PRÁTICAS DE MINERAÇÃO DE DADOS: CONSIDERAÇÕES SOBRE OS DADOS GOVERNAMENTAIS ABERTOS EM LONDRINA

Data mining practices: Considerations on open government data in Londrina
Prácticas de minería de datos: consideraciones sobre los datos gubernamentales abiertos en Londrina

Matheus Antunes Palmieri
Graduando, Universidade Estadual de Londrina (UEL), Londrina, Paraná, Brasil.
Lattes: http://lattes.cnpq.br/6625427150387648

Benjamin Luíz Franklin
Doutor, Professor asociado, Universidade Estadual de Londrina (UEL), Londrina, Paraná, Brasil
Lattes: http://lattes.cnpq.br/7192179175216123
ORCID: https://orcid.org/0000-0002-4807-8339

Resumo

Introdução: Este estudo explora o potencial analítico da crescente disponibilidade de documentos públicos online para profissionais da informação, em especial os arquivistas. **Objetivos**: Examinar o conjunto documental do portal de transparência da prefeitura de Londrina, convertendo a informação não estruturada em um Modelo Entidade-Relacionamento e, em seguida, em um modelo de rede. **Metodologia**: Envolve a aplicação de algoritmos de consolidação, agrupamento e centralidade para oferecer uma visão relacional de atores como Documentos, Secretarias e Credores. A pesquisa baseia-se no arcabouço teórico da Mineração de Dados Públicos e seus recursos analíticos. **Resultados**: Sugerem que a transparência na governança deve estar vinculada ao desenvolvimento de competências analíticas pelos profissionais da informação, considerando a crescente disponibilidade de recursos computacionais e tecnológicos. **Conclusão**: A ampliação da transparência e a análise de dados públicos são cruciais para a democracia e a prestação de contas. O município de Londrina pode se beneficiar das práticas de mineração de dados públicos. Futuras pesquisas devem focar no estreitamento do diálogo entre a Arquivologia e os métodos analíticos, visando uma crítica mais aprofundada à dinâmica democrática atual. Isso implica a necessidade de maior colaboração entre as áreas e na aplicação de técnicas analíticas para aprimorar a transparência e a eficácia dos processos de governança.

Palavras-chave: Arquivologia; Ciência de dados; teoria ator-rede; dados governamentais abertos.

Abstract

Introduction: *This study explores the analytical potential of the growing availability of online public records for information professionals, particularly archivists.* ***Objectives***: *To examine the documentary set of the transparency portal of the Londrina city hall, converting unstructured information into an Entity-Relationship Model and then into a network model.* ***Methodology***: *It involves the application of consolidation, clustering, and centrality algorithms to offer a relational view of actors such as Documents, Secretariats, and Creditors. The research is based on the theoretical framework of Public Data Mining and its analytical resources.* ***Results***: *They suggest that transparency in governance must be linked to the development of analytical skills by information professionals, considering the growing availability of computational and technological resources.* ***Conclusion***: *The expansion of transparency and the analysis of public data are crucial for democracy and accountability. The city of Londrina can benefit from public data mining practices. Future research should focus on narrowing the dialogue between Archivology and analytical methods, aimed at a deeper critique of the current democratic dynamics. This implies the need for greater collaboration between the areas and in the application of analytical techniques to improve the transparency and effectiveness of governance processes.*

Keywords: *Archivology; Data science; actor-network theory; government open data.*

Resumen

Introducción: *Este estudio explora el potencial analítico de la creciente disponibilidad de registros públicos en línea para profesionales de la información, en particular los archivistas.* ***Objetivos***: *Examinar el conjunto documental del portal de transparencia de la alcaldía de Londrina, convirtiendo la información no estructurada en un Modelo Entidad-Relación y luego en un modelo de red.* ***Metodología***: *Implica la aplicación de algoritmos de consolidación, agrupación y centralidad para ofrecer una visión relacional de actores como Documentos, Secretarías y Acreedores. La investigación se basa en el marco teórico de la Minería de Datos Públicos y sus recursos analíticos.* ***Resultados***: *Sugiere que la transparencia en la gobernanza debe vincularse al desarrollo de habilidades analíticas por parte de los profesionales de la información, considerando la creciente disponibilidad de recursos computacionales y tecnológicos.* ***Conclusión***: *La expansión de la transparencia y el análisis de datos públicos son fundamentales para la democracia y la rendición de cuentas. El municipio de Londrina puede beneficiarse de las prácticas de minería de datos públicos. La investigación futura debe centrarse en estrechar el diálogo entre la Archivística y los métodos analíticos, con miras a una crítica más profunda de las dinámicas democráticas actuales. Esto implica la necesidad de una mayor colaboración entre las áreas y en la aplicación de técnicas analíticas para mejorar la transparencia y la eficacia de los procesos de gobierno..*

