RICI Revista Ibero-Americana de Ciência da Informação

ISSN 1983-5213

Visualização de informação sobre preços de medicamentos da base de dados abertos da ANVISA com auxílio de análise de redes de informação

Lucas Vale
Universidade Federal do Espírito Santo, Programa de Pós Graduação em Ciência da Informação
Vitória, ES, Brasil
lucas.s.vale@edu.ufes.br

Henrique Monteiro Cristovão
Universidade Federal do Espírito Santo, Programa de Pós Graduação em Ciência da Informação, Vitória, ES, Brasil
henrique.cristovao@ufes.br

DOI: https://doi.org/10.26512/rici.v16.n1.2023.47582

Recebido/Recibido/Received: 2022-12-10
Aceitado/Aceptado/Accepted: 2023-03-14

ARTIGOS

Resumo
Investigou e revelou relações entre variáveis da base de dados abertos da Anvisa sobre preço de medicamentos. Empregou a metodologia qualitativa e aplicada, com 26310 registros correspondentes ao período de 2017 a 2021. Com o método da descoberta de conhecimento apoiado por técnicas de análise de redes complexas e com o suporte de softwares apropriados, observou-se que os medicamentos mais produzidos pelos laboratórios registrados são os de tarja vermelha, enquanto os medicamentos de tarja preta têm produção e demanda mais limitadas devido ao uso restrito. Observou-se também uma alta produção de classes terapêuticas em que o custo gira em torno de R$100, sugerindo que a maioria dos laboratórios tem como público-alvo as classes C e D. Além disso, apenas um laboratório produz todos os medicamentos com custo de produção acima de 1 milhão de reais. Contudo, é necessário mais esforços na análise da base de dados para identificar outras relações entre as variáveis. Observou-se também que elementos de ambas as áreas, Ciência de Dados e Ciência da Informação foram importantes para o desenvolvimento da pesquisa, como o aspecto interdisciplinar, as técnicas de ARS e a vocação do empenho de esforços direcionados à resolução de problemas reais da sociedade.
Palavras-chave: Visualização de informação. Análise de redes de informação. Descoberta de conhecimento. Ciência de dados. Ciência da Informação. Anvisa.

Visualization of information on drug prices from ANVISA'S open database with the aid of information network analysis

Abstract
It investigated and revealed relationships between variables from Anvisa's open database on drug prices. It used a qualitative and applied methodology, with 26,310 records corresponding to the period from 2017 to 2021. With the knowledge discovery method supported by complex network analysis techniques and with the support of appropriate software, it was observed that the drugs most produced by Registered laboratories are those with the red stripe, while the black stripe drugs have more limited

production and demand due to restricted use. There was also a high production of therapeutic classes in which the cost is around R$100, suggesting that most laboratories target classes C and D. In addition, only one laboratory produces all drugs with a cost of production above 1 million reais. However, more efforts are needed in the analysis of the database to identify other relationships between the variables. It was also observed that elements from both areas, Data Science and Information Science were important for the development of the research, such as the interdisciplinary aspect, the ARS techniques and the vocation of commitment to efforts aimed at solving real problems in society.
Keywords: Information visualization. Analysis of information networks. Knowledge discovery. Data science. Information Science. Anvisa.

Visualización de información sobre precios de medicamentos de la base de datos abierta de ANVISA con la ayuda del análisis de redes de información

Resumen

Investigó y reveló relaciones entre variables de la base de datos abierta de Anvisa sobre precios de medicamentos. Utilizó una metodología cualitativa y aplicada, con 26.310 registros correspondientes al período de 2017 a 2021. Con el método de descubrimiento de conocimiento apoyado en técnicas de análisis de redes complejas y con el apoyo de software apropiado, se observó que los medicamentos más producidos por laboratorios Registrados son las que tienen la franja roja, mientras que las drogas de la franja negra tienen una producción y demanda más limitada debido al uso restringido. También hubo una alta producción de clases terapéuticas en las que el costo es de alrededor de R$ 100, lo que sugiere que la mayoría de los laboratorios apuntan a las clases C y D. Además, solo un laboratorio produce todos los medicamentos con un costo de producción superior a 1 millón de reales. Sin embargo, se necesitan más esfuerzos en el análisis de la base de datos para identificar otras relaciones entre las variables. También se observó que elementos de ambas áreas, Data Science y Information Science, fueron importantes para el desarrollo de la investigación, como el aspecto interdisciplinario, las técnicas ARS y la vocación de compromiso con los esfuerzos encaminados a la solución de problemas reales de la sociedad.
Palabras clave: Visualización de información. Análisis de redes de información. Descubrimiento del conocimiento. Ciencia de los datos. Ciencias de la Información. Anvisa.

