Raphael Amancio pela sua orientação no presente trabalho, desempenhando papel fundamental no progresso e fluidez das atividades.

Agradeço aos colegas Carlos Eduardo Favaro e Cassio Henrique Jorge pelo seu empenho, colaboração e incentivo ao longo deste MBA.

Agradeço aos amigos Willian Dener de Oliveira e Igor Vitório Custódio pelos constantes debates, trocas de ideias e sugestões, que contribuíram muito para a realização do presente trabalho.

Agradeço ao Instituto de Ciências Matemáticas e de Computação, na pessoa do seu diretor Prof. Dr. André Carlos Ponce de Leon Ferreira de Carvalho, por fomentar e incentivar o desenvolvimento profissional dos funcionários.

Agradeço ainda à Universidade de São Paulo, na pessoa do Magnífico Reitor Prof. Dr. Carlos Gilberto Carlotti Junior.

"A criação bem-sucedida de inteligência artificial seria o maior evento na história da humanidade. Infelizmente, pode também ser o último, a menos que aprendamos a evitar os riscos."

Stephen Hawking

RESUMO

SAMPAIO, A.J.F. **Identificação de conceitos a partir de relatórios gerados pelo scriptLattes**. 2023. 41p. Monografia (MBA em Inteligência Artificial e Big Data) - Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2023.

O presente trabalho tem por objetivo estudar as ferramentas atuais de extração e processamento de informações da Plataforma Lattes, buscando integração com outras bases de dados científicas. A partir de relatórios gerados com a ferramenta *scriptLattes* foi possível recuperar informações adicionais na plataforma OpenAlex e assim gerar visualizações adicionais. Especialistas do domínio participaram na avaliação das visualizações geradas, validando sua relevância. Como resultado foi criada uma extensão à ferramenta *scriptLattes*, ampliando suas funcionalidades. Concluiu-se que é possível estender ferramentas de cientometria já existentes utilizando outras fontes de dados e aplicando técnicas de Inteligência Artificial e Big Data.

Palavras-chave: DOI. Lattes. Cientometria. Dissertação. scriptLattes. OpenAlex. Visualização. Trabalho de conclusão de curso (TCC).

LISTA DE FIGURAS

Figura 1 – Metodologia de trabalho 26
Figura 2 – Diagrama da coleta de dados 26
Figura 3 – Modelo de dados da plataforma OpenAlex 28
Figura 4 – Histograma de *concepts* 29
Figura 5 – Boxplot de *concepts* 29
Figura 6 – Distribuição de concepts após filtragem 30
Figura 7 – Concepts com maiores pontuações 31
Figura 8 – Nuvem de palavras 32
Figura 9 – Concepts por ano - Gráfico de barras verticais empilhadas 32
Figura 10 – Concepts por ano - Gráfico de barras horizontais empilhadas 33
Figura 11 – Concepts por ano - Gráfico de área empilhada 33
Figura 12 – Concepts por ano - wireframe 3D 34
Figura 13 – Concepts wireframe 2 35
Figura 14 – Concepts por ano - lâminas 3D 36
Figura 15 – Concepts por ano - Barras verticais agrupadas 37
Figura 16 – Concepts por ano - Barras verticais independentes 38

LISTA DE TABELAS

Tabela 1 – Avaliações dos gráficos 37

LISTA DE ABREVIATURAS E SIGLAS

DOI	Digital Object Identifier
CNPq	Conselho Nacional de Desenvolvimento Científico e Tecnológico
USP	Universidade de São Paulo
ICMC	Instituto de Ciências Matemáticas e de Computação
HTML	HyperText Markup Language
PLN	Processamento de Língua Natural
IA	Inteligência Artificial
AM	Aprendizado de Máquina

SUMÁRIO

1 INTRODUÇÃO 21

2 TRABALHOS RELACIONADOS 23
2.1 Scriptlattes 23
2.2 OpenAlex 23

3 DESENVOLVIMENTO 25
3.1 Materiais e métodos 25
3.1.1 Nuvem de palavras 25
3.1.2 Metodologia de trabalho 25
3.2 Identificação do problema 25
3.3 Pré processamento 26
3.3.1 Coleta de DOI´s de relatórios gerados pelo scriptLattes 26
3.3.2 Captura de concepts rotulados pela plataforma OpenAlex 27
3.3.3 Análise exploratória do conjunto de dados 28
3.3.4 Redução de dimensionalidade 29
3.4 Extração de padrões 30
3.4.1 Identificação dos principais concepts 31
3.4.2 Análise de concepts ao longo do tempo 32
3.5 Pós processamento 33
3.6 Utilização do conhecimento 36
3.7 Resultados obtidos 37
3.8 Discussão 37

4 CONCLUSÃO 39

REFERÊNCIAS 41

1 INTRODUÇÃO

Mensurar a produção científica é um desafio para a evolução das ciências desde a antiguidade, mas foi Laplace (1749-1827) o primeiro a estudar sistematicamente os problemas teóricos levantados pela aplicação das medidas nas ciências de sua época (SILVA J.A. E BIANCHI, 2001).

Cientometria é definida como o estudo da mensuração do progresso científico e tecnológico e se baseia na avaliação quantitativa e na análise das colaborações, produtividade e progresso científico. Em outras palavras, a cientometria consiste em aplicar técnicas numéricas analíticas para estudar a ciência da ciência (SILVA J.A. E BIANCHI, 2001).

Na atualidade, a cientometria trata da mensuração e quantificação do desenvolvimento científico baseada em indicadores bibliométricos extraídos das diferentes publicações científicas refletidas em artigos, livros e em revistas científicas editadas, e é de grande interesse para governos e instituições de pesquisa direcionarem seus esforços e programas de investimento. Essas informações também alimentam pesquisas em outras áreas de conhecimento, tais como historiadores, sociólogos e outros pesquisadores interessados na evolução da ciência (SILVA J.A. E BIANCHI, 2001).

As pesquisas nessa área tipicamente envolvem o tratamento de conjuntos de dados massivos e com conteúdos diversos e complexos, características que inviabilizam o processamento por humanos (MENA-CHALCO; M., 2009).

Realizar uma compilação ou sumarização de produções bibliográficas para um grupo de pesquisadores populoso pode exigir um grande esforço manual, largamente suscetível a falhas (MENA-CHALCO; M., 2013).

A Plataforma Lattes é resultado do esforço do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) na integração de bases de dados de Currículos, de Grupos de pesquisa e de Instituições em um único Sistema de Informações, e se tornou estratégica não só para as atividades de planejamento e gestão, mas também para a formulação das políticas do Ministério de Ciência e Tecnologia e Inovação e de outros órgãos governamentais relacionados ao desenvolvimento científico (CPNq, 2023). O chamado Curriculo Lattes é portanto considerado o padrão nacional de informações sobre realizações científicas e acadêmicas de estudantes, professores e pesquisadores, representando seu histórico profissional de atividades (MENA-CHALCO; M., 2009). Os registros de atividades na Plataforma Lattes são agrupados por indivíduo e não oferecem mecanismo para geração de relatórios referentes a grupos de pesquisadores, como departamentos, instituições ou núcleos de pesquisa.

Ferramentas voltadas à extração e compilação automatizadas de informações da

plataforma Lattes são objeto de estudo no Brasil há mais de dez anos, e algumas obtiveram sucesso na geração de relatórios, como o scriptLattes (MENA-CHALCO; M., 2009). Entretanto as informações contidas na plataforma dependem de alimentação manual dos próprios pesquisadores, na forma de texto editável, e portanto suscetível a falhas de digitação, incompletude dos dados e falta de padronização.