Palabras clave: *Arquivología; Cencia de datos; teoría actor-red; datos gubernamentales abiertos.*

DOI 10.22477/vi.widat.06

1. Introdução

O Portal da Transparência de Londrina visa promover a transparência e o controle social, disponibilizando informações detalhadas sobre orçamentos, gastos públicos e outras áreas, seguindo as exigências da Lei de Acesso à Informação e da Lei da Transparência (LONDRINA, 2022). No entanto, a apresentação desses dados em diferentes formatos cria "silos de informação" (JACINTO *et al.*, 2022), ou seja, aglomerados de dados não estruturados, limitando o acesso e o potencial reuso dos dados e da informação. Assim, os usuários precisam efetuar limpeza, tratamento e interligação com outras fontes de dados para obter conhecimentos mais aprofundados (JACINTO *et al.*, 2022). Este trabalho pretende colaborar na evidenciação dos benefícios de estruturar "silos de informação" para dinamizar e democratizar os processos de governança.

2. Procedimentos Metodológicos

A abordagem social propõe analisar o processo de mineração de dados públicos não apenas como técnico, mas também como social para entender a produção e geração de informações. Latour (2012) sugere determinar inicialmente o grupo e nível de análise a ser enfatizado, ou adotar os procedimentos dos atores envolvidos no domínio. Ao adotar tais procedimentos, é possível identificar protagonistas do movimento de formação e desmantelamento de grupos e contribuir para o empoderamento da sociedade, fiscalizando a transparência governamental.

Desse modo, a análise computacional de conteúdo é utilizada para mapear os relacionamentos de atores envolvidos nas contas pagas do portal da transparência do município de Londrina no período de 01/01/2022 a 09/09/2022. O desenvolvimento metodológico do estudo pressupõe o entendimento da análise de conteúdo, a qual segundo Bardin (2000), tem funções heurística e de administração de prova, auxiliando na identificação das principais ideias e conceitos presentes nos dados. Em contrapartida, Krippendorff (2004) destaca que, embora valiosa, a análise de conteúdo tradicional possui limitações, como a dificuldade de lidar com grandes volumes de dados. Por isso, a análise computacional de conteúdo surge como uma alternativa, transformando um conjunto de texto em representações que aproximam a resposta à pergunta do pesquisador.

Assim, a partir do ciclo analítico da mineração de dados (OLIVEIRA; GUERRA; MCDONNEL, 2018) as técnicas de análise computacional de conteúdo aplicadas aos dados governamentais abertos podem contribuir em diversas áreas da arquivologia, como a gestão documental, na preservação de documentos e na análise de documentos governamentais.

Uma das etapas da técnica de análise computacional é a mineração de dados públicos que, entendida como campo interdisciplinar, reúne técnicas de *machine learning*, reconhecimento de padrões, estatísticas, banco de dados e visualização (CAMILO; SILVA, 2009) tornando-se uma ferramenta útil para pesquisadores, governos, empresas e outros interessados em inferir insights com base em dados públicos disponíveis.

Para desenvolver o estudo, os seguintes sistemas de código aberto foram utilizados: *Python*[1], *LibreOffice Base* e *Calculator*[2], *Open Refine*[3] e o *Gephi*[4].

1 Disponível em: https://www.python.org/

2 Disponível em: https://pt-br.libreoffice.org/

3 Disponível em: https://openrefine.org/

4 Disponível em: https://gephi.org/

2.1 Coleta de dados

Os dados públicos foram coletados do Portal da Transparência do Município de Londrina/PR, focando na seção de contabilidade e finanças, a analisamos as despesas pagas disponibilizadas pela Prefeitura de Londrina entre 01/01/2022 e 09/09/2022, pois representam registros concretos dos gastos públicos e são relevantes para entender a destinação do dinheiro. Embora os dados do portal sejam oficiais e considerados íntegros de acordo os princípios estabelecidos pelo *Open Government Data Principles*, as contas pagas estavam disponíveis em PDF, o que levou à conversão para TXT. A análise buscou mapear a relação de latência entre credores e secretarias, priorizando o interesse da população em compreender a destinação dos recursos públicos.

