ANGELO CESAR COLOMBINI

RELATÓRIO FINAL DE PÓS-DOUTORADO

Extração de Conhecimento em Big Data

Relatório de Pós-doutorado realizado na Universidade Estadual Paulista (UNESP), DCCE – IBILCE – UNESP / São José do Rio Preto.

Supervisor(a): Prof. Dr. Carlos Roberto Valêncio

No do processo: 4176

São José do Rio Preto

2024

Sumário

1. INTRODUÇÃO

Na vanguarda da revolução digital, o fenômeno do Big Data surge como um catalisador de transformações abrangentes, estendendo-se muito além do vasto volume de dados para incorporar a velocidade vertiginosa de sua geração, a diversidade de formatos e o imenso valor intrínseco que tais informações carregam. Neste panorama, enfrentar os desafios de gerenciamento e análise de Big Data torna-se não apenas crucial, mas também uma oportunidade para inovações disruptivas em múltiplos setores da sociedade.

O papel estratégico do Big Data como um ativo fundamental é cada vez mais reconhecido, tanto no contexto empresarial, orientando decisões estratégicas, quanto no acadêmico, estimulando o desenvolvimento de teorias e metodologias analíticas inéditas. A complexidade técnica envolvida no armazenamento, no processamento e na análise de conjuntos de dados de grande escala é a força motriz para extrair valor e *insights* práticos de informações aparentemente complexos.

Simultaneamente, as dimensões sociais e éticas associadas ao Big Data emergem com força total, destacando-se questões de privacidade, segurança e equidade algorítmica. A concentração nesses elementos é crucial para estabelecer uma governança de dados que promova a confiança e equidade na era digital. As inovações tecnológicas, sobretudo em computação em nuvem, aprendizado de máquina e inteligência artificial, surgem como ferramentas poderosas para navegar neste oceano de dados, simplificando sua manipulação e potencializando a análise.

A aplicabilidade prática do Big Data manifesta-se por meio de uma ampla gama de domínios, desde a melhoria dos cuidados de saúde até a otimização de operações financeiras, cada um ilustrando o valor tangível que a análise de dados pode oferecer. No contexto acadêmico, a pesquisa em Big Data impulsiona a ciência de dados, e forja novos

frameworks e metodologias analíticas, fomentando condições para descobertas científicas e avanços tecnológicos.

Este relatório sintetiza os desafios e avanços identificados durante o projeto de pesquisa, com um olhar atento aos dados de saúde complexos e multifacetados. Os aportes delineados neste texto visam não somente aprimorar o arcabouço teórico e aplicado, mas também ressaltar a centralidade dos dados no embasamento de decisões conscientes e na condução de pesquisas científicas de vanguarda.

Através deste trabalho, delineia-se um futuro no qual o Big Data é não apenas uma entidade massiva, mas uma fonte inesgotável de conhecimento e inovação, cuja influência se estende por todas as esferas da atividade humana.

2. OBJETIVO GERAL

Este relatório tem como objetivo sintetizar as pesquisas realizadas no âmbito da preparação, limpeza, mineração e visualização de dados, destacando sua importância em contextos que desafiam os paradigmas de armazenamento tradicionais. A narrativa detalhada da revisão bibliográfica extensiva (resumida em tópico aqui, mas completa nos artigos no anexo) apoiam forntemente os resultados desse trabalho, estabelecendo um contraponto entre os métodos existentes e as novas abordagens emergentes.

O relatório foca na análise comparativa de algoritmos, investigando suas capacidades e limitações, e nas tecnologias emergentes, que são fundamentais para o avanço do campo do Big Data. Ao articular os avanços teóricos e práticos obtidos, o documento proporciona *insights* inovadores e soluções aplicáveis que demonstram um potencial considerável para a transformação de dados em um recurso estratégico e operacional.

A síntese das investigações vai além da mera compilação de resultados; ela reflete uma integração criteriosa de conhecimento multidisciplinar, evidenciando como a ciência de dados, impulsionada pelas tecnologias de armazenamento e análise juntas podem ampliar o limite do possível. A contribuição deste relatório ao campo do Big Data é, portanto, duplamente valiosa: promove novas frentes para práticas mais eficientes e inovadoras em armazenamento de dados e abre novas pespectivas de pesquisa que podem elucidar conhecimentos ocultos nos vastos volumes de dados característica marcante da era digital.

2.1 Objetivos Específicos Alcançados

O projeto atingiu alguns objetivos específicos notáveis, que contribuem significativamente para a ciência de dados aplicada ao Big Data e suas interseções com áreas críticas como a saúde.

1. **Extração de Conhecimento em Big Data**: a pesquisa avançou no desenvolvimento de metodologias de limpeza, prospecção e visualização de dados. A aplicação dessas técnicas a sistemas de saúde foi metodicamente testada, resultando na validação e aperfeiçoamento de novas metodologias. Este processo não apenas refinou as técnicas de gestão de dados em larga escala, mas também estabeleceu novos *benchmarks* para a eficácia em ambientes complexos de dados de saúde.
2. **Armazenamento de Dados Não Convencionais**: a investigação dedicou-se ao exame de arquiteturas de armazenamento alternativas. Ao fazer isso, o estudo revelou abordagens inovadoras, cada uma desenhada para superar os desafios impostos pelo aumento exponencial no volume e na complexidade dos dados. Essas novas abordagens foram meticulosamente documentadas e contribuíram para o estabelecimento de novos paradigmas de armazenamento de dados, propostos em publicações científicas.

