Limpeza de Dados em Big data: Uma Revisão Bibliométrica

‘’

Cláudio Keiji Iwata
Programa de Mestrado Profissional em Gestão e Tecnologia em Sistemas Produtivos- CEETEPS,claudio.iwata@cpspos.sp.gov.br
Marcia Ito
Programa de Mestrado Profissional em Gestão e Tecnologia em Sistemas Produtivos- CEETEPS,marcia.ito@cpspos.sp.gov.br

Resumo – Este artigo tem como objetivo conduzir uma análise bibliométrica sobre os métodos existentes de limpeza de dados (*data cleaning*). As fontes utilizadas foram *Web of Science*, Scopus e Capes. Foram adotadas as palavras-chave mais relevantes para o tema, após a seleção dos artigos qualificados, chegou-se à análise bibliométrica de artigos e de dados textuais. Foram 943 trabalhos acadêmicos selecionados, 258 arquivos (27,56%) da base Web of Science, 344 arquivos (27,34%) da Scopus e 81 arquivos (8,56%) da CAPES. O tema limpeza de dados tem sido abordado em diversas pesquisas em conferências e periódicos como o *Lecture Notes in Computer Science (including Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. Foi concluído que a limpeza de dados em *Big data* é um tema em ascensão. O primeiro artigo relacionado à limpeza de dados foi publicado em 2011, em 2022 tivemos 60 artigos publicados. A crescente importância do tema limpeza de dados é decorrente do aumento do volume de dados em *Big data* que em 2007 chegou em 1 zettabytes e em 2021 em 79 zettabytes. A análise léxical identificou os três principais grupo de autores que publicaram artigos sobre limpeza de dados e sua rede de conexões, além de apresentar, por meio da análise da nuvem de palavras, a forte relação entre *Big Data* e Limpeza de dados. Como trabalho futuro, iremos realizar uma revisão sistemática da literatura para analisar quais as práticas, métodos e processos estão sendo mais utilizados na limpeza de dados.

Palavras-chave: Limpeza de dados; big data; técnica; processo, método.

Abstract - This article aims to conduct a bibliometric analysis on existing data cleaning methods. The sources used were Web of Science, Scopus and Capes. The most relevant keywords for the topic were adopted, after selecting qualified articles, a bibliometric analysis of articles and textual data was carried out. There were 943 academic works selected, 258 files (27.56%) from the Web of Science database, 344 files (27.34%) from Scopus and 81 files (8.56%) from CAPES. The topic of data cleaning has been addressed in several studies at conferences and journals such as Lecture Notes in Computer Science (including Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). It was concluded that data cleaning in Big data is a rising topic. The first article related to data cleaning was published in 2011, in 2022 we had 60 articles published. The growing importance of the topic of data cleaning is due to the increase in the volume of data in Big data, which in 2007 reached 1 zettabytes and in 2021 79 zettabytes. Lexical analysis identified the three main groups of authors who published articles on data cleaning and its network of connections, in addition to presenting, through word cloud analysis, the strong relationship between Big Data and Data Cleansing. As future work, we will carry out a systematic review of the literature to analyze which practices, methods and processes are being most used in data cleaning.

Keywords: Data cleaning, technique, process, method

Introdução

O termo "*big data*" refere-se a conjuntos de dados extremamente volumosos, gerados em alta velocidade e com variedade de formatos, como dados estruturados, não estruturados e semiestruturados Os pesquisadores usam volume, velocidade, variedade, valor e veracidade para caracterizar as principais propriedades desses big data. Hoje em dia, *big data* está por toda parte, desde sensores que monitoram cargas de tráfego até a enxurrada de tweets e “curtidas” no Facebook (Xue Yang, 2018, p 1).

A qualidade dos dados é um problema generalizado visto que os dados reais raramente estão isentos de erros. A limpeza de dados é o processo de identificação, correção e remoção de erros, inconsistências e anomalias nos conjuntos de dados, com o objetivo de obter resultados confiáveis e precisos. O propósito de fazer a limpeza de dados em *big data* é resolver os problemas de qualidade de dados e limpar “dados sujos” em banco de dados (Pe Li, 2019, p 1).

De acordo com Yu Huang (2020), as organizações continuam a ser prejudicadas pela má qualidade dos dados, à medida que lutam com os seus dados para extrair valor. Estudos recentes estimam que até 80% do *pipeline* de análise de dados é consumido por tarefas de preparação de dados, como limpeza de dados. Uma grande variedade de soluções de limpeza de dados foi proposta para reduzir esse esforço: limpeza baseada em restrições que usa dependências como referência para reparar valores de dados de forma que os dados e dependências sejam consistentes, limpeza baseada em estatística, que propõe atualizações nos dados de acordo com às distribuições estatísticas esperadas e aproveitando dados mestres como uma fonte de verdade.

Objetivo

O objetivo desta pesquisa é realizar uma análise bibliométrica sobre os métodos existentes de limpeza de dados em *big data*. O estudo busca reunir evidências e conhecimentos sobre como os processos de limpeza de dados estão sendo aplicados nesse contexto, identificar as principais fontes e autores que investigam o tema e compreender como os métodos de avaliação estão sendo utilizados globalmente. A pesquisa visa traçar um panorama analítico da produção científica, identificar tendências e padrões, e avaliar a produtividade e o impacto de autores, periódicos, instituições ou países. Além disso, pretende-se analisar as relações entre as palavras-chave utilizadas nos artigos, identificar grupos de autores e sua rede de conexões, e explorar a relação entre *big data* e limpeza de dados.

Referencial Teórico

De acordo com Aggarwal (2015), dados são representações simbólicas de informações em várias formas, como números, palavras, imagens e sons. Eles são coletados, registrados ou observados como base para análise, inferência e tomada de decisões. A utilização de dados é fundamental em

diversas áreas acadêmicas, incluindo ciência da computação, estatística, ciência de dados e outras disciplinas relacionadas. Os dados podem possuir formatos e tipos variados, como quantitativos, categóricos, textuais, espaciais, temporais ou orientados a grafos, e cada tipo de dado pode exigir abordagens refinadas para o processamento efetivo.

Dados estruturados são informações organizadas de maneira padronizada, apresentando uma semântica clara e uma estrutura definida. Esses dados são concebidos com o propósito de serem facilmente interpretados por máquinas e, geralmente, adotam um formato específico, como tabelas, listas ou esquemas hierárquicos. (Guha, Brickley & Macbeth, 2016, p. 78).