Nos últimos anos os esforços globais em busca de uma base sólida de registros de produção científica levaram à criação do Projeto OpenAlex, um catálogo aberto que reúne mais de duzentos milhões de trabalhos científicos, associados a seus autores e instituições (PRIEM J.; PIWOWAR, 2022). Processamento de Língua Natural (PLN) é uma subárea da Inteligência Artificial (IA) que visa habilitar as máquinas a lidar com a linguagem humana, realizando tarefas como sumarização de textos (RASTEIRO B. H.; MONTEIRO, 2017 apud JURAFSKY D.; MARTIN, 2009). Ferramentas e técnicas de PLN podem ser aplicadas a grandes volumes de dados para observar similaridades ou correlações, o que evidentemente gera enorme contribuição às atividades da cientometria.

Observando-se os recentes progressos apontados acima apresenta-se uma janela de oportunidade para revisitar o processamento automatizado de currículos da Plataforma Lattes, com vistas ao aperfeiçoamento e aprimoramento dos resultados entregues pelas ferramentas atuais, através da aplicação de técnicas modernas de PLN.

O presente trabalho tem por objetivo estudar as ferramentas atuais de extração e processamento de informações da Plataforma Lattes e a partir daí propor aprimoramentos por meio da aplicação de técnicas e ferramentas de IA e PLN, buscando integração com outras bases de dados e exploração do conteúdo completo das publicações, quando disponíveis.

2 TRABALHOS RELACIONADOS

Trabalhos anteriores já se ocuparam da exploração de dados associados a publicações científicas, trazendo relevantes contribuições à área. O presente trabalho agrega novos recursos em uma ferramenta já existente denominada *scriptLattes*, a partir da coleta de dados adicionais em uma plataforma online também já existente denominada OpenAlex. Ao agregar informações de diferentes fontes tornou-se possível disponibilizar novas funcionalidades para os usuários do *scriptLattes*.

2.1 Scriptlattes

O *scriptLattes* é uma ferramenta de software livre desenvolvida para a extração e compilação automática de listas de produções acadêmicas a partir das informações coletadas na Plataforma Lattes (MENA-CHALCO; M., 2009). A partir de uma lista de membros pré-selecionados o software faz o download e extração automatizados dos currículos Lattes e constrói relatórios de saída na forma de páginas navegáveis no formato HTML.

Algum tempo depois da criação do *scriptLattes* o CNPq mudou a política de acesso a currículos da plataforma Lattes, impedindo o download automatizado das informações. Isso prejudicou fortemente a funcionalidade do software. O próprio CNPq oferece a possibilidade de extração de dados utilizando uma ferramenta denominada Extrator Lattes, mas que exige um fluxo de cadastro e autorização [GOVBR, 2023], e portanto não é totalmente aberto e disponível para qualquer indivíduo acessar livremente.

Apesar de extremamente útil e funcional, o *scriptLattes* apresenta ao menos duas limitações conhecidas, que uma vez sanadas poderiam potencializar o uso da ferramenta:

- Não gera uma lista das áreas de conhecimento tratadas pelas publicações processadas. Mesmo gerando uma lista das publicações produzidas, a ferramenta é incapaz de identificar os tópicos e/ou áreas de conhecimento abordadas.

- Não oferece mecanismo de comparação entre grupos. Uma vez que um grupo tenha sido gerado, não há mecanismos para compreender a similaridade com outros grupos.

2.2 OpenAlex

OpenAlex, cujo nome foi inspirado pela biblioteca ancestral de Alexandria, é um catálogo aberto e compreensivo de publicações acadêmicas, autores, instituições e mais, composto por centenas de milhões de entidades interconectadas [OpenAlex, 2023]. Trata-se

de uma base de dados científicos global e de acesso fácil, contando com interface de consultas bem definida.

A construção do catálogo é feita de forma automatizada, utilizando técnicas de PLN para analisar o conteúdo dos trabalhos, evitando portanto a contaminação de informações digitadas por humanos.

3 DESENVOLVIMENTO

Este capítulo descreve a execução do trabalho, descrevendo as operações realizadas durante o seu desenvolvimento.

3.1 Materiais e métodos

3.1.1 Nuvem de palavras

Nuvem de palavras é uma forma de representação visual de palavras, capaz de destacar os termos de acordo com suas ocorrências em determinado contexto. Nuvem de palavras é um tipo de visualização ponderada de uma lista de dados, de forma que os elementos com maior peso recebam maior destaque na visualização. O resultado é uma imagem de fácil compreensão por humanos.

3.1.2 Metodologia de trabalho

Para o desenvolvimento do trabalho foi adotada a metodologia apresentada durante o curso (REZENDE, 2003), cujo diagrama pode ser observado na Figura 1. As seções seguintes desde capítulo espelham o modelo proposto pela metodologia.

- Identificação do problema: definir os objetivos do trabalho
- Pré-processamento: obter o conjunto de dados e realizar os tratamentos necessários (filtragem, normalização, conversões)
- Extração de padrões: aplicar técnicas de AM sobre o conjunto de dados visando atingir os objetivos propostos
- Pós-processamento: interpretação dos resultados
- Utilização do conhecimento: aplicação prática dos resultados

3.2 Identificação do problema

O scriptLattes é capaz de gerar relatórios quantitativos a partir de uma lista de pesquisadores, mas sua única fonte de dados (plataforma Lattes) identifica as áreas de conhecimento abordadas apenas através de informação declarada pelo próprio pesquisador. Assim, não se tem nos relatórios produzidos informações consistentes dos temas sobre os quais as publicações versam.

Portanto se estabeleceu como objetivo primário: estender os relatórios produzidos pelo scriptLattes, adicionando informações sobre as principais áreas de conhecimento

Figura 1 – Metodologia de trabalho

abordadas. Esses rótulos devem ser atribuídos de forma automatizada, considerando o conteúdo dos trabalhos processados.

3.3 Pré processamento

A etapa de pré processamento consistiu na obtenção do conjunto de dados, análise exploratória desse conjunto e tratamento dos dados.

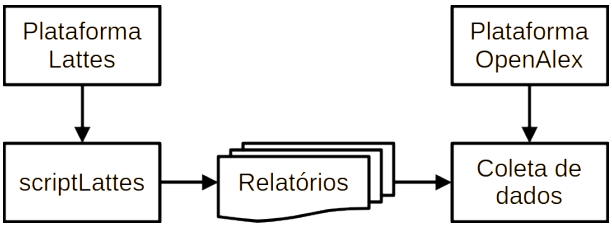

Figura 2 – Diagrama da coleta de dados

3.3.1 Coleta de DOI´s de relatórios gerados pelo scriptLattes

Dentre os vários relatórios gerados como saída pelo scriptLattes, um deles lista todos os trabalhos científicos publicados pelo grupo de pesquisadores indicados. Essa lista exibe os resultados em formato semelhante a referências bibliográficas, e portanto contém título do trabalho, ano de publicação, lista de autores, entre outros atributos. Para os objetivos do presente trabalho, o atributo mais relevante dessa lista é a URL do DOI, que representa um identificador único de cada trabalho.

Para o desenvolvimento do presente trabalho foi utilizado um relatório gerado pelo ScriptLattes para um grupo de pesquisadores vinculados a uma instituição de ensino superior especializada nas áreas de ciências de computação, matemática e estatística. Isso

significa que é esperado que os trabalhos tenham correlação prioritária com essas áreas de conhecimento.

O relatório processado é representado por páginas no formato HTML, composto por conteúdo textual, gráficos, elementos visuais e outras funcionalidades adequadas para navegação web. Uma vez que o objetivo seria capturar apenas os identificadores DOI, as páginas coletadas foram processadas com o uso de expressões regulares, extraindo apenas as URL´s dos DOI associadas. Essa estratégia gerou uma lista de entrada composta de 3314 identificadores DOI, que correspondem, portanto, ao mesmo número de trabalhos únicos.