1 Introdução

A visualização da informação é um dos principais componentes para uma recuperação de informação significativa, pois tende a aumentar a percepção do usuário no entendimento quanto às informações recuperadas. A Recuperação de Informação é um dos campos mais importantes da Ciência da Informação (CAPURRO; HJØRLAND, 2003) que, por sua vez, tem uma forte dimensão social e humana, está ligada à tecnologia da informação, sendo também uma ciência participante e ativa na evolução da sociedade da informação (SARACEVIC, 1995). De outro modo, ou de forma complementar, a Ciência de Dados tem como principal mote a transformação de dados em valor real (AALST, 2016), ao qual pode ser fornecido na forma de previsões, decisões automatizadas, modelos baseados em dados ou qualquer tipo de visualização de dados que ofereçam informações úteis. Ambas as áreas de pesquisa, Ciência da Informação e Ciência de Dados, são fortemente interdisciplinares, segundo os autores Saracevic e Aalst, respectivamente.

Apesar da forte aproximação das duas ciências em nível de aplicação na resolução de problemas sociais, conforme exemplificado na pesquisa realizada por Virkus e Garoufallou (2019), segundo levantamento bibliográfico de Martins (2022), no Brasil "[...] a Ciência de Dados

no contexto da Ciência da Informação ainda parece estar em sua fase inicial de apropriação pelos pesquisadores da área". Ainda assim, a união das duas áreas têm sido complementares e têm proporcionado trabalhos interessantes como, por exemplo, na área da descoberta do conhecimento e mais especificamente na temática da Análise de Redes Sociais (ARS).

Na área da ARS a Ciência da Informação pode cuidar da parte da organização do conhecimento, de informações e de dados, entregando um conteúdo mais bem tratado e semanticamente bem identificado para o processamento da parte específica da análise de dados, tratada por técnicas da Ciência da Dados. A ARS tem cunho fortemente interdisciplinar e "[...] tem sido aplicada em diversos campos da ciência, com múltiplas finalidades, auxiliando no estudo de diferentes fenômenos sociais, em especial em análise da difusão de inovações, jornalismo investigativo, mapeamento de redes terroristas, mapeamento de epidemias, mobilidade demográfica e, particularmente, no campo administrativo, em estudos de processos decisórios e gestão do conhecimento em redes interorganizacionais" (SOUZA; QUANDT, 2008).

As redes de informação, consideradas como redes complexas, utilizam-se em grande parte das ferramentas e técnicas da ARS que, segundo Wasserman e Faust (1994), tem foco no entendimento das ligações entre entidades sociais e as implicações destas ligações. As redes de informação são mais amplas do que as redes sociais, pois podem lidar com atores sociais, não sociais e também com dados oriundos de bases de dados.

Nesse contexto, de um mundo orientado a dados, retratado pelo fenômeno denominado de *data driven,* organizações caminham em direção ao aproveitamento das potencialidades advindas do uso de dados coletados, estruturados e devidamente tratados. É concomitante ao desenvolvimento da civilização o acúmulo de dados onde, ter acesso à informação desejada de maneira precisa é um diferencial que pode fazer prosperar negócios em áreas muito distintas. Isso faz dos profissionais responsáveis pelo processo de tomada de decisão dependentes de ferramentas assertivas para os procedimentos de coleta, organização, processamento e utilização da informação (COSTA *et al.*, 2009).

Temas relacionados à saúde e ao bem-estar são sempre sensíveis e relevantes para a sociedade. Em 2008, o Instituto Brasileiro de Geografia e Estatística (IBGE) estimou que a população de idosos no Brasil equivaleria a 9,5% da população total, com projeção deste grupo corresponder a aproximadamente 30% da população até 2050. Segundo a OMS, o envelhecimento populacional traz algumas implicações para as sociedades, como o aumento natural do número de indivíduos que vivem com doenças crônicas, em especial aquelas com tratamento contínuo, como hipertensão arterial sistêmica e diabetes *mellitus* (BALDONI & PEREIRA, 2010). Essa transformação acelerada acaba por dificultar a atuação do Estado, visto

que alterações nas proporções das faixas etárias geram perfis epidemiológicos distintos que por sua vez demandam políticas e investimentos específicos (CHAIMOWICZ, 1997).

Como elucidado por Macedo *et al*. (2006), metade da população brasileira consumiu ou está consumindo medicamentos na última semana, sendo a idade, o sexo e a presença de doenças crônicas os principais fatores relacionados ao uso. Pessoas com idade superior a 65 anos tendem a consumir 5 medicamentos a mais do que adultos com idade inferior a esta, condição que se relaciona também com a alta ocorrência de reações adversas em função do uso de medicamentos neste grupo (BLASCO PATIÑO et al., 2008). A busca por serviços de saúde pelas mulheres é maior, o que explica também o maior consumo de medicamentos (KAUFMAN *et al*., 2002).