Esses avanços foram seguidos por uma análise rigorosa dos processos e práticas atuais, destacando onde as melhorias podem ser implementadas e como as novas técnicas podem ser escaladas e adaptadas para outros domínios. Além disso, esses trabalhos podem alavancar tecnologias emergentes e oferecer um caminho para pesquisas futuras, desenvolver e implementar soluções inovadoras que abordem as necessidades em evolução do campo de Big Data.

3. METODOLOGIA APLICADA

A metodologia implementada foi concebida para abordar a complexidade e a variedade inerentes aos conjuntos de dados de Big Data, com uma atenção especial aos dados de saúde. Esta abordagem multifacetada englobou várias etapas, cada uma meticulosamente desenhada para garantir que os resultados alcançados fossem não apenas teoricamente sólidos, mas também práticos e aplicáveis.

1. **Desenvolvimento de Metodologias de Limpeza de Dados**: a pesquisa aprofundou-se na criação de processos de limpeza de dados que vão além da remoção de imprecisões, focando na melhoria da qualidade dos dados para análise subsequente. Isto envolveu a utilização de algoritmos avançados para a identificação e correção de inconsistências, o que é particularmente crucial em dados de saúde, onde a precisão é vital.
2. **Prospecção de Dados e Visualização Avançada**: a prospecção de dados foi conduzida com o auxílio de técnicas de mineração de dados de última geração, que permitiram a identificação de padrões significativos e tendências ocultas. A visualização de dados foi aprimorada através do uso de ferramentas de mineração

visual de dados, que transformaram conjuntos de dados complexos em representações visuais intuitivas, facilitando a interpretação e análise.

3. **Exploração de Arquiteturas de Armazenamento Alternativas**: o projeto explorou soluções inovadoras para o armazenamento de dados não convencionais. Foram estudadas e comparadas diferentes arquiteturas de armazenamento, avaliando sua capacidade de lidar com as demandas de Big Data, particularmente em termos de escalabilidade, desempenho e segurança.

Cada um desses métodos foi testado e refinado através de um processo iterativo que garantiu que as soluções propostas não apenas atendessem aos requisitos teóricos, mas também superassem os desafios práticos encontrados em ambientes de dados reais. A metodologia adotada proporcionou uma base sólida para a continuidade do trabalho, sugerindo áreas para futuras investigações e desenvolvimentos no campo de Big Data.

4. REVISÃO DA LITERATURA

Aqui apresenta-se um compilado da revisão da literatura aplicada na confecção dos artigos, como esse tema seria longo e repetitivo, apenas uma compilação será apresentada, deixando a pesquisa completa para os artigos anexados neste relatório.

Artigo 1 - Generation of new bilingual dictionaries through inference techniques supported by a Graph Oriented Database Management System

Nesta revisão da literatura o foco foi na busca pelo estado da arte em conceitos-chave e estudos relevantes em na área da Linguística Computacional, notadamente em tradução léxica, representação de dicionários bilíngues através de grafos e a prática de inferência para descoberta de novas traduções. Também procurou se aprofundar na pesquisa em áreas que potencializaram o trabalho, como a área de Linguística Computacional explorando suas suas principais subáreas, com especial atenção à tradução automática e sua aplicação prática em dicionários online. Discutiu-se também como os dicionários são estruturados como grafos, permitindo uma representação dinâmica das relações lexicais e facilitando o armazenamento e a inferência em bancos de dados orientados a grafos.

Identificou-se também através do levantamento bibliográfico alguns desafios e as soluções para a ambiguidade lexical e a polissemia na tradução automática. O uso de sistemas como o Neo4j é mencionado para a otimização do armazenamento e da exploração de dados. A plataforma Apertium foi identificada como um caso de uso no contexto da tradução automática, destacando-se a conversão de dicionários bilíngues para o modelo RDF e sua integração na Linguistic Linked Open Data Cloud, promovendo interoperabilidade de dados.

Além das pesquisas que direcionaram o trabalho, pesquisou-se também trabalhos correlatos, ressaltando metodologias de inferência avançadas em dicionários bilíngues/multilíngues e enfatizando as contribuições significativas ao campo da tradução automática. Foram analisados estudos que propõem novas técnicas de inferência e heurísticas para enriquecimento de dicionários e geração de novos dicionários, oferecendo uma visão ampla do estado da arte e direções futuras para pesquisa e também foi possível enquadrar de forma mais precisa as contribuições desse trabalo.