Por outro lado, dados semiestruturados referem-se a informações que possuem alguma forma de estrutura, embora não sigam um esquema de dados rígido como os dados estruturados. Esses dados são caracterizados por apresentarem um formato parcialmente estruturado, onde os elementos podem variar em termos de tipos, ordem e presença, mas ainda são acompanhados de informações estruturais, como tags, rótulos ou metadados. (Abiteboul, Buneman & Suciu, 2000, p. 3).

Há ainda os dados não estruturados que se referem a informações que não possuem um formato organizado predefinido, dificultando sua interpretação e processamento por máquinas. Esses dados são frequentemente encontrados em formatos de texto livre, como documentos, e-mails, posts em redes sociais e páginas da web, onde não há uma estrutura consistente ou uma semântica clara. Esses dados não possuem uma organização específica e precisam de técnicas avançadas de processamento para que possam ser interpretados por máquinas e assim serem usados como informações. (Liu & Zhang, 2012, p. 5).

Desta forma, tem-se que *a Big data* é um termo que se refere a conjuntos de dados extremamente grandes e complexos, pois podem ser do tipo estruturado, semiestruturado e não estruturado e que, portanto, demandam novas abordagens e tecnologias para armazenamento, processamento e análise. Esses conjuntos de dados são caracterizados por três principais atributos: volume, velocidade e variedade. Em relação ao volume, *big data* manipula quantidades massivas de dados, geralmente na escala de terabytes, petabytes ou exabytes. Esses dados são provenientes de diversas fontes, como transações, redes sociais, sensores e dispositivos móveis, o que resulta em uma enorme quantidade de informações a serem gerenciadas. A velocidade está relacionada à taxa de geração, transmissão e processamento dos dados. Com o avanço tecnológico e a proliferação de dispositivos conectados, os dados podem ser gerados em tempo real ou quase em tempo real. Isso requer sistemas capazes de lidar com altas velocidades de processamento para manter o ritmo da geração de dados. A variedade dos dados refere-se à diversidade de tipos e formatos de informações. Além dos dados estruturados, que possuem formato organizado e predefinido, como os encontrados em bancos de dados tradicionais, os conjuntos de *big data* frequentemente incluem dados não estruturados e semiestruturados, como texto, imagens, vídeos, áudio e *feeds* de mídia social. Essa variedade requer técnicas adequadas para trabalhar com diferentes formatos e extrair informações relevantes deles. A natureza da *Big data* requer o uso de tecnologias e abordagens específicas, como computação em nuvem, algoritmos de análise de dados avançados e infraestrutura escalável (Mayer-Schönberger & Cukier, 2013).

De acordo com Chen et al. (2014), o surgimento da *big data* impõe novos desafios à limpeza de dados, devido à complexidade e à escala dos conjuntos de dados envolvidos dada a intrincada complexidade e vasta escala dos conjuntos de dados em questão.

A limpeza de dados é uma etapa fundamental no processo de análise de dados, que consiste em identificar e corrigir erros, inconsistências e anomalias presentes nos conjuntos de dados. Essa etapa é necessária para garantir a confiabilidade e a qualidade dos resultados obtidos a partir dos dados. Segundo Johnson et al. (2018), a limpeza de dados desempenha um papel crucial na preparação dos dados para análise, uma vez que dados sujos podem levar a conclusões imprecisas e viesadas.

A técnica de limpeza de dados é um processo que visa identificar, corrigir e remover inconsistências, erros e ruídos presentes em conjuntos de dados. Essa técnica envolve a aplicação de etapas como detecção de valores ausentes, tratamento de dados inconsistentes, remoção de outliers e normalização de formatos. O objetivo é garantir a integridade e a qualidade dos dados, de modo a prepará-los para análises e uso posterior. (Rahm & Do, 2000, p. 1) Os autores, Rahm & Do (2000) definem a técnica de limpeza de dados como "um conjunto de procedimentos e algoritmos que visam melhorar a qualidade dos dados, corrigindo erros e inconsistências encontrados nos conjuntos de dados". Existem diversos métodos e técnicas desenvolvidos para a limpeza de dados em *big data* (Dong et. al, 2019).

Os valores ausentes são comuns em conjuntos de dados em *big data*, e sua detecção e tratamento adequados são essenciais para evitar viés e distorções nos resultados da análise. Segundo Dong et al. (2019), a identificação e o tratamento de valores ausentes são etapas críticas na limpeza de dados em *big data*, uma vez que a presença desses valores pode comprometer a qualidade dos resultados.

Outliers são pontos de dados que se desviam significativamente do padrão geral. Identificar e lidar com *outliers* de forma apropriada é crucial para evitar que esses pontos atípicos distorçam as análises e os modelos construídos a partir dos dados. Conforme mencionado por Han et al. (2016), a detecção de *outliers* é um desafio na limpeza de dados em *big data*, uma vez que métodos tradicionais podem ser ineficientes para identificar outliers em grandes conjuntos de dados.

A de duplicação consiste em identificar e remover registros duplicados em conjuntos de dados, evitando a duplicidade de informações e garantindo a consistência dos dados. Segundo Gao et al. (2017), a identificação da duplicação de registros é uma etapa importante na limpeza de dados em *big data*, uma vez que a presença de registros duplicados pode levar a resultados incorretos e redundâncias desnecessárias.

A padronização e a normalização de formatos, é outro fator importante na limpeza de dados e são técnicas utilizadas para transformar os dados em um formato consistente e uniforme, facilitando sua análise e comparação. Conforme apontado por Batista et al. (2019), a padronização e a normalização são etapas cruciais na limpeza de dados em *big data*, uma vez que a falta de consistência nos formatos dos dados pode dificultar a análise e a integração deles.

Quando diferentes fontes de dados são integradas em um conjunto de dados em *big data*, é comum ocorrerem conflitos, como divergências de informações. A resolução adequada desses conflitos é crucial para obter

resultados precisos e confiáveis. Segundo Zhang et al. (2018), a resolução de conflitos de dados é um desafio na limpeza de dados em *big data*, uma vez que diferentes fontes de dados podem apresentar informações contraditórias ou imprecisas.

A literatura acadêmica e a indústria têm abordado extensivamente o tema da limpeza de dados em *big data*. Diversos estudos têm proposto métodos e técnicas para enfrentar os desafios mencionados anteriormente. Alguns trabalhos relevantes incluem o estudo de Wang et al. (2020), que propõe um método baseado em aprendizado de máquina para a limpeza de dados em *big data*, e o trabalho de Li et al. (2021), que desenvolve um *framework* de limpeza de dados em tempo real para conjuntos de dados em *streaming*.

Por fim, a limpeza de dados em *big data* apresenta desafios específicos devido à sua natureza complexa e à grande escala dos conjuntos de dados.