2.2 Tratamento

Procuramos identificar as limitações de *hardware* e *software* para definir o montante total de documentos restringido o estudo para adequar-se à capacidade computacional disponível. Os documentos disponibilizados em PDF apresentavam múltiplos itens documentais e foram divididos em notas de pagamento individuais. Os arquivos TXT foram carregados no *Open Refine* para limpeza e tratamento usando expressões GREL e linguagem *Regex*. As entidades e atributos foram identificados e estruturados por meio de análise exploratória.

Devido aos problemas de processamento os dados foram filtrados para permitir análise e mapeamento da relação de latência entre os atores envolvidos. Os dados estruturados foram levados ao *LibreOffice Base*, construindo relações de latência entre as entidades e calculando arestas e nós para compor um grafo.

A visualização dos dados foi realizada utilizando o *software Gephi*, aplicando algoritmos de modularidade e visualização geral da rede para calcular grau médio e peso de cada nó. Métricas obtidas foram levadas ao *LibreOffice Base* para identificar termos com maior grau e definir grupos a partir das relações dos termos candidatos.

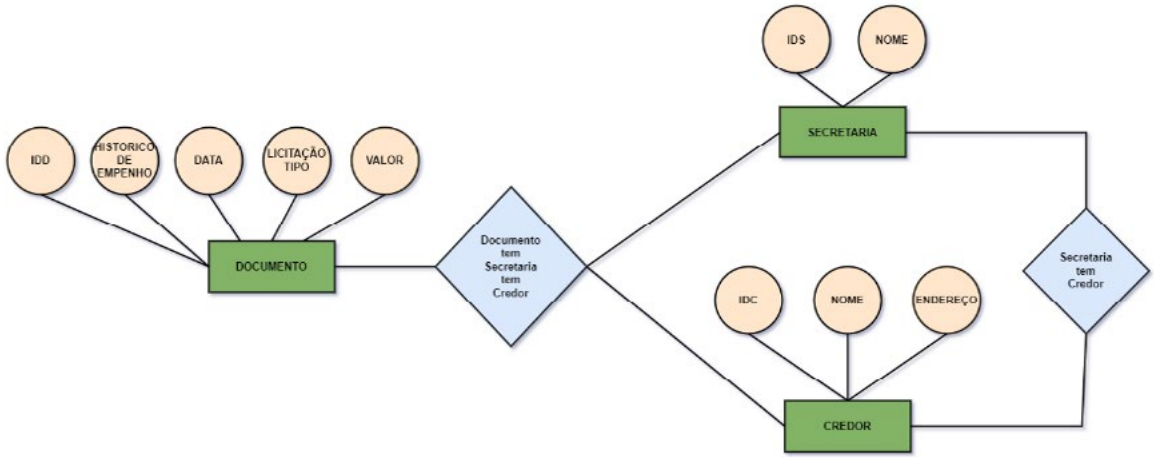
Os resultados foram comunicados através de grafos, mapeando a relação de latência dos credores e secretarias dentro do recorte efetuado para viabilizar o processamento.

3. Resultados

Após o tratamento dos dados, para compreensão dos atores e da tríade entidade, atributo e valor empregou-se o modelo entidade-relacionamento, a partir do qual identificou-se entidades como Fornecedor e Secretaria e atributos como data, tipo de licitação, endereço, CPF/CNPJ, telefone, histórico de empenho e classificação de despesa foram identificados. Um modelo de entidade-relacionamento (*MER*), foi criado, composto por entidades, relacionamentos e atributos (CHEN, 1976).