Artigo 2 – Spatial Data Mining algorithm: a proposal with Big Data characteristics

Este artigo analisa uma série de trabalhos com foco na aplicação de técnicas de paralelização, como o método MapReduce, para otimizar o desempenho em algoritmos de mineração de dados espaciais. O estudo de Yue et al. (2016) introduz um algoritmo de agrupamento usando K-Medoids, que é robusto em relação ao ruído. Foi necessário implementar o MapReduce para melhorar seu desempenho em grandes conjuntos de dados. Manjula e Narsimha (2018) desenvolveram um processo eficaz de classificação de solo usando uma base de dados de solo e otimização por rede neural. Daniel (2016) apresenta dois algoritmos, VDBSCAN-MR e OVDBSCAN-MR, para melhorar desempenho e escalabilidade em grandes bases de dados. Por último, Sheena Smart et al. (2021) propõem uma nova solução para classificação de dados espaciais que aplica uma técnica baseada em Pareto Exponentiated, resultando em uma classificação eficiente e uma redução na complexidade do tempo.

Artigo 3 - Algoritmo para data cleaning baseado em séries temporais

Este artigo analisa uma série referências abordando tópicos como a aplicação de séries temporais e sua análise em diversos campos, evidenciando a importância da precisão dos dados, especialmente os coletados por sensores.

Discute também a limpeza de dados e a detecção de *outliers* destacando técnicas relevantes ao próposito do trablaho, como técnicas de suavização, análise estatística e restrições, bem como a complexidade de tratar grandes volumes de dados com variabilidade e potenciais erros.

A escalabilidade é outro ponto relevante abordado no levantamento bibliográfico tendo sido explorado como elemento essencial para processamento em Big Data, com o Apache *Spark* exemplificando uma

ferramenta poderosa para gestão de dados em grande escala. Aborda discussões a cerca de algoritmos avançados para detecção de *outliers* e estratégias de limpeza de dados em séries temporais, ressaltando a necessidade de integridade e minimização da perda de informação durante a limpeza, e a utilidade de identificar dados discrepantes para aprimorar modelos de predição e análise.

5. RESULTADOS ALCANÇADOS E IMPACTOS RELACIONADOS AOS TRABALHOS

5.1 RESULTADOS E IMPACTOS DO ARTIGO - GENERATION OF NEW BILINGUAL DICTIONARIES THROUGH INFERENCE TECHNIQUES SUPPORTED BY A GRAPH ORIENTED DATABASE MANAGEMENT SYSTEM

- O artigo trouxe contribuições significativas no campo da linguística computacional, especificamente na criação de dicionários bilíngues por meio de técnicas de inferência suportadas por um Sistema de Gerenciamento de Banco de Dados Orientado a Grafos (GDMS). A robustez dos algoritmos utilizados possibilitou não apenas um aumento na quantidade de traduções disponíveis, mas também uma melhoria substancial na qualidade das conexões lexicais.
- Ao identificar relações lexicais complexas com um grau de precisão elevado, o estudo expandiu de maneira notável os dicionários existentes, indicando um avanço metodológico significativo. A implementação no GDMS Neo4j resultou em melhorias na eficiência dos processos de busca e recuperação de traduções, com a estrutura de dados baseada em grafos provando ser excepcionalmente adequada para manipular a complexidade dos dados lexicais.

- O impacto dos resultados se estende para além do acadêmico, influenciando o desenvolvimento de ferramentas aplicadas, como softwares de tradução automática e aplicativos de aprendizagem de idiomas, e pavimentando o caminho para novas pesquisas onde a inferência baseada em dados ocupa uma posição central.
- A validação dos dicionários inferidos foi realizada com rigor, utilizando comparação com padrões de dicionários estabelecidos e análise manual, garantindo a coerência e relevância contextual das novas traduções. Esta etapa confirmou a utilidade prática dos dicionários inferidos, atestando a eficácia da metodologia adotada.
- O estudo também destaca o potencial das abordagens baseadas em grafos para resolver problemas linguísticos, demonstrando a capacidade de gerar novo conhecimento a partir de dados pré-existentes, um marco de eficiência e inovação.
- Em suma, o artigo não apenas avança o entendimento dos métodos de inferência em tradução, mas também estabelece precedentes para futuras investigações, explorando as potencialidades dos GDMS na área de tradução e processamento de linguagem natural.

5.2 Resultados e impactos do artigo - Spatial data mining algorithm: a proposal with Big Data characteristics

- O artigo apresentou avanços notáveis no campo do Big Data, particularmente no processamento e análise de dados espaciais. O estudo desenvolveu e aplicou um novo algoritmo, com base na técnica MapReduce e no ambiente de computação distribuída Apache Spark, para otimizar a detecção de padrões e a formação de clusters em grandes conjuntos de dados espaciais.

- A pesquisa demonstrou que a aplicação de técnicas de paralelização, em conjunto com o algoritmo proposto, resultou em uma redução significativa do tempo de processamento em comparação com métodos anteriores. A implementação eficaz do MapReduce resultou em uma diminuição marcante na quantidade de dados a serem processados, o que se refletiu em uma melhoria na eficiência geral do algoritmo, especialmente relevante em contextos de Big Data com atributos espaciais complexos.
- Os impactos deste estudo são profundos, pois fornecem uma abordagem escalável e eficiente para a análise de dados espaciais, fundamental para campos como a geoinformática, planejamento urbano e ambiental, e outras áreas que dependem da análise de grandes volumes de dados geoespaciais. As técnicas desenvolvidas neste trabalho têm o potencial de serem adaptadas para outros contextos de Big Data, melhorando a capacidade de extração de conhecimento e insights a partir de dados complexos e variados.