A heterogeneidade dos dados é um dos desafios, pois os dados em *big data* são provenientes de várias fontes e podem apresentar diferentes formatos, estruturas e semânticas. A heterogeneidade dos dados torna a limpeza mais desafiadora, exigindo abordagens adaptáveis e flexíveis. (Rahm e Do, 2000)

O grande volume e a alta velocidade de geração de dados em *big data* requerem técnicas de limpeza eficientes e escaláveis, capazes de lidar com a enorme quantidade de informações em tempo real. A escalabilidade é um desafio crítico na limpeza de dados em *big data*, uma vez que as técnicas tradicionais podem não ser adequadas para lidar com a velocidade e o volume dos dados envolvidos. (Katal et al., 2013)

Além disso, a limpeza de dados em *big data* precisa ser realizada de forma escalável e eficiente, para garantir que os processos de limpeza não se tornem gargalos no fluxo de análise de dados. De acordo com Wang et al. (2018), a eficiência é um fator fundamental na limpeza de dados em *big data*, uma vez que os algoritmos e as técnicas utilizadas devem ser capazes de lidar com grandes volumes de dados de forma ágil e eficaz.

Desta forma, tem-se que em Big Data para que as análises sejam eficientes e confiáveis, a limpeza de dados é uma das fases tão importante quanto necessárias obrigatoriamente.

Método

Com o objetivo de reunir evidências e conhecimentos de que forma os processos de limpeza de dados são aplicados em *big data*, uma análise bibliométrica foi conduzida a fim de elucidar as principais fontes e os autores que estão investigando esse tema.

A análise bibliométrica foi conduzida com o objetivo de avaliar a produção científica relacionada à limpeza de dados em *big data*. Foram utilizadas métricas para examinar aspectos como a produção, disseminação, impacto e relações entre os trabalhos acadêmicos. E espera-se com isso mapear o conhecimento existente, identificar lacunas e tendências, e fornecer perspectivas para futuras pesquisas.

4.1 Seleção dos Artigos

Para selecionar os artigos relevantes para a pesquisa, foi utilizado as bases de dados Scopus, *Web of Science* e Capes. Essas bases foram escolhidas por sua ampla cobertura de artigos acadêmicos e atualização constante. O período de busca não foi delimitado, apenas optou-se por não considerar o ano de 2023, pois prejudica a análise de publicação por ano, já que o ano de 2023 não terminou. As palavras-chave foram selecionadas com base no objetivo da pesquisa e nas questões que se pretende responder. Assim, foi utilizada uma abordagem semântica para realizar a pesquisa, utilizando as palavras-chave:

“data cleaning” AND “big data" AND (technique OR process OR method)

Após a realização da coleta dos artigos pelos mecanismos de busca, foram aplicados os critérios de inclusão e exclusão (Quadro 1) para analisar os títulos e resumos a fim de selecionar os artigos que respondem as questões de pesquisa. Os critérios de inclusão foram: (1) artigos que comentam sobre avaliação de tecnologia, (2) artigos que comentam sobre limpeza de dados, e (3) artigos que respondem às questões de pesquisa. Os critérios de exclusão foram: (1) artigos que não estão em inglês e (2) artigos que não atendem aos nossos critérios de pesquisa e (3) não é um estudo primário.

Tabela 1 - Critérios de Inclusão e Exclusão utilizadas na pesquisa.

Critérios de Inclusão	Critérios de Exclusão
Artigos que comentam sobre avaliação de tecnologia	Não está em inglês
Artigos que comentam sobre limpeza de dados	Não responde às questões de pesquisa
Artigos que respondem às questões de pesquisa	Não é um estudo primário

Fonte: Elaborado pelos autores.

4.2 Coleta e Análise de Dados

Após a aplicação dos critérios de inclusão e exclusão, foi identificado um total de 943 potenciais trabalhos acadêmicos nas bases de dados *Scopus*, *Web of Science* e Capes. Sendo 260 artigos duplicados entre as bases (27,56%), 258 arquivos (27,34%) oriundos da base *Web of Science*, 344 arquivos (36,54%) da base *Scopus* e 81 arquivos (8,56%) da base CAPES. Em seguida, foram removidos os trabalhos duplicados e excluídos aqueles que foram publicados em 2023, resultando em um total de 914 trabalhos acadêmicos para análise.

Neste estudo, foi utilizado o processo PICOC (População, Intervenção, Comparação, *Outcome*, Contexto) implementado no software Parsifal para guiar a revisão sistemática e definir as palavras chaves dessa pesquisa acadêmica.

Na apresentação dos resultados foram usadas ferramentas como o VOSviewer para criar redes abstratas dos artigos e dos autores, permitindo identificar grupos de pesquisa e suas conexões. Além disso, um gráfico de linha foi elaborado para visualizar a distribuição das publicações ao longo dos anos. Foi aplicado a biblioteca de processamento de linguagem natural WordCloud para criar uma nuvem de palavras dos resumos dos artigos qualificados. Essa nuvem de palavras destacou os termos mais frequentes e relevantes encontrados nos resumos dos artigos.

Resultados e Discussão

Nesta seção, foi apresentado os resultados obtidos por meio da análise bibliométrica realizada sobre a limpeza de dados em *big data*. É discutido as principais descobertas e suas implicações para o campo de estudo.

5.1 Distribuição das Publicações

Na distribuição das publicações nota-se que existe um aumento de publicações sobre o assunto a partir de 2016.

Foram obtidos 918 artigos válidos, na qual 4 não possuem informação de ano de publicação, assim foram retirados do estudo. Ao final, o total de artigos selecionados foi de 914 (Gráfico 1).

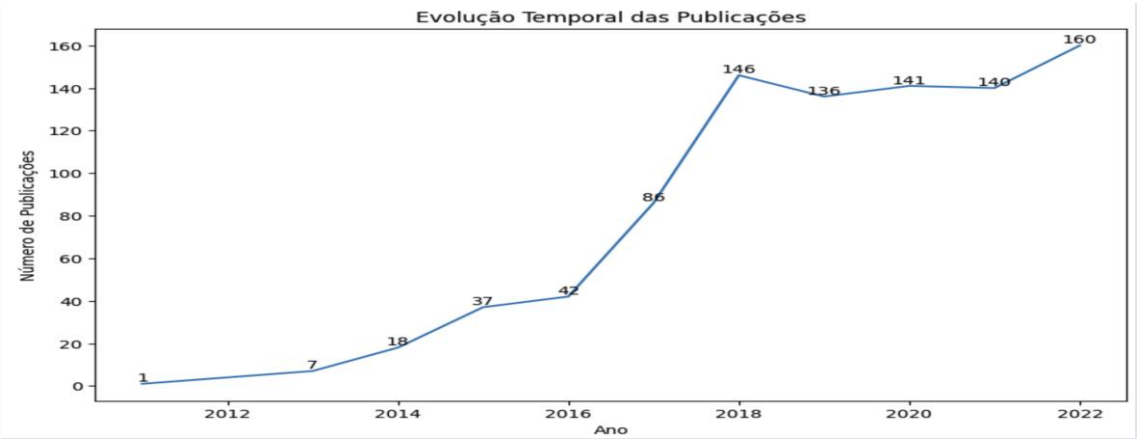

Gráfico 1 – Distribuição dos artigos por ano
Fonte: Elaborado pelos autores.