3.3.2 Captura de concepts rotulados pela plataforma OpenAlex

Os vários endereços DOI servem como chave de busca na plataforma OpenAlex, identificando unívocamente cada trabalho catalogado na plataforma. Utilizando essa informação podem ser realizadas requisições para a API da plataforma OpenAlex, submetendo como único parâmetro o endereço DOI. A plataforma então devolve como resposta um objeto JSON com todas as características do trabalho em questão, incluindo lista de autores, ano de publicação, editora, páginas, etc. O diagrama exibido na Figura 3 ilustra o modelo de objetos adotado pela plataforma.

A maioria desses atributos já era conhecida anteriormente no escopo deste trabalho, a partir do relatório gerado pelo scriptLattes. Mas essa resposta inclui também uma lista de *concepts* associados a pesos normalizados. Todo o desenvolvimento do presente trabalho se concentra na exploração e utilização desses *concepts* na forma de atributos dos trabalhos.

Os *concepts* são ideias abstratas (ou rótulos) sobre os quais os trabalhos tratam, e são organizadas em uma estrutura hierárquica na forma de árvore. Existem 19 *concepts* no primeiro nível, e outras 6 camadas de descendentes a partir deles, resultando em 65 mil rótulos no total (OPENALEX, 2023). Essa estrutura é baseada no modelo proposto por Zhihong S.; Hao (2018).

Cada trabalho é associado a múltiplos *concepts* pela plataforma OpenAlex a partir de seu título, resumo e veículo de publicação, tarefa realizada por um classificador treinado utilizando um *corpus* especializado (OPENALEX, 2020).

Foi construído um script utilizando a linguagem Python, capaz de ler uma lista de URL´s de DOI´s e, para cada elemento, disparar consulta na plataforma OpenAlex para obter dados sobre aquele trabalho. Dos 3314 identificadores consultados a plataforma OpenAlex não encontrou resultados para 17 deles, valor que representa aproximadamente 0,51% do total. Esses registros foram ignorados e portanto removidos do escopo deste trabalho.

A técnica de coleta descrita resultou em uma tabela de dados composta por

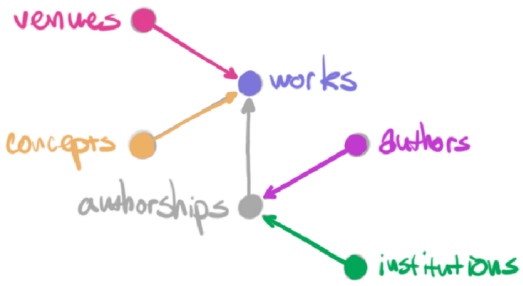

Figura 3 – Modelo de dados da plataforma OpenAlex

3297 linhas e 4891 colunas. Cada linha corresponde a um trabalho único, enquanto as colunas correspondem às características dos trabalhos. As quatro primeiras colunas armazenam, respectivamente, o DOI do trabalho, o título, o ano de publicação e o número de citações, enquanto as demais colunas representam os vários *concepts* obtidos na plataforma OpenAlex.

3.3.3 Análise exploratória do conjunto de dados

A análise exploratória dos dados tem por objetivo examinar e estudar as características de um conjunto de dados, e pode envolver descrições estatísticas como contagem de registros, valores médios e medianas das colunas, bem como representações gráficas que facilitem a interpretação inferencial.

Considerando que a informação escolhida como mais relevante neste trabalho são os *concepts*, é exibido na Figura 4 um histograma demonstrando a frequência de cada *concept* associado às publicações. Para evitar poluição visual desnecessária foram exibidos apenas os 200 resultados mais representativos (com maior valor).

É possível observar que alguns *concepts* apresentam grande quantidade de ocorrências, enquanto que muitos deles tem apenas 1 única ocorrência.

Também foi elaborado um gráfico *boxplot*, apresentado na Figura 5, para avaliar a distribuição dos *concepts*, destacando os *outliers* e quartis. Novamente é possível observar a enorme incidência de *outliers*, demonstrando que muitos *concepts* ocorrem com baixa frequência

O resultado da extração e preparação de dados foi construído em um *Pandas Dataframe*, que por sua vez foi armazenado na forma de um arquivo CSV para posterior processamento, tirando vantagem da reutilização e evitando que a coleta tivesse que ser refeita a cada passo na evolução do trabalho.

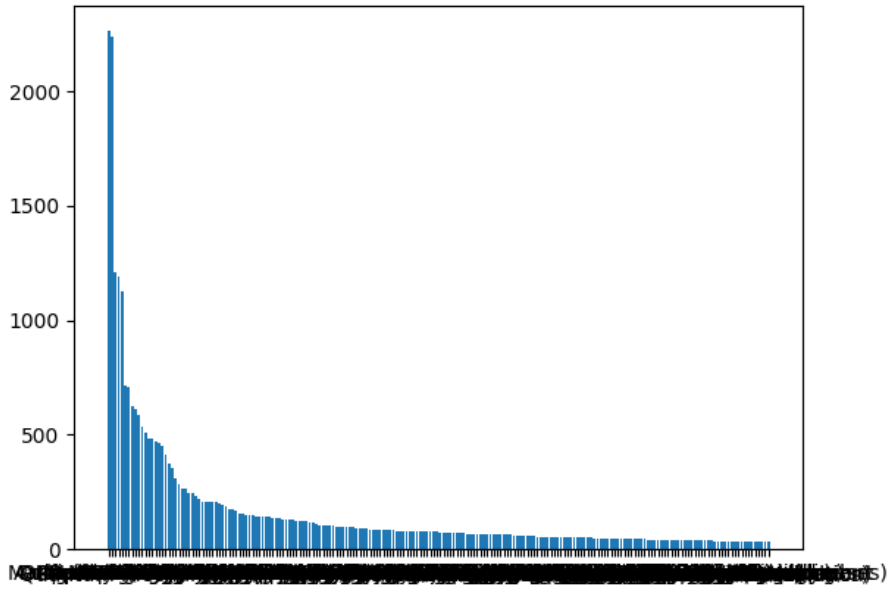

Figura 4 – Histograma de *concepts*

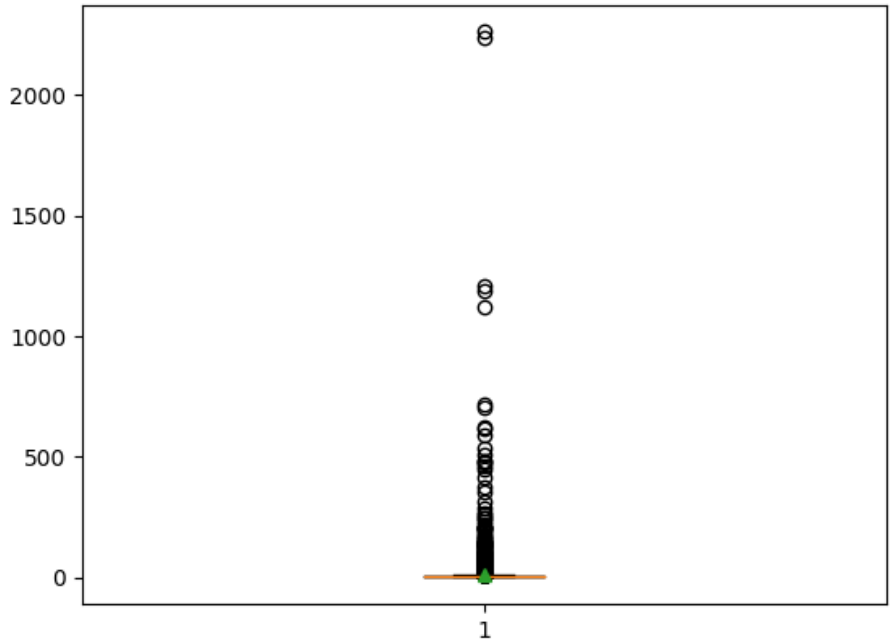

Figura 5 – Boxplot de *concepts*

3.3.4 Redução de dimensionalidade

Conforme já descrito, a massa de dados construída conta com grande quantidade de atributos (4.891), levando a um problema denominado *maldição da dimensionalidade* (BELLMAN, 1957). É sabido que esse problema causa degradação da performance dos algoritmos conforme o aumento do volume de dados tratados. Também são conhecidas técnicas de redução de dimensionalidade, que buscam mitigar o problema descrito e serão aplicadas ao longo deste trabalho.