4.5. RESULTADOS E IMPACTOS DO ARTIGO - ALGORITMO PARA DATA CLEANING BASEADO EM SÉRIES TEMPORAIS

- O artigo constitui um marco no aprimoramento dos processos de limpeza de dados em séries temporais, com ênfase particular na detecção e correção de outliers. Através da implementação de técnicas computacionais avançadas, utilizando o Apache Spark, o estudo alcançou uma metodologia que processa grandes volumes de dados com rapidez e precisão notáveis.
- Este progresso é essencial no universo do Big Data, especialmente em áreas que se apoiam em séries temporais, como os mercados financeiros e o monitoramento de

infraestruturas. A metodologia desenvolvida não apenas melhora o armazenamento e a análise de dados, mas também eleva o padrão de confiabilidade para decisões baseadas em dados.

- O impacto deste trabalho se manifesta na melhoria da integridade dos dados em Big Data. O uso de computação paralela e distribuída traz um avanço em eficiência e expande a capacidade de análise de conjuntos de dados volumosos. Ao assegurar a precisão dos dados após a limpeza de outliers, as análises subsequentes se baseiam em informações confiáveis, fundamentais para decisões críticas em setores sensíveis como finanças e saúde.
- Além disso, a metodologia oferece uma abordagem escalável, adaptável a outros contextos de Big Data, o que reflete a adaptabilidade e a escalabilidade essenciais em sistemas de dados extensivos. Portanto, o estudo apresentado no Artigo 3 estabelece um novo padrão para a integridade dos dados, incentivando avanços em pesquisa e práticas de Big Data, e reforçando a confiança e a utilidade dos insights derivados de dados complexos.

6. CONCLUSÃO

Este projeto de pesquisa trouxe contribuições siginificativas no campo de Big Data, com repercussões específicas na área da saúde, que se beneficia imensamente da precisão e eficiência dos avanços alcançados. As metodologias inovadoras aqui apresentadas, que abrangem desde a inferência e meta-heurística até a limpeza de dados e visualização avançada, não apenas cumprem as expectativas iniciais, como também superam, promovendo um novo patamar de excelência e aplicabilidade.

As estratégias desenvolvidas, testadas e validadas, destacam-se por sua escalabilidade e flexibilidade, características fundamentais que permitem sua adaptação a uma multiplicidade de cenários de dados. Este aspecto é muito importante, pois o volume e a complexidade dos dados em Big Data continuarão a crescer, exigindo soluções que possam ser customizadas para atender às necessidades emergentes.

Além disso, a contribuição acadêmica deste projeto não pode ser subestimada. Ao elevar o padrão de rigor e inovação, o trabalho pode ser visto como um referencial para futuras pesquisas, incentivando outros cientistas de dados a desenvolverem soluções que abordem problemas atuais e futuros com igual diligência e criatividade. A aplicabilidade prática das contribuições deste projeto são evidentes, mas seu verdadeiro valor será percebido à medida que essas técnicas forem adaptadas e integradas em novos sistemas e processos, essa escalabilidade é uma contribuição intresseca deste projeto que não havia sido prevista no seu início.

O diálogo entre as descobertas práticas e teóricas emergentes deste trabalho ilustram uma simbiose entre teoria e prática. As soluções propostas para os desafios de Big Data não apenas moldam o atual panorama da análise de dados, mas também os fundamentos para uma transformação mais ampla em diversas disciplinas que dependem da interpretação e gestão eficaz de grandes conjuntos de dados.

Num olhar de futuro, o projeto não apenas estabelece uma direção para pesquisas subsequentes, mas também desafia o *status quo*, incentivando a adoção de tecnologias emergentes e metodologias avançadas na rotina da análise de dados.

Os esforços contínuos neste campo serão essenciais para resolver questões complexas de Big Data, as quais terão um papel central na inovação e no progresso em várias esferas das atividades humana.

Em suma, o projeto de pesquisa alcançou seus objetivos com distinção, transcendendo-os em algumas áreas, e contribuindo substancialmente para a ciência de dados. As contribuições deste trabalho irão

indubitavelmente apoiar e inspirar as pesquisas na área, as quais continuarão a trabalhar com o incrível desafio da busca do conhecimento e insights do crescente volume de dados Big Data, buscando soluções que impulsionem ainda mais a fronteira do conhecimento e da inovação tecnológica.

7. REFERÊNCIAS

[1]. Aggarwal, C. C. and Yu, P. S. (2001). Outlier detection for high dimensional data. Proceedings of the ACM SIGMOD International Conference on Management of Data.

[2]. Aggarwal, C. C. and Yu, P. S. (2008). Outlier detection with uncertain data. Society for Indus- trial and Applied Mathematics - 8th SIAM International Conference on Data Mining 2008, Proceedings in Applied Mathematics 130, 2.