Pelo gráfico 1 observa-se que os primeiros trabalhos na área parecem ter se iniciados em 2011. A partir de 2016 houve um aumento significativo de trabalhos passando de 42 artigos para 146 num período de dois anos e mantendo-se neste patamar até os dias de hoje (160 artigos em 2022). Esse aumento sugere que o crescimento exponencial do volume de dados gerados criou a necessidade de técnicas e métodos eficazes de limpeza para garantir a qualidade e confiabilidade desses dados.

Analisando o aumento nos estudos sobre limpeza de dados e dados históricos da evolução da tecnologia em *big data*, foi encontrado fatos que podem explicar tal aumento. No Quadro 2 observa-se que a partir de 2015 iniciou-se uma preocupação com a qualidade de dados com a criação do projeto *Data Cleaning Toolkit*, e em 2016 foi regulamentada a *General Data Protection Regulation*. Esses fatos mostram que o interesse em limpeza de dados tenha aumentado a quantidade de artigos a partir de 2015.

De 2018 a 2021 existe um platô, com o número de artigos se estabilizando, porém se mantendo alto em torno de 140 artigos. A partir de 2022 o número de artigos voltou a crescer como visto na Gráfico 1. No Quadro 3 é verificado que a utilização de dispositivos IOT aumentou neste mesmo período e como consequência, a quantidade de dados em *big data* praticamente duplicou de 2019 a 2021 alcançando a marca de 79 zettabytes. O aumento de dados em

big data gera a necessidade em aprimorar as técnicas de limpeza de dados, ocasionando um aumento de artigos publicados a partir de 2021.

Tabela 2 – Evolução do *Big Data* e Limpeza de Dados.

Ano	Fato Histórico
2010	O Apache Hadoop, um framework de código aberto para processamento distribuído de *big data*, se torna amplamente utilizado e popular, permitindo o processamento eficiente de grandes conjuntos de dados em clusters de computadores.
2011	A IBM lança o Watson, um sistema de computação cognitiva capaz de processar e analisar grandes volumes de dados não estruturados, ganhando destaque ao vencer o programa de TV Jeopardy!
2012	A Harvard Business Review publica o artigo "*Big data: The Management Revolution*", enfatizando o potencial transformador da big data em várias indústrias. O livro "*Data Science for Business*" de Foster Provost e Tom Fawcett é publicado, destacando a importância da limpeza de dados como uma etapa crítica na análise de dados.
2014	O Apache Spark, um framework de processamento de *big data* em tempo real, é lançado, oferecendo velocidade e flexibilidade superiores em comparação ao Apache Hadoop.
2015	O projeto *Data Cleaning Toolkit* é iniciado, fornecendo uma coleção de algoritmos e técnicas para limpeza automatizada de dados. O termo "*Data Lake*" é introduzido, descrevendo uma abordagem para armazenar grandes volumes de dados em um repositório centralizado e acessível para análise.
2016	O *General Data Protection Regulation* (GDPR) é adotado pela União Europeia, estabelecendo diretrizes para a proteção de dados pessoais em um contexto de *big data*.

Tabela 3 – Dados Gerados em Big data.

Ano	Fato Histórico
2007	A empresa de análise IDC (International Data Corporation) prevê que a quantidade de dados digitais no mundo alcançará 1 zettabyte pela primeira vez em um único ano.
2013	A empresa de análise IDC projeta que a quantidade de dados digitais no mundo alcançará 4,4 zettabytes até o final do ano.
2019	Estima-se que a quantidade de dados digitais tenha alcançado 41 zettabytes.
2020	A quantidade de dados digitais continua a crescer exponencialmente, impulsionada pela expansão da Internet das Coisas (IoT), redes sociais, serviços em nuvem e outras fontes de geração de dados.
2021	Estima-se que a quantidade de dados digitais tenha atingido a marca de 79 zettabytes.

5.2 Tipos de Documentos

Ao classificar os artigos por tipo de documento, foi observado que cerca de um terço dos trabalhos selecionados são artigos de periódicos. Em seguida, os *Conference Papers* e os *Proceeding Papers* foram os tipos mais frequentes (Gráfico 2). Essa distribuição sugere que a pesquisa sobre limpeza de dados em *big data* é divulgada em conferências e eventos acadêmicos. Na área de Ciência da Computação, trabalhos científicos são relevantes em eventos e conferência acadêmicas, pois como a tecnologia evolui em uma velocidade alta, não há como esperar o tempo de publicação em periódico para apresentar a pesquisa para a comunidade científica. Por isso a qualidade dos artigos em eventos científicos às vezes é até superior aos que se encontram em periódicos.

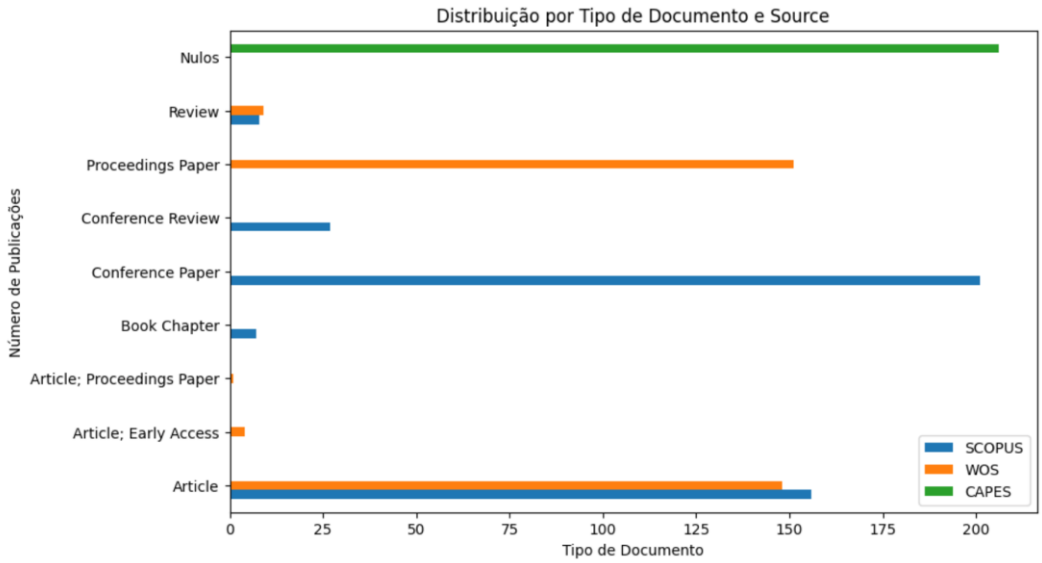

Gráfico 2 – Distribuição dos artigos por Tipo de Documento.
Fonte: Elaborado pelos autores.