A técnica de filtragem de dados consiste na eliminação de linhas e/ou colunas de menor representatividade na massa de dados. Essas linhas e colunas podem ser selecionadas com base na contagem de valores ou com base em métricas como média ou mediana dos atributos.

Para a massa de dados em análise foi escolhida com métrica a soma dos valores por coluna (ou *concepts*), pois essa técnica leva em consideração também a quantidade de trabalhos associados a determinada área de conhecimento. Foram escolhidos os vinte *concepts* que apresentaram maior soma, concentrando-se então nas áreas de conhecimento mais discutidas no escopo dos trabalhos processados. Esse valor foi escolhido por apresentar resultados com volume de informações compreensível por usuários humanos, buscando um ponto de equilíbrio entre riqueza de informações e poluição visual. As colunas restantes foram descartadas. Com a filtragem de colunas restaram trabalhos sem nenhum *concept* associado, pois se concentravam em áreas menos representadas dentro do conjunto de dados. Esses trabalhos foram também descartados.

Após a aplicação desse filtro o *Pandas Dataframe* passou a ter 976 linhas e 24 colunas. A Figura 6 ilustra a frequência de cada *concept* após a aplicação dos filtros.

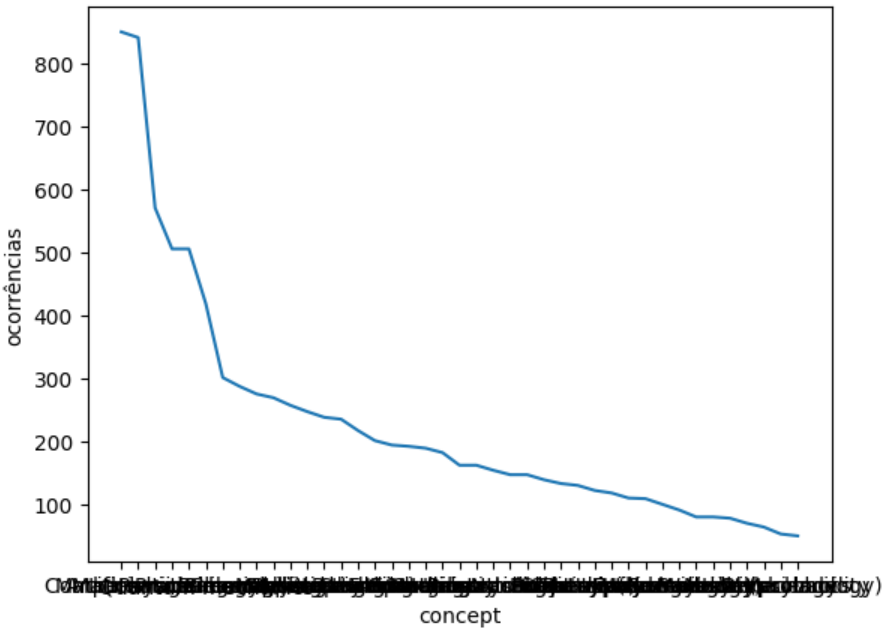

Figura 6 – Distribuição de concepts após filtragem

3.4 Extração de padrões

Conhecendo o conjunto de dados disponível teve início a aplicação de técnicas de AM para extrair padrões e torná-los acessíveis para os usuários finais.

3.4.1 Identificação dos principais concepts

Cada registro do conjunto de dados continha milhares de atributos, ao passo que o usuário que consome os relatórios só precisaria saber as principais áreas de conhecimento abordadas. Por isso foi utilizado o conjunto de dados gerado nos passos anteriores deste trabalho, com os filtros já aplicados, portanto considerando apenas os vinte *concepts* mais relevantes, identificados a partir da soma das pontuações de cada trabalho.

Sobre esse conjunto reduzido de dados foram aplicadas diferentes técnicas de visualização, permitindo que o usuário escolha as mais eficazes para seus propósitos e conjunto de dados específico.

Foi gerado um gráfico de barras simples para visualizar os 20 *concepts* com maior pontuação, que representam as principais áreas de conhecimento abordadas pelos trabalhos analisados. Esse gráfico pode ser visto na Figura 7.

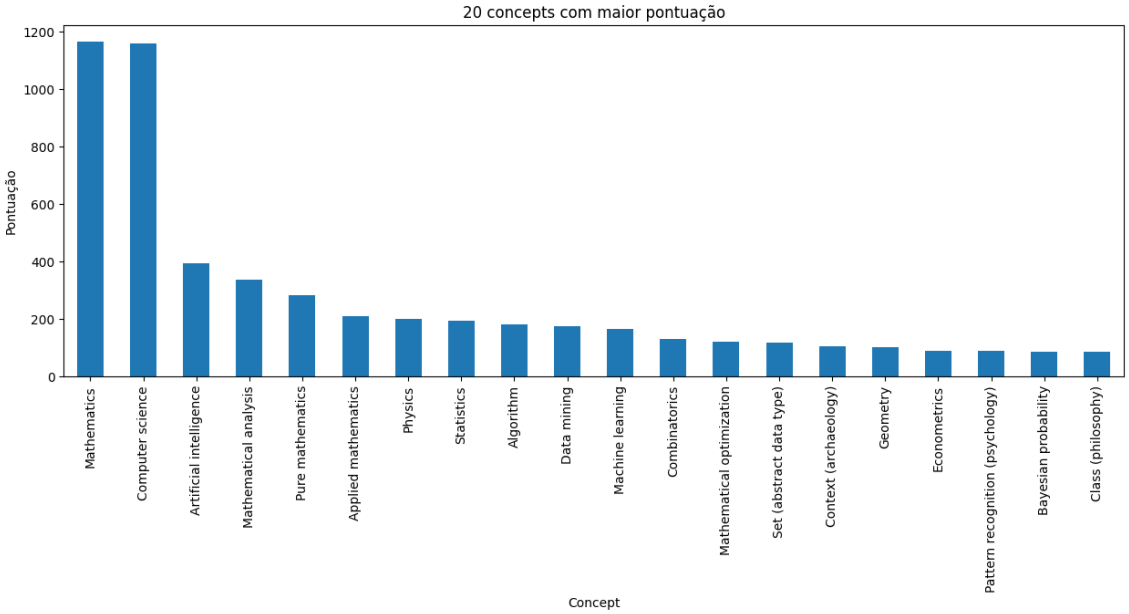

Figura 7 – Concepts com maiores pontuações

Apesar de conter as informações pertinentes e exibi-las de forma limpa e organizada, a figura ainda depende de algum esforço interpretativo do usuário, que deve ler os rótulos e compreender o significado dos elementos.

O uso de uma nuvem de palavras no contexto do presente trabalho ofereceu percepção visual sobre os *concepts* cobertos pelos trabalhos processados. Foi atribuído um valor a cada um dos *concepts*, calculado a partir da soma de todas as pontuações associadas. Esses valores foram utilizados para dar peso às palavras na construção da nuvem. O resultado pode ser visto na Figura 8.