[3]. Alocci, D., Mariethoz, J., Horlacher, O., Bol- leman, J. T., Campbell, M. P., and Lisacek, F. (2015). Property graph vs rdf triple store: A comparison on glycan substructure search. PLoS ONE, 10.

[4]. Alper, M. (2017). Auto-generating bilingual dictionaries: Results of the tiad- 2017 shared task baseline algorithm. CEUR Workshop Proceedings, 1899.

[5]. Angles, R. and Gutierrez, C. (2018). An Introduction to Graph Data Management: Fundamental Issues and Recent Developments. Angles, R., Thakkar, H., and Tomaszuk, D. (2020). Mapping rdf databases to property graph databases. IEEE Access, 8.

[6]. Apache (2021). Apache. Brillinger, D. R. (2001). Time Series: Data Analysis and Theory, volume 1. Society for Industrial and Applied Mathematics.

[7]. Baldwin, T., Pool, J., and Colowick, S. M. (2010). Panlex and lextract: Translating all words of all languages of the world.

[8]. Coling 2010 - 23rd International Conference on Computational Linguistics, Proceedings of the Conference, 2.

[9]. Chiarcos, C., Klimek, B., F¨ath, C., Declerck, T., and McCrae, J. P. (2020). On the linguistic linked open data infrastructure. Proceedings of the 1st International Workshop on Language Technology Platforms.

[10]. Cheragui, M. A. (2012). Theoretical overview of machine translation. 867.

[11]. Chomicki, J. and Marcinkowski, J. (2005). Minimal-change integrity maintenance using tuple deletions. Information and Computation, 197:90–121.

[12]. Cimiano, P., Chiarcos, C., McCrae, J. P., and Gracia, J. (2020). Linguistic linked open data cloud.

[13]. COLING 2016 - 26th International Conference on Computational Linguistics, Proceedings of COLING 2016: Technical Papers.

[14]. DANIEL, G. P. "Otimização de algoritmos de agrupamento espacial baseado em densidade aplicados em grandes conjuntos de dados." (2016).

[15]. Dasu, T., Duan, R., and Srivastava, D. (2016). Data quality for temporal streams. IEEE Computer Society Technical Committee on Data Engineering. Duggimpudi, M. B., Abbady, S., Chen, J., and Raghavan, V. V. (2019).

[16]. Demsar, U., Harris, P., Brunsdon, C., Fotheringham, A. S., & McLoone, S. (2013). Principal component analysis on spatial data: an overview. Annals of the Association of American Geographers, 103(1), 106-128.

[17]. Dranca, L. (2020). Multi-strategy system for translation inference across dictionaries. Dunkel, P. and Grishman, R. (1987). Computational linguistics: An introduction. TESOL Quarterly, 21.

[18]. Dziemianko, A. (2017). Dictionary form in decoding, encoding and retention: Further insights. ReCALL, 29.

[19]. Fagin, R., Kimelfeld, B., and Kolaitis, P. G. (2015). Dichotomies in the complexity of prefer- red repairs. Proceedings of the ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, 31.

[20]. Fox, A. J. (1972). Outliers in time series. Journal of the Royal Statistical Society: Series B (Methodological), 34:350–363.

[21]. Gardner, E. S. (2006). Exponential smoothing: The state of the art-part ii. International Journal of Forecasting, 22:637–666.

[22]. Garrido, C. and Gutierrez, C. (2016). Dictionaries as networks: Identifying the graph struc- ture of ogden's basic english.

[23]. GOEL, S., GRACIA, J., and L., F. M. (2021). Bilingual dictionary generation and enrichment via graph exploration.

[24]. Gracia, J., Kabashi, B., Kernerman, I., Lanau Coronas, M., and Lonke, D. (2019). Results of the translation inference across dictionaries 2019 shared task. volume 2493.

[25]. Gracia, J., Villegas, M., G´omez-P´erez, A., and Bel, N. (2018). The apertium bilingual dictionaries on the web of data. Semantic Web, 9.

[26]. Gupta, M., Gao, J., Aggarwal, C. C., and Han, J. (2014). Outlier detection for temporal data: A survey. IEEE Transactions on Knowledge and Data Engineering, 26.

[27]. Have, C. T., Jensen, L. J., and Wren, J. (2013). Are graph databases ready for bioinformatics? Bioinformatics, 29.

[28]. Jeffery, S. R., Alonso, G., Franklin, M. J., Hong, W., and Widom, J. (2006). Declarative sup- port for sensor data cleaning. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture No- tes in Bioinformatics), 3968 LNCS:83–100.

[29]. Hawkins, D. M. (1980). Identification of Outliers. Chapman and Hall. Hill, D. J. and Minsker, B. S. (2010). Anomaly detection in streaming environmental sensor data: A data-driven modeling approach. Environ- mental Modelling and Software, 25:1014–1022.

[30]. Kaisler, S., Armour, F., Espinosa, J. A., & Money, W. (2013, January). Big data: Issues and challenges moving forward. In 2013 46th Hawaii international conference on system sciences (pp. 995-1004). IEEE.

[31]. Karkouch, A., Mousannif, H., Moatassime, H. A., and Noel, T. (2016). Data quality in internet of things: A state-of-the-art survey. Journal of Network and Computer Applications, 73:57–71.