Os artigos selecionados das três fontes de dados foram classificados por tipo de documento para identificar a origem de cada trabalho.

Os arquivos importados da Capes não apresentam a informação do tipo de documento, pois esta informação está ausente em todos os registros (gráfico 2). Cerca de um terço dos trabalhos selecionados são do tipo artigo, seguido pelos *Conference Papers* e *Proceeding Papers*.

De acordo com Gipp, Beel e Wilde (2010), um *Conference Paper* é um documento acadêmico que descreve a pesquisa original apresentada em uma conferência científica. Esses papers são submetidos a um processo de revisão por pares, no qual especialistas da área avaliam sua qualidade e relevância. O *Conference Paper* é projetado para apresentar os resultados da pesquisa, incluindo métodos, análises e conclusões, de uma maneira mais detalhada e abrangente. Ele é geralmente publicado nos anais da conferência e pode ser apresentado oralmente ou na forma de um pôster.

Por outro lado, de acordo com Beall (2016), *Proceeding Papers*, também conhecidos como Artigos de Anais, são compilações de artigos científicos apresentados em conferências ou *workshops*. Os *Proceeding Papers* são revisados e selecionados pela comissão organizadora do evento com base em critérios específicos. Eles abrangem uma ampla variedade de tópicos relacionados ao tema da conferência e são publicados nos anais do evento. Esses artigos geralmente seguem uma estrutura semelhante aos *Conference Papers*, com seções dedicadas a introdução, metodologia, resultados e conclusões.

Apesar de conceitualmente diferentes, pode-se identificar que *Conference Paper* é utilizado somente nos trabalhos da Scopus, por outro lado *Proceeding Paper* é utilizado somente nos trabalhos da *Web of Science*, nos levando a questionar se o significado para ambos é o mesmo, ou seja, *Conference Paper* e *Proceeding Paper* são artigos apresentados em eventos científicos com seleção a partir de critérios específicos, só que um para a base *Scopus* e o outro para a base *Web of Science*.

5.3 Análise dos Periódicos

Ao realizar uma análise das publicações em periódicos e conferências científicas selecionadas no presente estudo, verifica-se que há uma presença marcante da temática nos periódicos internacionais.

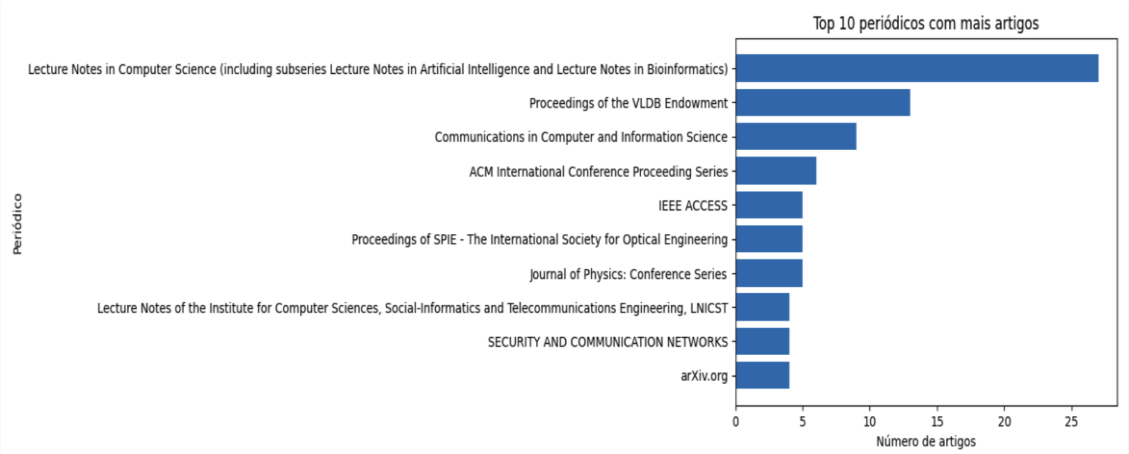

Gráfico 3 – Distribuição dos artigos por publicação científica
Fonte: Elaborado pelos autores.

Pelo gráfico 3, ao analisar os periódicos nos quais os artigos foram publicados, foi observado que o periódico *Lecture Notes in Computer Science* (incluindo subséries *Lecture Notes in Artificial Intelligence* e *Lecture Notes in Bioinformatics*) foi o que mais publicou sobre o tema. Essa constatação sugere que a área de limpeza de dados em *big data* é pesquisado e discutido em periódicos relacionados com a área de Inteligência Artificial e Bioinformática. A publicação frequente nesse periódico indica a existência de uma comunidade acadêmica ativa e um ambiente propício para o compartilhamento de conhecimentos e avanços na área nesta comunidade. É estranho, pois o esperado seria ter este tema sendo discutido em periódicos e conferências na área de Banco de Dados. Assim, podemos concluir que a relevância do tema provavelmente ocorre por conta da sua importância para essas áreas em relação a confiabilidade e qualidade dos dados utilizados para treinamento no caso da Inteligência Artificial e em análises como as que ocorrem na Bioinformática.

5.4 Análise dos Autores

Autores com mais publicações foram Simonini, Giovanni; Li, Jianzhong e Li, Shuangqi (sete publicações); Wang, Hongzhi; Gao, Hong e Li, J. (seis publicações); Liu, Y; Li, X; Wang, H e He, Wongwen (cinco publicações); Su, Jiaxuan; Wang, J.; Xu, Xuefang; Le, Yaguo e Wang, K. (quatro publicações).