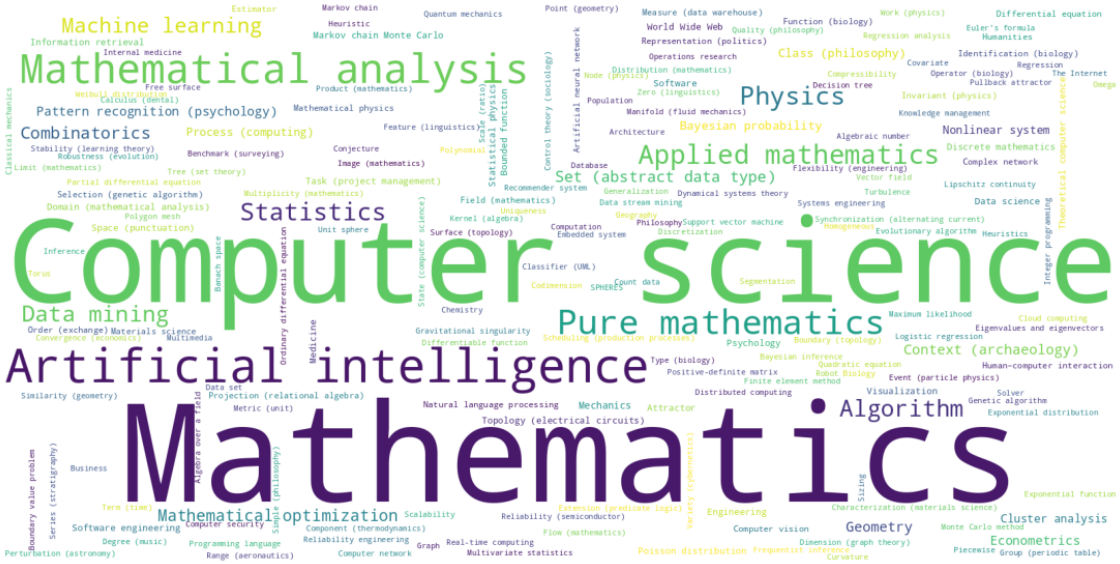

Figura 8 – Nuvem de palavras

3.4.2 Análise de concepts ao longo do tempo

O conjunto de dados utilizado contém o ano de publicação de cada trabalho, permitindo que sejam exploradas visualizações capazes de evidenciar as tendências ao longo do tempo.

O gráfico de barras verticais empilhadas demonstra a evolução dos *concepts* ao longo do tempo, destacando a participação de cada *concept* para o resultado final. Esse gráfico pode ser visto na Figura 9.

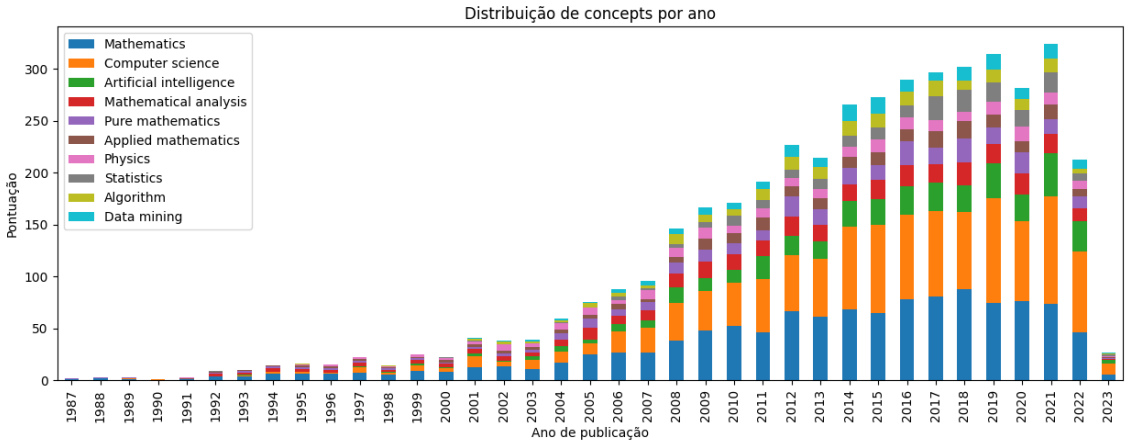

Figura 9 – Concepts por ano - Gráfico de barras verticais empilhadas

O mesmo gráfico de barras pode ser construído na orientação horizontal, que apesar de exibir exatamente as mesmas informações do gráfico de barras verticais, pode facilitar a percepção visual de algumas caracterísicas, em virtude da proporção das barras. Esse gráfico pode ser visto na Figura 10.

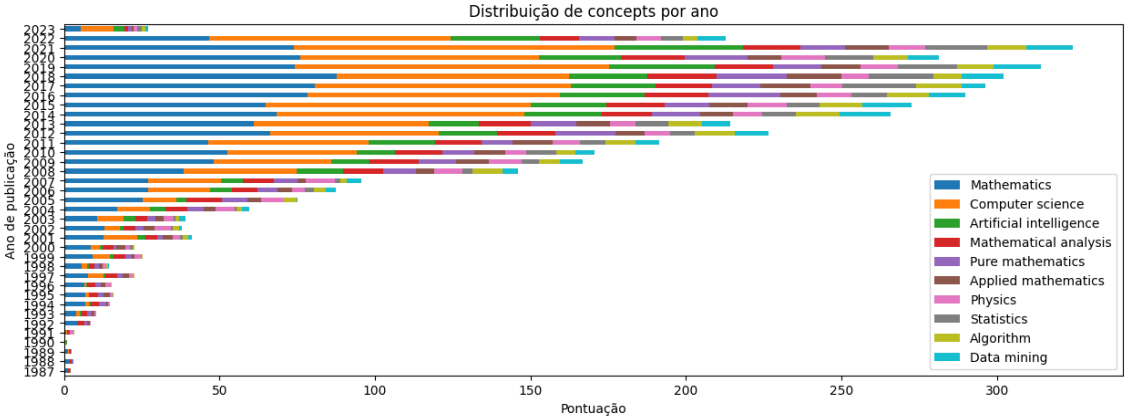

Figura 10 – Concepts por ano - Gráfico de barras horizontais empilhadas

Um gráfico de área empilhada é também muito parecido com o de barras verticais empilhadas, mas a ausência dos espaços em branco entre as barras pode afetar a percepção visual, causando sensação de fluidez. Esse gráfico pode ser visto na Figura 11.

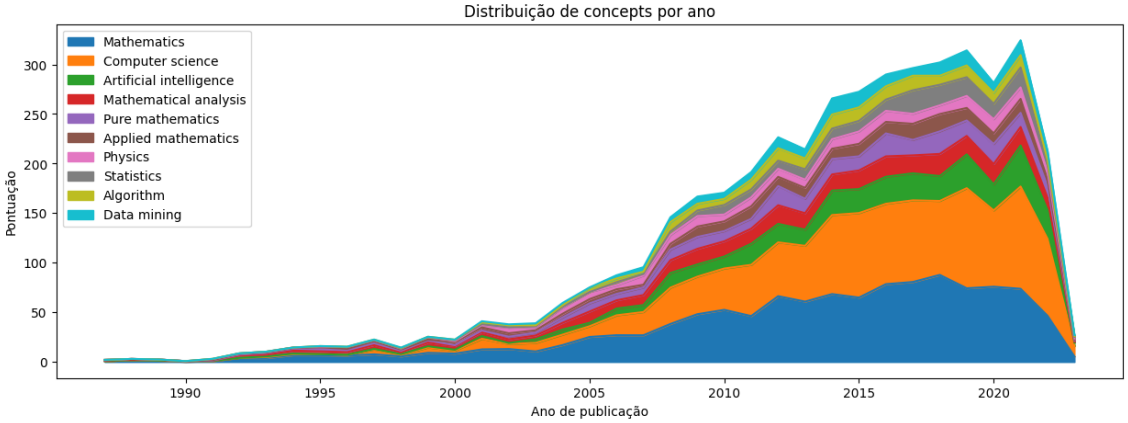

Figura 11 – Concepts por ano - Gráfico de área empilhada

Outras técnicas exploram a visualização em três dimensões, favorecendo a comparação entre *concepts* a cada ano. Gráficos desse tipo podem ser vistos nas Figuras 12, 13 e 14.