[32]. Keogh, E., Lin, J., and Fu, A. (2005). Hot sax: Efficiently finding the most unusual time series subsequence. Proceedings - IEEE International Conference on Data Mining, ICDM.

[33]. Maimon, O. Z. (2005). Data mining and knowledge discovery handbook. Springer. Mavridis, I. and Karatza, H. (2017). Performance evaluation of cloud- based log file analysis with apache hadoop and apache spark. Journal of Systems and Software, 125.

[34]. Matsumoto, S., Yamanaka, R., and Chiba, H. (2018). Mapping rdf graphs to property graphs. CEUR Workshop Proceedings, 2293.

[35]. McCrae, J., de Cea, G. A., Buitelaar, P., Cimi- ano, P., Declerck, T., G´omez-P´erez, A., Gracia, J., Hollink, L., Montiel-Ponsoda, E., Spohr, D., and Wunner, T. (2012). Interchanging lexical resources on the semantic web. Language Re- sources and Evaluation, 46.

[36]. Medeiros, C. A. D. (2014). Extração de conhecimento em bases de dados espaciais: algoritmo CHSMST+.

[37]. Oujdi, Sihem, Hafida Belbachir, and Faouzi Boufares. "C4. 5 decision tree algorithms for spatial data, alternatives and performances." Journal of computing and information technology 27.3 (2019): 29-43.

[38]. Proisl, T., Heinrich, P., Evert, S., and Kabashi, B. (2017). Translation inference across dictionaries via a combination of graph-based methods and co- occurrence statistics. volume 1899.

[39]. Punnoose, R., Crainiceanu, A., and Rapp, D. (2012). Rya: A scalable rdf triple store for the clouds. ACM International Conference Proceeding Series.

[40]. Rahm, E. and Do, H. (2000). Data cleaning: Problems and current approaches. IEEE Data Eng. Bull., 23.

[41]. Ramírez-Gallego, S., Fernández, A., García, S., Chen, M., & Herrera, F. (2018). Big Data: Tutorial and guidelines on information and process fusion for analytics algorithms with MapReduce. Information

Fusion, 42, 51-61. Academy of Natural Sciences: MAL. Available in: https://doi.org/10.15468/dl.8nxkz6. Access in 01 dec. 2022.

[42]. Sheena Smart, P. D., Thanammal, K. K., & Sujatha, S. S. (2021). A novel linear assorted classification method-based association rule mining with spatial data. Sādhanā, 46(1), 1-12.

[43]. Singh, D. and Reddy, C. K. (2015). A survey on platforms for big data analytics. Journal of Big Data, 2:1–20.

[44]. Song, S., Zhang, A., Wang, J., and Yu, P. S. (2015). Screen: Stream data cleaning under speed constraints. Proceedings of the ACM SIGMOD International Conference on Management of Data, 2015-May:827–841

[45]. Spatio-temporal outlier detection algorithms based on computing behavioral outlierness factor. Data and Knowledge Engineering, 122:1–24.

[46]. Tanaka, K. and Umemura, K. (1994). Construction of a bilingual dictionary intermediated by a third language.

[47]. Vilar, D., Xu, J., D'Haro, L. F., and Ney, H. (2006). Error analysis of statistical machine translation output. Proceedings of the 5th International Conference on Language Resources and Evaluation, LREC 2006.

[48]. Villegas, M., Melero, M., Gracia, J., and Bel, N. (2016). Leveraging rdf graphs for crossing multiple bilingual dictionaries. Proceedings of the 10th International Conference on Language Re- sources and Evaluation, LREC 2016.

[49]. Wang, X. and Wang, C. (2020). Time series data cleaning: A survey. IEEE Access, 8. Wei, L., Keogh, E., and Xi, A. (2006). Saxually explicit images: Finding unusual shapes. Proceedings - IEEE International Conference on Data Mining, ICDM.

[50]. Yue, X., Man, W., Yue, J., & Liu, G. (2016). Parallel k-medoids++ spatial clustering algorithm based on mapreduce arXiv preprint arXiv:1608.06861. Manjula, Aakunuri, and G. Narsimha. "Using an

efficient optimal classifier for soil classification in spatial data mining over big data." Journal off Intelligent Systems 29.1 (2020): 172-188.

[51]. Zhuang, Y., Chen, L., Wang, X. S., and Lian, J. (2007). A weighted moving average-based approach for cleaning sensor data. Proceedings - International Conference on Distributed Computing Systems, page 38.

8. ANEXOS – ARTIGOS ACEITOS E EM CORREÇÃO

ARTIGO1 - aceito para ser publicado no evento: ICEIS2023.

Termo de aceite: 13/09/2023 – ICKG 2023 SUBMISSION kG252 – *accepted*.