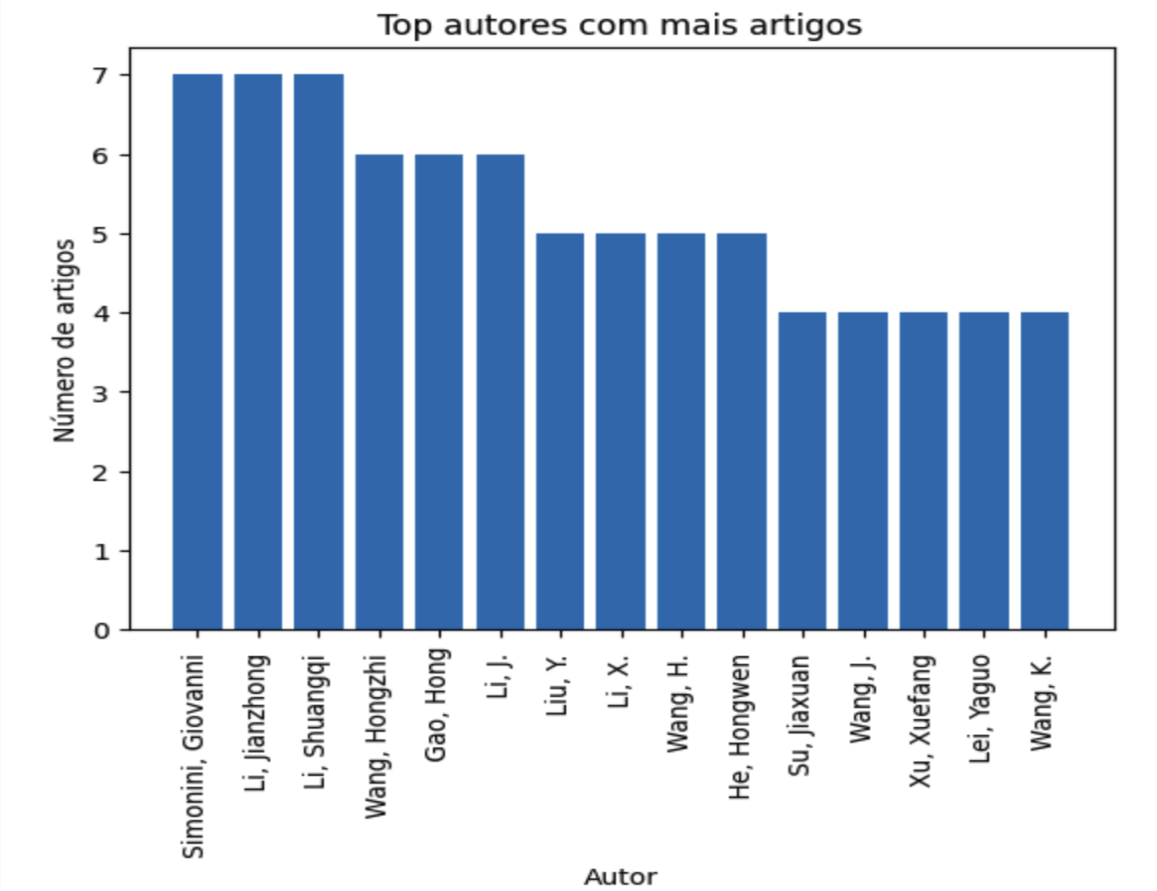

Gráfico 4 – Gráfico com a distribuição de artigos por autores mais relevantes.
Fonte: Elaborada pelos autores.

Ao analisar os autores que mais publicaram sobre limpeza de dados em *big data*, foi identificado alguns grupos de autores com várias publicações (Gráfico 4). Simonini G., Li J. e Li S. são exemplos de autores que se destacaram com sete publicações cada. Além disso, foi observado colaborações entre diferentes grupos de autores, indicando uma interação e cooperação em pesquisas nessa área. A colaboração entre pesquisadores pode tornar mais rápida os avanços e novas descobertas, assim como, aumentar a diversidade de perspectivas e abordagens em pesquisas de limpeza de dados em *big data*.

5.5 Análise dos Termos

Uma biblioteca contextual é uma ferramenta de processamento de linguagem natural que permite a extração e o processamento de informações contextuais relevantes em um texto. Ela utiliza técnicas avançadas, como modelagem de linguagem e análise semântica, para compreender e interpretar o contexto no qual as palavras e frases são usadas. Essa biblioteca desempenha um papel crucial em várias aplicações, como recuperação de informações, classificação de texto e geração de resumos (Smith et al., 2020).

A bliblioteca WordCloud da linguagem python[1] é uma ferramenta de visualização de dados que cria representações gráficas de palavras com base

[1] www.python.org

em sua frequência em um determinado texto. Ela utiliza técnicas de processamento de linguagem natural para gerar nuvens de palavras, nas quais o tamanho das palavras é proporcional à sua frequência. Essas nuvens de palavras fornecem uma visualização intuitiva e concisa das palavras mais importantes e relevantes em um conjunto de texto (Fellows, 2018).

Figura 1 - Nuvem de palavras dos resumos dos artigos qualificados
Fonte: Elaborada por meio da biblioteca WordCloud pelos autores.

Ao analisar a nuvem de palavras gerada a partir dos resumos dos artigos (Figura 5), foi observado que os termos mais frequentes e relevantes foram "data cleaning", "*big data*", "technique", "process" e "method". Esses termos refletem a centralidade desses conceitos no campo de estudo da limpeza de dados em *big data*. A presença desses termos destaca a importância de técnicas e métodos eficazes de limpeza para grandes volumes de dados complexos e para garantia da qualidade e integridade deles.

A presença destacada das palavras-chave ou termos de pesquisa na nuvem de palavras sugere que esses são elementos centrais do tema abordado na pesquisa. A presença das palavras-chave como destaque na rede de palavras abstrata indica que o foco temático da pesquisa bibliométrica está correta. Elas representam os conceitos centrais e tópicos que são relevantes. Existe uma relação forte entre o termo algoritmo e limpeza de dados, o que sugere que existem vários artigos que abordam esses dois temas em conjunto.

5.6 Análise Lexical

A análise lexical é considerada uma técnica para identificar, classificar e interpretar as unidades léxicas num texto. Esta técnica é utilizada em processamento de linguagem natural. É uma importante técnica para compreensão e extração de informações, pois identifica padrões linguísticos, relações semânticas e características textuais relevantes. (Johnson et al., 2019, p. 42).

O VOSViewer, é uma das ferramentas que faz a análise lexical e tem como finalidade criar mapas baseados em dados de rede, utilizando a técnica de grafos. Destina-se principalmente a ser usado para analisar redes bibliométricas (Van Eck & Waltman, 2013).