Outras formas de visualização podem destacar o desempenho de cada *concept* isolado ao longo do tempo, em vez de compara-los uns com os outros. Gráficos desse tipo podem ser vistos nas Figuras 15 e 16.

3.5 Pós processamento

Uma vez que os resultados produzidos foram tabelas e figuras representando diferentes visualizações das informações e que seu principal objetivo é oferecer melhor entendimento por parte dos usuários consumidores dos relatórios, foram convidados cinco

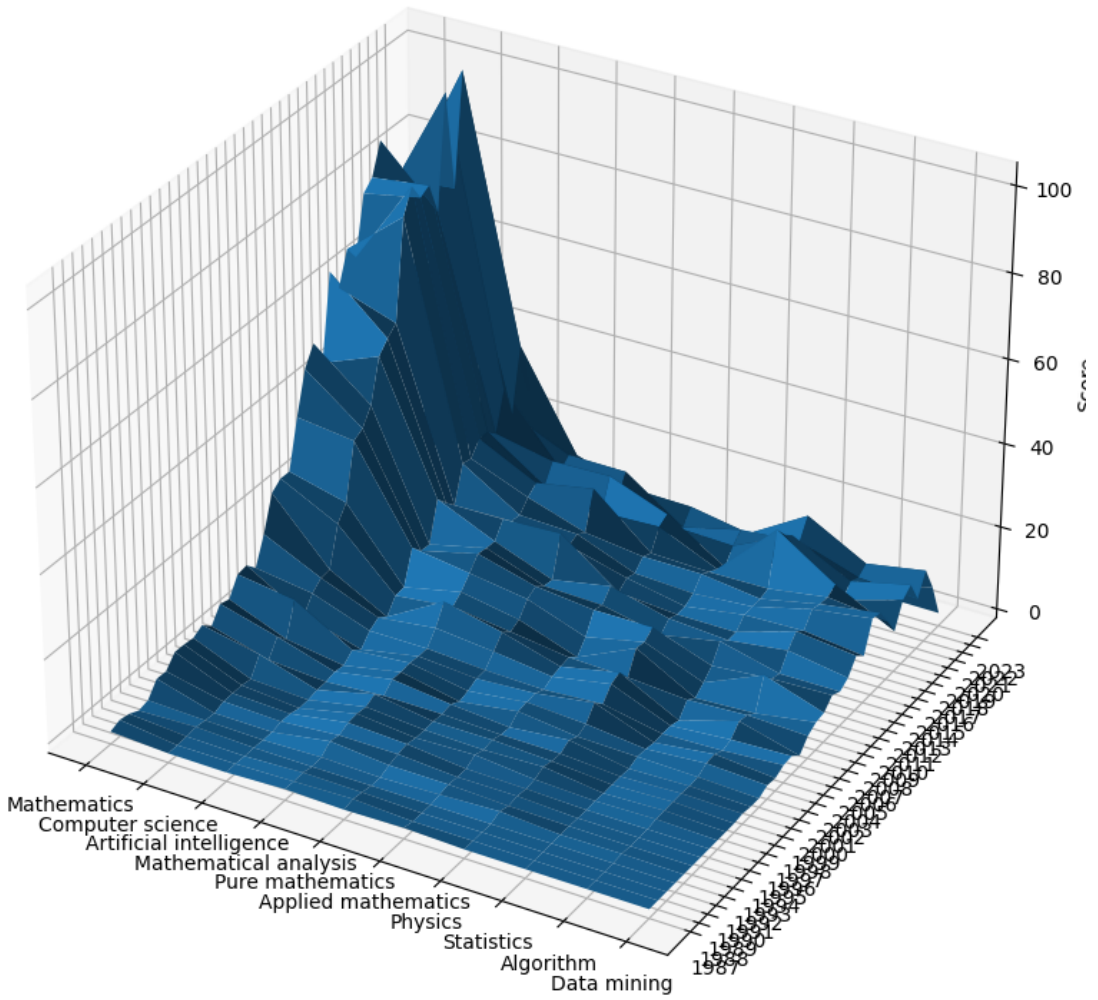

Figura 12 – Concepts por ano - wireframe 3D

usuários avançados do scriptLattes (denominados U1 até U5), especialistas na área, para avaliar a utilidade prática de cada visualização. Cada usuário atribuiu notas de relevância para cada visualização, no intervalo de 0 a 10, produzindo assim um ranqueamento baseado na soma das notas.

Para facilitar a tabulação dos resultados foi atribuído um identificador para cada visualização, da seguinte maneira:

- A - Figura 7 - Concepts com maiores pontuações
- B - Figura 8 - Nuvem de palavras
- C - Figura 9 – Concepts por ano - Gráfico de barras verticais empilhadas
- D - Figura 10 – Concepts por ano - Gráfico de barras horizontais empilhadas

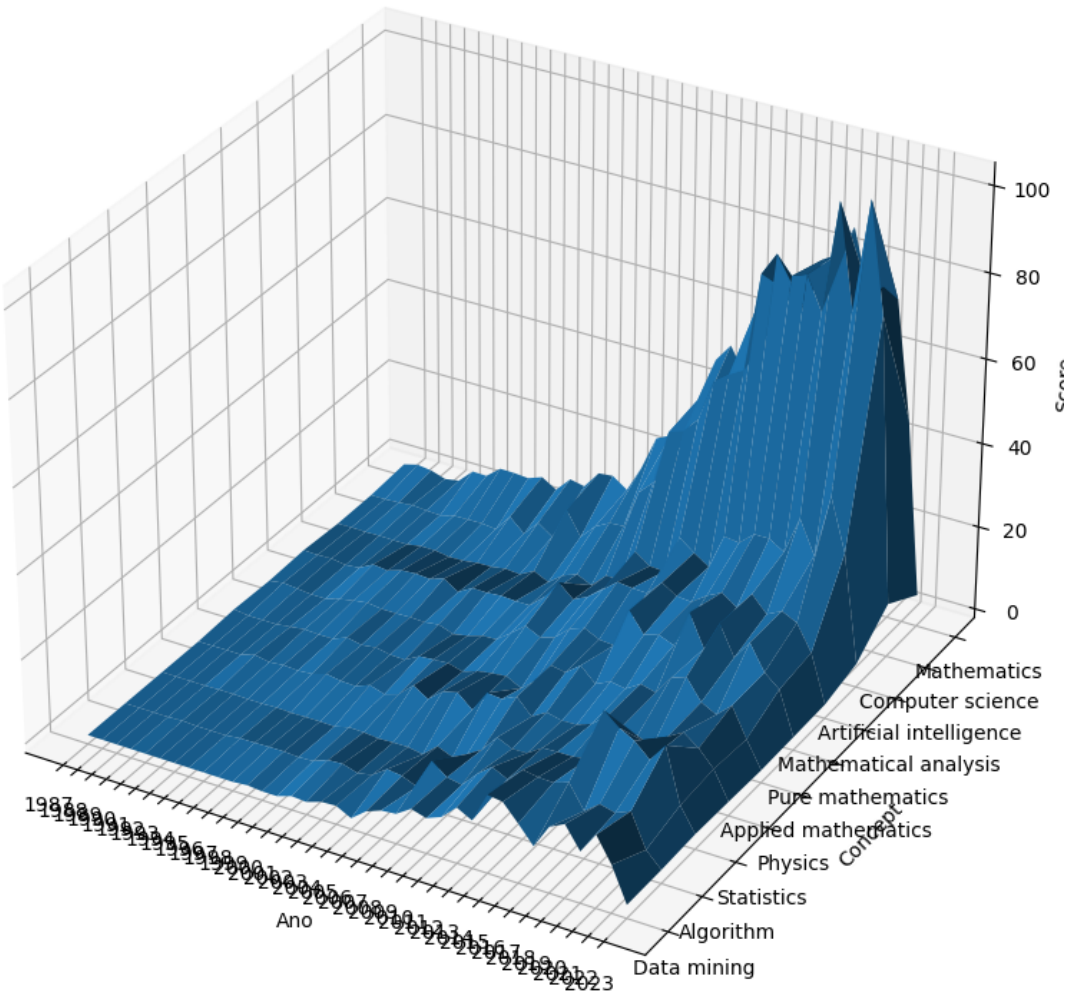

Figura 13 – Concepts wireframe 2

- E - Figura 11 – Concepts por ano - Gráfico de área empilhada
- F - Figura 12 – Concepts por ano - wireframe 3D
- G - Figura 13 – Concepts wireframe 2
- H - Figura 14 – Concepts por ano - lâminas 3D
- I - Figura 15 – Concepts por ano - Barras verticais agrupadas
- J - Figura 16 – Concepts por ano - Barras verticais independentes

Conforme resultados da pesquisa apresentados na Tabela 1, as visualizações consideradas mais relevantes pelos usuários foram as Figuras 8, 7, 9, 16 e 14. Portanto essas visualizações foram selecionadas como resultado final do presente trabalho.