Generation of new bilingual dictionaries through inference techniques supported by a Graph Oriented Database Management System

João Pedro Quadrado
Department of Computer Science and Statistics
IBILCE UNESP Universidade Estadual Paulista - Júlio de Mesquita Filho
São José do Rio Preto-SP, Brazil
jp.quadrado@unesp.br

Carlos Roberto Valêncio
Department of Computer Science and Statistics
IBILCE UNESP Universidade Estadual Paulista - Júlio de Mesquita Filho
São José do Rio Preto-SP, Brazil
carlos.valencio@unesp.br

Adriane Orenha Ottaiano
Department of Modern Letters
IBILCE UNESP Universidade Estadual Paulista - Júlio de Mesquita Filho
São José do Rio Preto-SP, Brazil
adriane.ottaiano@unesp.br

Geraldo Francisco D. Zafalon
Department of Computer Science and Statistics
IBILCE UNESP Universidade Estadual Paulista - Júlio de Mesquita Filho
São José do Rio Preto-SP, Brazil
geraldo.zafalon@unesp.br

Luís Marcello Moraes Silva
Department of Computer Science and Statistics
IBILCE UNESP Universidade Estadual Paulista - Júlio de Mesquita Filho
São José do Rio Preto-SP, Brazil
luis.marcello@unesp.br

Angelo Cesar Colombini
Department of Electrical Engineering
UFF Universidade Federal Fluminense
Niterói-RJ, Brazil
accolombini@id.uff.br

***Abstract*—The popularization of the internet allowed the use of different technologies for storing and retrieving bilingual dictionaries. An example is the Resource Description Framework (RDF) datasets. The storage of bilingual dictionaries in this type of structure makes it possible to discover indirectly connected translations through inference. However, such datasets are not efficient in graph search operations, which are important actions for inference techniques in the discovery of translations. On the other hand, database management systems based on the graph model offer a collection of resources that facilitate the tasks that involve such operations. In this way, a group of bilingual dictionaries from an RDF dataset was adapted to a Graph Database Management System (GDMS) which made it possible to reduce and simplify the structures of these dictionaries. Subsequently, 2 translation inference techniques were implemented in which two new bilingual English-Portuguese dictionaries were automatically generated. Thus, such inference algorithms are efficient in generating new dictionaries through GDMS. The scientific contribution is the application of a new technology for data storage and retrieval in support of bilingual dictionaries, in which there are advantages together with inference techniques in the generation of bilingual dictionaries and not addressed by related papers in literature.**

***Keywords*—Bilingual dictionaries, Inference techniques, Translation graph, RDF dataset, Graph Oriented Database Management Systems.**

I. INTRODUCTION

Dictionaries are essential tools in learning, whether to understand the meaning or spelling of a word in monolingual dictionaries, or to identify the translation of an expression in bilingual and multilingual dictionaries. The popularization of the internet allowed the development of online dictionary applications, with the possibility of using different technologies for the storage and retrieval of your data [1].

One of the ways to store the information present in bilingual dictionaries is the use of Resource Description Framework (RDF) datasets, which are graph data managers [2]. An example is the dictionaries found in the Linguistic Linked Open Data Cloud (LLODC), an infrastructure that stores several databases of dictionaries, fragments of written texts and metadata [3]. The storage of bilingual dictionaries allows their structure to be explored in different ways, one of which is the discovery of new translations that are indirectly connected, also known as translation by inference [4].

One of the problems RDF datasets have to deal with is the lack of optimizations for traversal operations on graphs, operations that are present in methods to infer new translations [5]. On the other hand, Graph-oriented Data Management System (GDMS) are dedicated to operations on graphs and used when the dimension of the graph data structure gains important proportions [6]. One can even use the properties of the GDMS to reduce the model of bilingual dictionaries, where their information is better compressed. Furthermore, because the GDMS stores a graph model, it is also possible that different inference techniques are applied in their structure in the generation of new translations based on existing translations. This makes the construction of new bilingual dictionaries or the enrichment of existing bilingual dictionaries possible [4].

A. Objectives and Methods

Given the importance of bilingual dictionaries and their exploration for the discovery of new translations, the goals of this article are:

ARTIGO 2 – submetido para no dia 31/01/2024 para o ICEIS – sem retorno até o momento.

Spatial data mining algorithm: a proposal with Big Data characteristics

Wellington Reguera Gouveia[1], Carlos Roberto Valêncio[1][a], Geraldo Francisco Donega Zafalon[1][b], Luís Marcello Moraes Silva[1][c], Angelo Cesar Colombini[2][d] and Márcio Zamboti Fortes[2]
[1]IBILCE UNESP Universidade Estadual Paulista - Júlio de Mesquita Filho, R. Cristóvão Colombo 2265 - Jardim Nazareth, São José do Rio Preto - SP, Brazil
[2]UFF Universidade Federal Fluminense, R. Mário Santos Braga 30 - Centro, Niterói - RJ, Brazil
{wellington.reguera, carlos.valencio, geraldo.zafalon, luis.marcello}@unesp.br,
{accolombini,mzamboti}@id.uff.br

Keywords: CHSMST+MR, Spatial Data Mining, Clustering algorithm based on Hyper Surface and Minimum Spanning Tree - CHSMST, CHSMST+.