Assim, o modelo simplificado elaborado será apresentado abaixo:

Figura 1 – Modelo simplificado de Entidade-Relacionamento

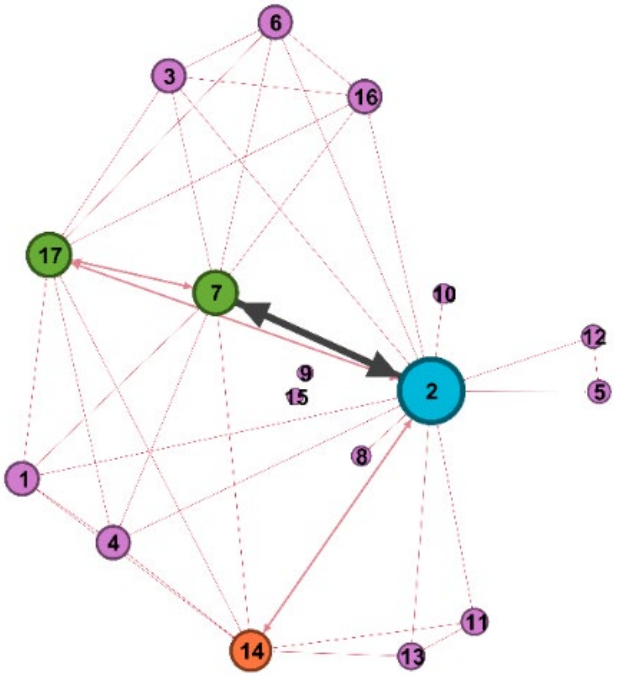


Fonte: Dados da pesquisa (2022)

A partir das entidades observadas nos documentos, identificou-se como atores as secretarias e os credores/fornecedores envolvidos nas notas de pagamento. Assim, procurou-se mapear a relação de latência existente entre os credores e as secretarias, isto é, uma vez que cada nota de pagamento apresentava um único credor/fornecedor e uma única secretaria, procurou-se identificar como os credores/fornecedores se relacionavam através das secretarias para mapear esses relacionamentos no período delimitado e oferecer uma visualização gráfica do relacionamento entre os atores do fundo documental em estudo.

Assim, a partir da aplicação de algoritmos de centralidade, distribuição e agrupamento, sintetizamos o grafo da Figura 2.

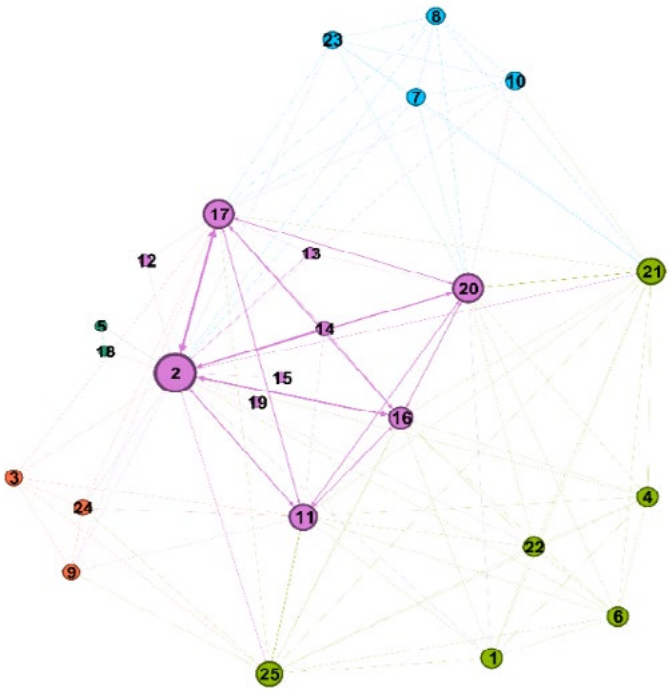
Figura 2 – Mapeamento da relação de latência entre credores e secretarias a partir da seleção de documentos contendo o termo "*software*" no histórico de empenho



Fonte: Dados da pesquisa (2022)

Partindo dos credores identificados dentro dos dados extraídos dos PDFs relativos à prestação de serviços de *software* contratados pela prefeitura, o grafo da Figura 2 objetivou identificar padrões e tendências de gastos relacionados a essa área.