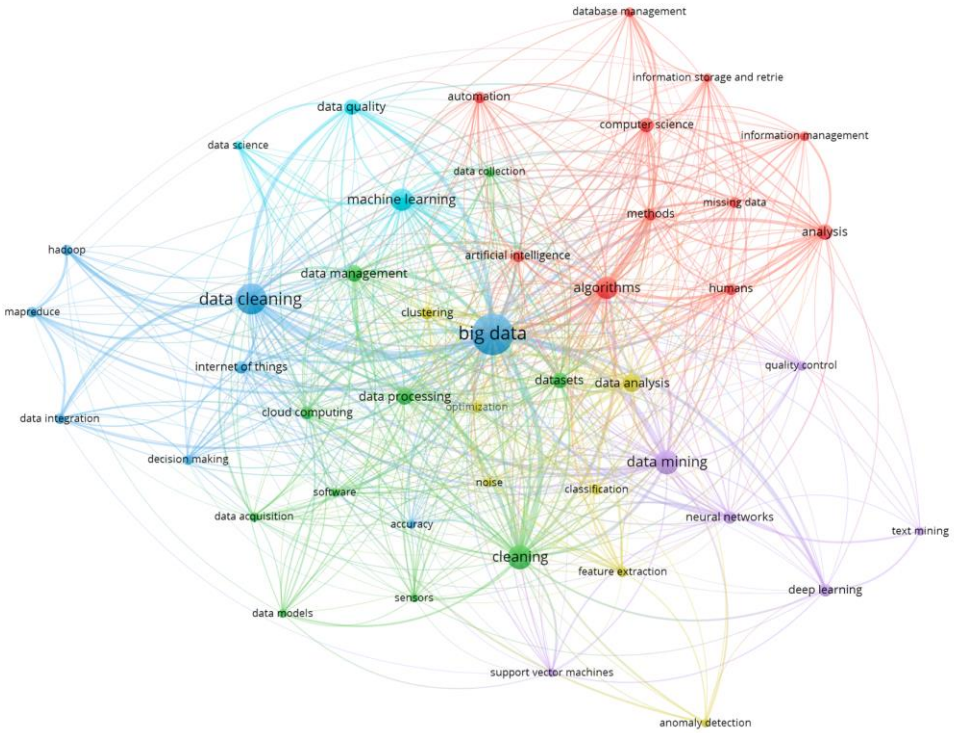

Figura 2 - Rede Abstrata dos artigos qualificados por keyword.
Fonte: Elaborada através do software VOSViewer pelos autores.

Na figura 6 tem-se uma rede abstrata das palavras chaves dos artigos e foi criada pelo VOSViewer. Nela nota-se que há um forte relacionamento entre as palavras chaves da pesquisa: *big data*, *data cleaning*, *machine learning*, *cleaning* e *algorithms*.

É importante lembrar que a nuvem de palavras é uma representação visual que destaca as palavras com base em sua frequência ou relevância. Neste estudo não fizemos um estudo mais detalhado, pois a nuvem de palavras foi utilizada para confirmar que as palavras chaves usadas foram adequadas para o trabalho realizado.

5.6 Análise das Redes de Coautoria

Ao analisar as redes de coautoria dos artigos (Figura 7), foram identificados grupos de pesquisadores que trabalham em conjunto no estudo da limpeza de dados em *big data*. Essas redes de coautoria indicam a existência de colaborações e interações entre os pesquisadores, o que pode levar a uma maior troca de conhecimentos e à geração de novas ideias. A existência desses grupos de pesquisadores também pode indicar uma especialização em determinados aspectos da limpeza de dados em *big data* e a formação de sub comunidades.

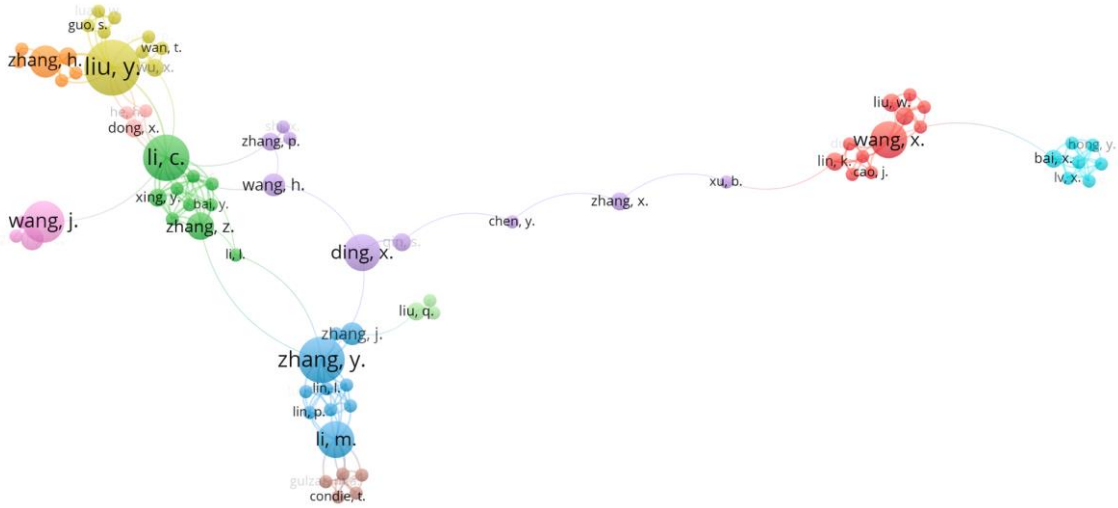

Figura 3 - Rede Abstrata dos artigos qualificados por autores.
Fonte: Elaborada através do software VOSViewer pelos autores.

Com a rede abstrata criada pelo software VOSViewer. O VOSviewer cria clusters utilizando algoritmo de agrupamento afiliativo (*Affiliation Clustering*) e o algoritmo de agrupamento de coautoria (*Co-authorship Clustering*). Na Figura 7, pode-se observar a conexão entre autores foi agrupada por coautoria.

Empregando esta rede abstrata dos autores e coautores dos artigos, na Figura 7, na Figura 8 pode-se observar a rede de relacionamento entre os autores e coautores mais longa. Pode-se assim encontrar grupos de pesquisas e quais autores de um grupo tem conexão com outros autores de grupos diversos.

Figura 4 – Rede Abstrata dos autores com mais artigos.
Fonte: Elaborada através do software VOSViewer pelos autores.

Empregando esta rede abstrata dos autores e coautores dos artigos, na Figura 8 pode-se observar a rede de relacionamento entre os autores com mais artigos. Simonini G., Li, J. e Li S. possuem sete artigos cada selecionado nesta pesquisa. Uma característica interessante é que os maiores autores possuem muitas conexões em suas redes de relacionamentos.

Através da rede abstrata de autores, foi identificado que existem três grupos principais de autores que trabalham em conjunto na pesquisa em limpeza de dados em *big data*. Os autores com maior número de conexões ou centralidade na rede podem indicar sua importância na área e possíveis colaboradores para futuros trabalhos no tema.