Abstract: Data mining in spatial Big Data is more complex compared to conventional data mining because of intrinsic relations, self-correlation and also the difficulty of extracting patterns from this data. The massive volume of spatial data information generates problems such as exponential growth of data volume, contamination and computational complexity. In this sense, parallelization techniques are constantly used in order to optimize the processing time of this large amount of data. Thus, to provide some contribution within this context, this paper proposes a new algorithm called CHSMST+MR, based on CHSMST, where distributed and parallel processing techniques were used. In comparison with its literature counterpart, an average reduction of approximately 51.36% in the time of the proposed algorithm was observed. Therefore, it was intended as a scientific contribution, the use of distribution and parallelization to create a solution with superior performance in order to extract behaviors and patterns in spatial data and with Big Data characteristics, given that, since more complex data is being dealt with, it requires more computational processing resources.

1 INTRODUCTION

The term Big Data is defined as the amount of data that goes beyond the ability of the technology to efficiently store, manage and process it. These actions are only revealed by a robust analysis of its own data, which requires multiple amounts of data processing and the use of tools, whether hardware, software or methods, which would allow the complete analysis of its own information. As it often happens when one faces a problem, the process of figuring out how to proceed can lead to the creation of new tools to perform new tasks. (Kaisler et al, 2013).

One of the types of data that benefit from the Big Data area is spatial data, which represent all types of data objects present in a geographic space. Their classification depends on the analysis of spatial objects associated with their spatial characteristics, i.e. areas of the region, roads, lakes or rivers. (Sheena; Thanammal; Sujatha, 2020). Spatial objects are characterized by a geometric representation and relative positioning that implicitly define both spatial relations and properties. In addition, these types of information are organized into a set of spatially linked thematic layers, which represent discrete or continuous features. (Oujdi; Belbachir; Boufares, 2019).

Furthermore, spatial data mining is the process of extracting knowledge and relations, which are stored in spatial databases. Extracting valuable patterns from spatial datasets is more difficult due to the complexity of data types, relations and autocorrelation (Sheena; Thanammal; Sujatha, 2020).

However, such type of mining is largely derived from conventional data mining techniques for spatial classification, which can be used to explain or predict a phenomenon based on the analysis of the properties of the geographic

[a] https://orcid.org/0000-0002-9325-3159
[b] https://orcid.org/0000-0003-2384-011X
[c] https://orcid.org/0000-0002-7201-1257
[d] https://orcid.org/0000-0002-8906-4128

ARTIGO 3 – artigo sendo preparado para submissão – em fase final de revisão.

Algoritmo para data cleaning baseado em séries temporais

Beatriz Ferreira de Lima, Carlos Roberto Valêncio *
Angelo Cesar Colombini, Geraldo Francisco Donega Zafalon
Márcio Zamboti Fortes

Resumo

As séries temporais têm se mostrado presentes em diversas áreas de grande valor econômico, como o mercado de ações e a indústria. Em circunstância de características Big Data, pode-se ter os dados processados em tempo real (*real time*) ou quase em tempo real (*near real time*), e fontes capazes de gerar volumes elevados de dados, o que impõe a fase de preparação a necessidade de execução da limpeza destes dados de forma eficaz e eficiente ao lidar com estes requisitos. Existem diversos algoritmos que podem ser utilizados para realizar esse processo, porém esses podem conter limitações como baixo rendimento, distorções dos dados, tempo elevado de processamento, entre outros. Um dos problemas a ser tratado na preparação dos dados é a detecção de *outliers*, dados que podem refletir distorções e implicar em custos adicionais na fase de limpeza dos dados. Assim, este trabalho teve como objetivo propor um algoritmo que realize a detecção de *outliers* e, posteriormente, a limpeza dos dados discrepantes de forma eficaz e eficiente, buscando assegurar a integralidade da informação obtida através dos dados provenientes de séries temporais.

Palavras-chave: Banco de Dados. Séries Temporais. Algoritmos paralelos. Limpeza de dados. Detecção de *outliers*. Apache Spark.

Resumo

Time series have been present in several areas of great economic value, such as the stock market and industry. In circumstances of Big Data characteristics, data can be processed in real time or near real time, and sources capable of generating high volumes of data, which impose the preparation phase to need to perform cleaning of these data effectively and efficiently when dealing with these requirements. There are several algorithms that can be used to carry out this process, but they may have limitations such as low yield, data distortion, high processing time, among others. One of the problems to be addressed in data preparation is the detection of outliers, data that can reflect distortions and that can imply additional costs in the data cleaning phase. Thus, this work aimed to propose an algorithm that performs the detection of outliers and, subsequently, the cleaning of outliers in an effective and efficient way, which sought to maintain the completeness of the information obtained through time series data.

Keywords: Database. Time Series. Parallel Algorithms. Data cleaning. Outlier detection. Apache Spark.

* "Carlos Roberto Valêncio - valencio@ibilce.unesp.br - IBILCE UNESP Universidade Estadual Paulista - Júlio de Mesquita Filho - R. Cristóvão Colombo, 2265 - Jardim Nazareth, São José do Rio Preto - SP, 15054-000; accolombini@id.uff.br - UFF - Universidade Federal Fluminense, geraldo.zafalon@unesp.br - IBILCE UNESP, mzamboti@id.uff.br - UFF"