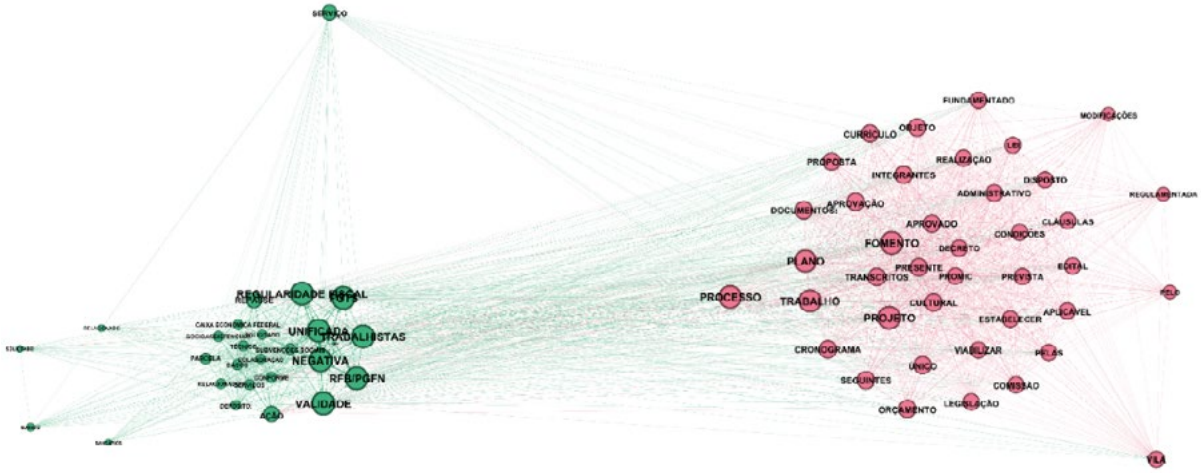
Figura 3 – Mapeamento da relação de latência entre credores e secretarias a partir da seleção de documentos contendo o termo "*software*" e "Impressora" no histórico de empenho



Fonte: Dados da pesquisa (2022)

O grafo da Figura 3 demonstra possíveis sinergias entre as áreas de *software* e impressão, possibilitando a otimização de recursos e a redução de custos.

Figura 4 – Mapeamento da relação de latência semântica entre termos candidatos a partir da seleção de documentos contendo o termo "Subvenções Sociais" no histórico de empenho



Fonte: Dados da pesquisa (2022)

A partir desse contexto documental, o grafo da Figura 4 mostra os termos que estão mais relacionados entre si para definir quais são os mais relevantes para a análise em questão e quais aqueles que não tem relevância de agrupamento, podendo ser, portanto, ignorados.

Assim, dentro do histórico de empenho envolvendo as subvenções sociais, o emprego dos algoritmos de consolidação, agrupamento e centralidade resultaram em dois grupos distintos, que denominamos de Grupo 1 (G1) e Grupo 2 (G2). Esse dado sugere a existência de fatores específicos influindo na distribuição de recursos para cada um dos grupos o que pode indicar a necessidade de uma revisão mais cuidadosa dos processos de tomada de decisão envolvidos nesse tipo de política pública.

Portanto, considerando a teoria ator-rede aplicada aos dados governamentais abertos como um resultado da interação e do engajamento de múltiplos atores e a interação entre gestores públicos, cidadãos, instituições governamentais, sistemas de informação e tecnologias utilizadas para divulgar os dados enquanto pressuposto da transparência não sendo apenas uma responsabilidade dos gestores públicos, mas um processo que envolve uma ampla rede de atores e elementos.

Para avançar em direção a uma gestão pública mais transparente e responsável por meio dos postais de transparência é necessário superar os desafios dos silos de informação, adotar uma postura ética e responsável na divulgação dos dados, promover a integração dos atores envolvidos na rede de transparência e fortalecer a disponibilização de dados governamentais abertos. Assim será possível promover uma gestão pública mais eficaz, participativa e comprometida com o interesse público.

4. Considerações Finais

Considerando o exposto, a utilização das técnicas de análise computacional mencionadas se configura como uma proposta para lidar com o aumento constante do volume de dados e informações gerados e armazenados por organizações públicas, privadas e do terceiro setor, especialmente no campo da arquivologia. Neste sentido, cabe a arquivologia desenvolver metodologias e ferramentas para análise e interpretação de dados arquivísticos por meio da mineração de dados públicos, garantindo sua preservação e acessibilidade a longo prazo.