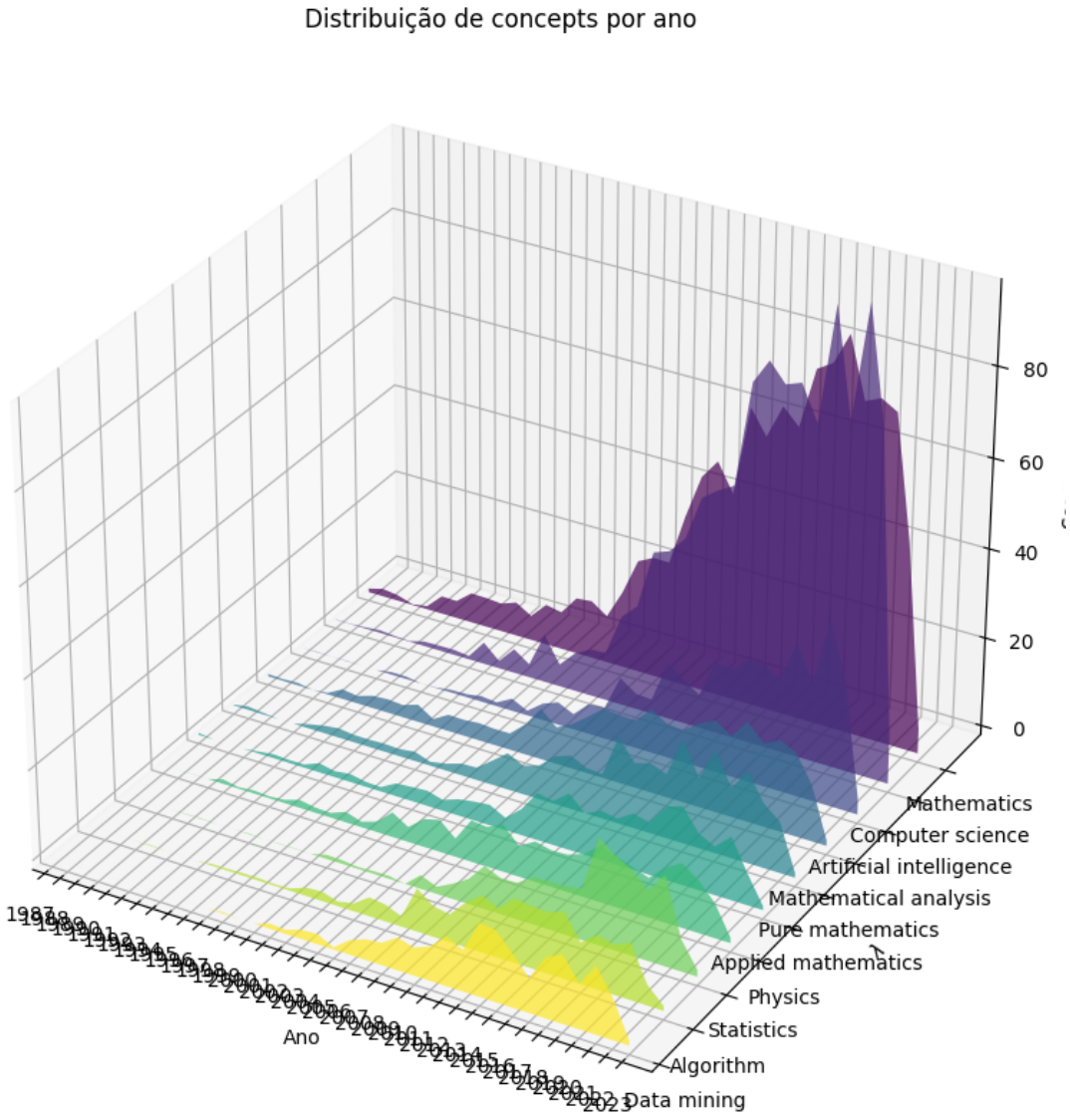

Figura 14 – Concepts por ano - lâminas 3D

3.6 Utilização do conhecimento

As visualizações que obtiveram melhor posição na classificação foram integradas aos relatórios produzidos pelo scriptLattes, tornando-se recurso padrão e estando disponível para outros usuários. A integração foi bastante simples, já que o *scriptLattes* é construído utilizando a linguagem de programação Python - a mesma adotada no desenvolvimento do presente trabalho.

O código fonte do scriptLattes foi modificado para incluir uma chamada adicional para um novo script ao final de seu processamento regular. Esse script realiza as operações descritas no presente trabalho e armazena as informações produzidas em arquivos localizados no mesmo diretório dos relatórios gerados pelo scriptLattes, tornando a visitação/consumo do conteúdo semelhante aos outros relatórios nativos da ferramenta.

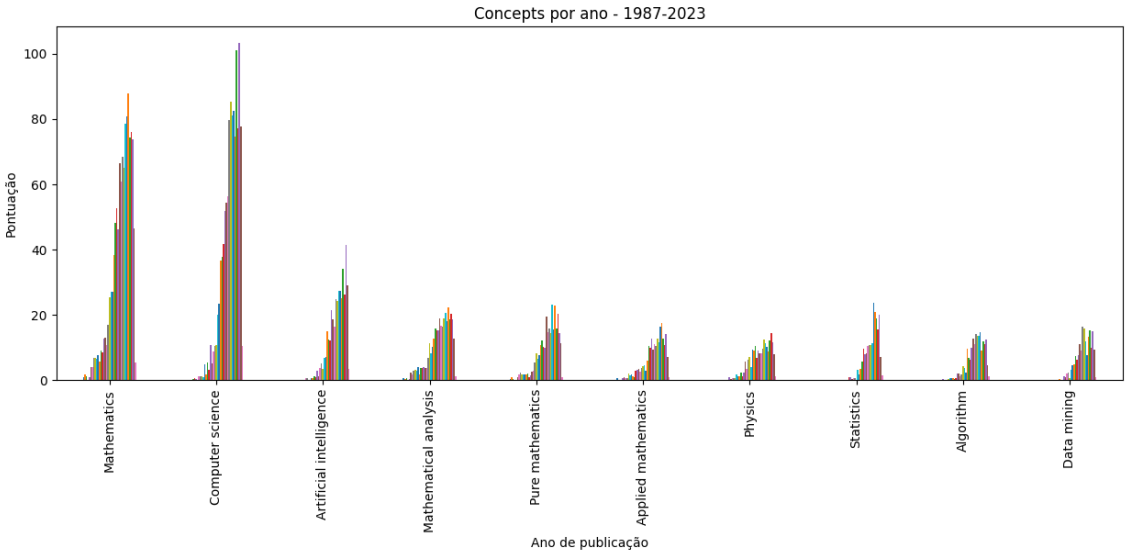

Figura 15 – Concepts por ano - Barras verticais agrupadas

	U1	U2	U3	U4	U5	Soma
A	7	10	10	0	10	37
B	10	9	10	7	10	46
C	8	3	6	10	10	37
D	2	3	6	1	2	14
E	8	10	6	1	1	26
F	5	3	8	1	1	18
G	5	3	6	1	0	15
H	3	7	9	7	2	28
I	0	0	6	1	7	14
J	4	1	9	9	8	31

Tabela 1 – Avaliações dos gráficos

3.7 Resultados obtidos

A junção de dados da plataforma OpenAlex aos relatórios produzidos pelo scriptLattes permitiu estender os relatórios originalmente gerados pela ferramenta, acrescentando informações relevantes. Foram criadas diversas visualizações do conjunto de dados, favorecendo a observação dos principais *concepts* sob diferentes perspectivas.