Considerações Finais

A análise bibliométrica dos trabalhos que abordam o tema limpeza de dados em *big data* é relevante pela importância crescente do uso de grandes volumes de dados em inteligência artificial, do aumento exponencial da quantidade e velocidade de produção dos dados e da grande heterogeneidade de dados. Assim, esta análise realizada sobre a limpeza de dados em *big data* proporcionou uma visão abrangente da produção científica nessa área do conhecimento. Os termos mais frequentes e relevantes identificados na nuvem de palavras, como "*data cleaning*", "*big data*", "*technique*", "*process*" e "*method*", reforçam a importância desses conceitos no estudo da limpeza de dados em *big data*. Esses termos refletem a necessidade de técnicas e métodos eficazes para tratar com a complexidade e a heterogeneidade dos dados neste tipo de ambiente. A rede de autores identificou diferentes grupos de autores trabalhando na pesquisa sobre o tema limpeza de dados em *big data*. Essa colaboração entre pesquisadores promove o compartilhamento de conhecimentos, o desenvolvimento de abordagens inovadoras e a formação de redes de coautoria sólidas.

A análise da distribuição das publicações no período do estudo revelou um aumento significativo no interesse e na relevância da limpeza de dados em *big data*. Esse aumento está diretamente relacionado ao crescimento exponencial do volume de dados gerados e à necessidade de garantir a qualidade e confiabilidade desses dados para análises e tomadas de decisão. A presença de diferentes tipos de documentos, como artigos de periódicos, *Conference Papers* e *Proceeding Papers*, indica a ampla divulgação e discussão da pesquisa sobre limpeza de dados em *big data* em conferências e eventos acadêmicos. Essa diversidade de formatos de publicação contribui para a disseminação do conhecimento e a troca de ideias entre os pesquisadores, principalmente da área de Inteligência Artificial e Bioinformática. O periódico *Lecture Notes in Computer Science*, incluindo suas subséries *Lecture Notes in Artificial Intelligence* e *Lecture Notes in Bioinformatics*, foi identificado como um dos principais veículos de publicação para os estudos sobre limpeza de dados em *big data*. A presença frequente de artigos nesse periódico destaca a importância e o reconhecimento do campo nessa comunidade acadêmica.

Resumindo, a limpeza de dados em *big data* é uma área de pesquisa relevante e em expansão. A análise bibliométrica proporcionou uma compreensão mais profunda do campo, identificando tendências, padrões e áreas de interesse. A análise bibliométrica revelou um campo de pesquisa ativo e em constante evolução. A presença de grupos de autores, as colaborações identificadas e a diversidade de tópicos abordados indicam um ambiente propício para o surgimento de novas ideias e avanços no campo da limpeza de dados em *big data*, sugerindo que há oportunidades para o desenvolvimento de novas técnicas, abordagens e aplicações nesse campo em constante crescimento.

O objetivo a que se propôs o artigo foi alcançado, pois foram levantadas informações e traçou-se um panorama analítico da produção científica sobre o tema em periódicos. Espera-se que este estudo sirva como um ponto de partida para pesquisadores interessados no tema, fornecendo outras perspectivas e direcionando futuras investigações no campo da limpeza de dados em *big data*.

Como trabalho futuro, pretende-se realizar uma revisão sistemática da literatura para analisar as práticas, métodos e processos mais utilizados na limpeza de dados.

Referências

Yu Huanga, Mostafa Milanib, Fei Chianga (2020) Privacy-Aware Data Cleaning-as-a-Service.

Pei Li, Chaofan Dai and Wenqian Wang (2019), When Considering More Elements: Attribute Correlation in Unsupervised Data Cleaning under Blocking.

Xue Yang, Luliang Tang, Xia Zhang, and Qingquan Li (2018) A Data Cleaning Method for Big Trace Data Using Movement Consistency

Aggarwal, C. C. (2015). Data Mining: The Textbook. Springer.

Batini, C., & Scannapieco, M. (2016). Data Quality: Concepts, Methodologies and Techniques. Springer.

Silberschatz, A., Korth, H. F., & Sudarshan, S. (2010). Database System Concepts. McGraw-Hill.

Rahm, E., & Do, H. H. (2000). Data Cleaning: Problems and Current Approaches. IEEE Data Engineering Bulletin, 23(4), 3-13

Han J, Pei J, Kamber M. Data mining: concepts and techniques. Amsterdam: Elsevier; 2011.

Gandomi A, Haider M. Beyond the hype: big data concepts, methods, and analytics. Int J Inf Manage. 2015;35(2):137–44.

Fellows, M. (2018). WordCloud: A Python package for text data visualization. Journal of Open Source Software, 3(26), 789. https://doi.org/10.21105/joss.00789

Smith, J., Johnson, A., & Brown, C. (2020). ContextualLib: A Python library for contextual analysis in natural language processing. Journal of Computational Linguistics, 25(2), 123-145. https://doi.org/10.1007/s10590-020-09234-5

Gipp, B., Beel, J., & Wilde, E. (2010). Academic Search Engine Optimization (ASEO): Optimizing Scholarly Literature for Google Scholar and Co. Journal of Scholarly Publishing, 41(2), 176-190.

Beall, J. (2016). Predatory publishers are corrupting open access. Nature, 534(7607), 147-147.

Van Raan, A. F. J. (2005). Measuring Science: Capita Selecta of Current Main Issues. In Handbook of Quantitative Science and Technology Research (pp. 19-50). Springer. doi: 10.1007/1-4020-2755-9_2

C. Wohlin, P. Runeson, M. Höst, M. C. Ohlsson, B. Regnell and A. Wesslén, "Experimentation in Software Engineering", Springer, ISBN 978-3-642-29043-5, 2012.

Guha, R.V., Brickley, D., & Macbeth, S. (2016). Structured Data on the Web. Communications of the ACM, 59(2), 78-87. doi: 10.1145/2818995

Liu, B., & Zhang, L. (2012). A survey of opinion mining and sentiment analysis. In Mining Text Data (pp. 415-463). Springer, Boston, MA. doi: 10.1007/978-1-4614-3223-4_12

Abiteboul, S., Buneman, P., & Suciu, D. (2000). Data on the Web: From Relations to Semistructured Data and XML. Morgan Kaufmann Publishers. ISBN: 978-1558606222.

Rahm, E., & Do, H. (2000). Data Cleaning: Problems and Current Approaches. IEEE Data Engineering Bulletin, 23(4), 3-13.

International Data Corporation. (2021). Global DataSphere Forecast. Recuperado de IDC: https://www.idc.com/