A aplicação de algoritmos e técnicas de visualização possibilitariam, assim, identificar padrões e tendências nos dados públicos, facilitando a gestão eficiente de documentos e informações relacionadas, especialmente em áreas como gestão de políticas públicas. Desse ponto de vista, a transparência governamental e a melhoria da análise quantitativa das relações propostas estão condicionadas à análise computacional de silos de informação.

Estendendo a análise, observamos que os estudos sobre mineração de dados públicos podem beneficiar o município de Londrina a identificar problemas e promover soluções eficientes aos problemas identificados. Desta forma, essas iniciativas podem melhorar a tomada de decisões, oferecer informações mais específicas e estruturadas e promover a transparência e, ainda, fomentar a participação cidadã. Visualizamos, também, que o agrupamento de dados governamentais abertos pode desencadear uma economia de recursos, colaborando com a população, gestores públicos e melhoria da eficiência dos serviços prestados, contribuindo para o bem-estar social.

Por outro lado, o estudo demonstra que a técnica de análise computacional de conteúdo, pode auxiliar na identificação de palavras-chave e temas recorrentes, facilitando, assim, a categorização e classificação documental. Neste sentido, visualizamos que o conceito de mineração de dados públicos pode vir a auxiliar o pesquisador na identificação de problemas e no apoio a decisões mais informadas, impactando no entendimento da transparência governamental e, portanto, culminando na eficiência dos serviços prestados à população.

Dito de outra forma é crucial que gestores públicos e pesquisadores reconheçam a importância desse campo de estudo e invistam em estudos e práticas relacionadas à mineração de dados e análise computacional de conteúdo. Somente dessa forma a produção de conhecimento contribuirá para a melhoria da transparência governamental, eficiência dos serviços prestados e bem-estar social. Destarte, essas tecnologias e metodologias podem transformar a coleta, análise e uso de informações para fins públicos.

Em última análise, a arquivologia tem papel de suma importância no desenvolvimento de políticas públicas, práticas de transparência e acesso à informação, promovendo a participação cidadã e a democracia.

Referências

BARDIN, Laurence. **Análise de conteúdo.** [Lisboa]: Edições 70, 2000.

CAMILO, Cássio Oliveira; SILVA, João Carlos da. **Mineração de dados**: conceitos, tarefas, métodos e ferramentas. Goiânia: Universidade Federal de Goiás, 2009. Disponível em: https://rozero.webcindario.com/disciplinas/fbmg/dm/RT-INF_001-09.pdf. Acesso em: 25 abr. 2023.

CHEN, Peter Pin-Shan. The entity-relationship model: toward a unified view of data. **ACM Transactions on Database Systems,** New York, v. 1, n. 1, p. 9-36, mar. 1976. Disponível em: https://dl.acm.org/doi/10.1145/320434.320440. Acesso em: 25 abr. 2023.

JACINTO, Afonso Serafim *et al.* Aumentando a eficiência da fiscalização social sobre gastos públicos por meio de uma aplicação web baseada em dados abertos. *In*: SENHORAS, Elói Martins (org.). **Ciência, tecnologia e inovação:** geração de emprego e democratização de oportunidades. Ponta Grossa: Atena, 2022. p. 39 - 44. Disponível em: https://www.atenaeditora.com.br/catalogo/ebook/ciencia-tecnologia-e-inovacao-geracao-de-emprego-e-democratizacao-de-oportunidades/. Acesso em: 9 dez. 2022.

KRIPPENDORFF, Klaus. **Content analysis:** an introduction to its methodology. 2nd. ed. Thousand Oaks: Sage, 2004.

LATOUR, Bruno. **Reagregando o social**: uma introdução a teoria do ator-rede. Salvador: EDUFBA; Bauru: EDUSC, 2012.

LONDRINA. Prefeitura de Londrina. **Portal da transparência**: sobre. Londrina: Prefeitura do Município de Londrina, 2022. Disponível em: https://portal.londrina.pr.gov.br/index.php/sobre-transparencia/. Acesso em: 20 abr. 2023.

OLIVEIRA, Paulo Felipe de; GUERRA, Saulo; MCDONNEL, Robert. Ciência de Dados com R: introdução. Brasília: Editora IBPAD, 2018.

OPEN GOVERNMENT DATA PRINCIPLES. 2007. Disponível em: https://public.resource.org/8_principles.html/ Acesso em: 20 abr. 2023.