3.8 Discussão

O pré-processamento é passo imprescindível e deve ser cuidadosamente planejado de acordo com a massa de dados a ser processada. A técnica adotada para o processamento dos dados depende fortemente da sequência de execução e dos parâmetros aplicados em cada algoritmo.

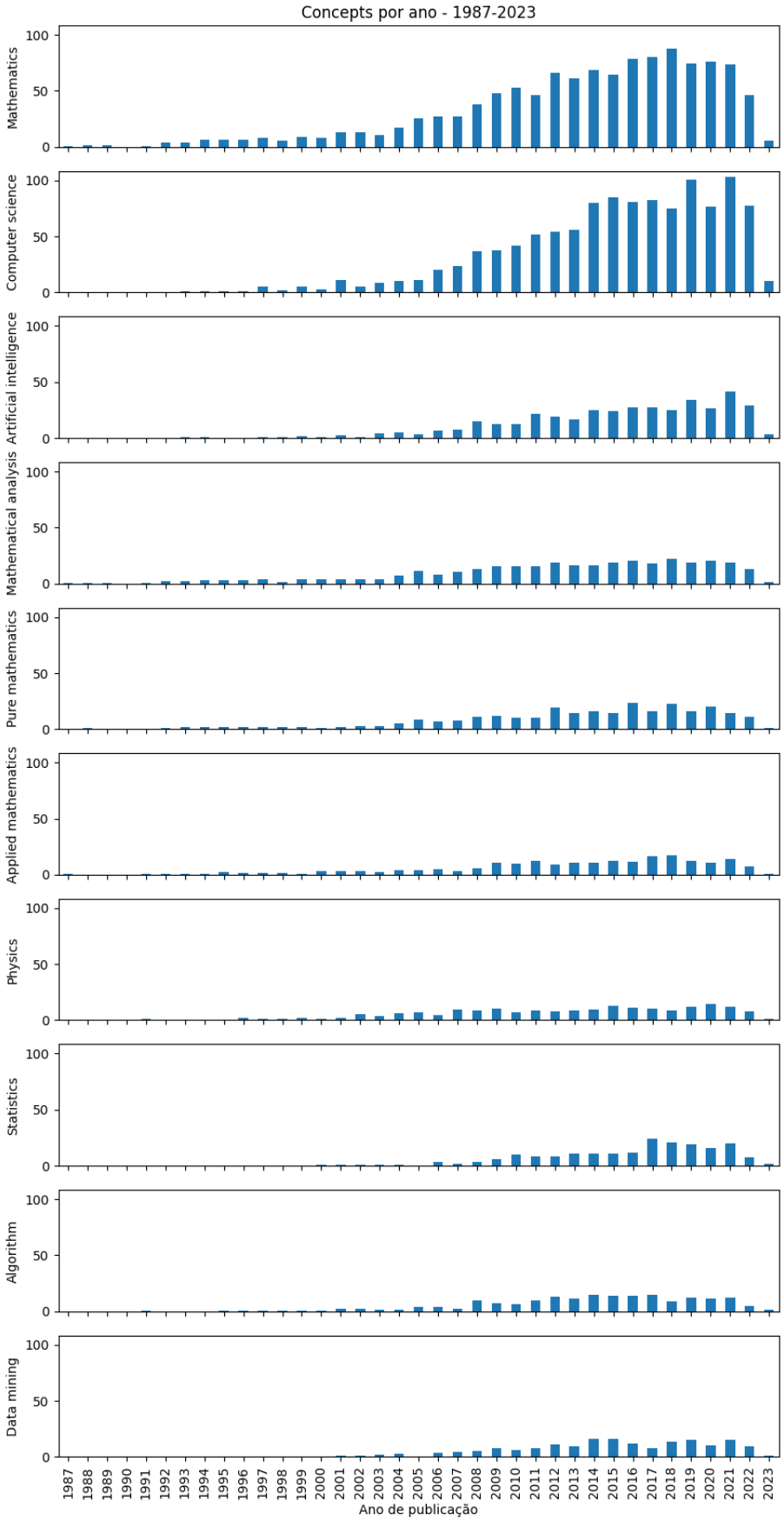

Figura 16 – Concepts por ano - Barras verticais independentes

4 CONCLUSÃO

O presente trabalho demonstrou uma forma de estender a ferramenta denominada scriptLattes, utilizando fontes de dados abertas (como OpenAlex) e aplicando técnicas de Aprendizado de Máquina.

A extração de dados da base OpenAlex trouxe volume massivo de informações, permitindo análise estatística e aplicação de algoritmos de AM. Em seguida a análise exploratória se mostrou fundamental para compreensão dos dados e escolha das técnicas de pré-processamento adequadas. Como apresentação do produto final, a aplicação de técnicas de visualização ofereceu suporte à melhor compreensão possível dos dados processados.

REFERÊNCIAS

BELLMAN, R. **Programação Dinâmica**. [*S.l.: s.n.*], 1957.

JURAFSKY D.; MARTIN, J. **Speech and Language Processing**. [*S.l.: s.n.*]: Prentice Hall, 2009.

MENA-CHALCO, J. P.; M., C.-J. R. scriptlattes: An open-source knowledge extraction system from the lattes platform. **Journal of the Brazilian Computer Society**, v. 15, n. 4, p. 31–39, 2009.

MENA-CHALCO, J. P.; M., C.-J. R. Prospecção de dados acadêmicos de currículos lattes através de scriptlattes. **Bibliometria e Cientometria: reflexões teóricas e interfaces**, v. 15, n. 4, p. 109–128, 2013.

OPENALEX. **Automated concept tagging for OpenAlex, an open index of scholarly articles**. 2020. Available at: https://docs.google.com/document/d/1OgXSLriHO3Ekz0OYoaoP_h0sPcuvV4EqX7VgLLblKe4/edit. Access at: 09 jun. 2023.

OPENALEX. **OpenAlex: The open catalog to the global research system**. 2023. Available at: https://openalex.org/. Access at: 09 jun. 2023.

PRIEM J.; PIWOWAR, H. O. R. Openalex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. **26th International Conference on Science, Technology and Innovation Indicators**, 2022. Available at: https://arxiv.org/abs/2205.01833.

RASTEIRO B. H.; MONTEIRO, R. P. T. Processamento de linguagem natural na educação superior: Comparando automaticamente currículos de cursos de computação. **V Workshop de Iniciação Científica em Tecnologia da Informação e da Linguagem Humana**, 2017. Available at: https://sites.icmc.usp.br/taspardo/TILic2017-RasteiroEtAl.pdf. Access at: 17 mar. 2023.

REZENDE, S. O. **Sistemas inteligentes: fundamentos e aplicações**. [*S.l.: s.n.*]: Manole, 2003.

SILVA J.A. E BIANCHI, M. Cientometria: a métrica da ciência. v. 20, n. 4, p. 5–10, 2001.

ZHIHONG S.; HAO, M. K. W. A web-scale system for scientific knowledge exploration. 2018. Available at: https://doi.org/10.48550/arXiv.1805.12216.