Frente ao exposto, torna-se elementar o entendimento de que o tratamento farmacológico é fator determinante na melhoria da qualidade de vida relacionada à saúde e na devida promoção do estado de bem-estar de todos os sexos e faixas etárias (MOHAMMED; MOLES; CHEN, 2016) e, a partir disto, entende-se que no Brasil e no mundo há uma alta demanda pela racionalização do uso e distribuição de medicamentos, o que vai ao encontro com o desenvolvimento de sistemas informatizados de informação sobre o consumo, grupos de risco etc. (GOMES; SILVA; GALVÃO, 2017). Neste contexto, entende-se que é salutar a preocupação com um Sistema Nacional de Saúde eficiente e robusto (BÁRRIOS *et al*., 2020) e, no Brasil, o Sistema Único de Saúde (SUS) tem integrado a si a autarquia Agência Nacional de Vigilância Sanitária (ANVISA), responsável por, entre outras coisas, a gestão do controle de qualidade e da vigência de preços de medicamentos.

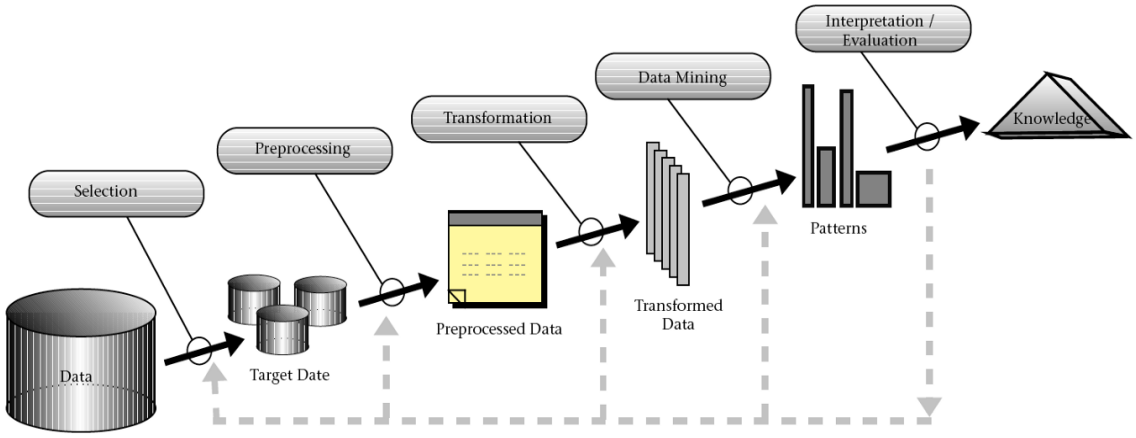
Dessa forma, o objetivo da presente pesquisa é investigar e revelar relações entre variáveis da base de dados abertos da ANVISA sobre preços de medicamentos. Utilizando-se de metodologia apropriada para descoberta de conhecimento e ferramentas computacionais adequadas é possível revelar relacionamentos não evidentes entre variáveis da base de dados escolhida. Uma vez revelados padrões ou tendências não evidentes ou despercebidas, é também possível entender melhor a problemática como também estabelecer encaminhamentos de ações para melhoria de políticas sociais que possam trazer algum ganho para os indivíduos diretamente envolvidos com o preço de medicamentos no Brasil.

O artigo está organizado em cinco seções. A seção 2 trata do tema da descoberta do conhecimento, abordando a análise de redes de informação, a visualização de dados e de informação. A seção 3 apresenta os procedimentos metodológicos e suas etapas: pré-processamento, transformação, mineração de dados, interpretação e avaliação. A seção 4 apresenta e discute os resultados. Finalmente, a seção 5 trata das considerações finais.

2 Descoberta de Conhecimento

O processo de descoberta de conhecimentos (KDD - *Knowledge Discovery in Database*) é um ponto de partida metodológico para a identificação de padrões em um determinado conjunto de dados em que exista potencial de compreensão e utilidade para o usuário. O processo sugerido por Fayyad *et al*. (1996), Figura 1, estabelece cinco etapas, partindo de uma base de dados (*Data*) até a formação, ou descoberta, de conhecimento (*Knowledge*).

Figura 1 - Etapas do processo KDD



Fonte: Fayyad *et al*. (1996).

As etapas do processo de descoberta de conhecimento, segundo Fayyad et al. (1996), são:

(1) Seleção (*Selection*) dos dados e identificação de possíveis problemas e/ou objetivo da análise. Nesta etapa pode-se também aprender mais sobre o significado dos dados para proceder na correta seleção dos dados que poderão contribuir de forma mais efetiva com o processo da descoberta do conhecimento.

(2) Pré-processamento (*Preprocessing*) dos dados, preparação, limpeza e ajustes diversos com o propósito de solucionar problemas, corrigir erros entre outros. De modo geral, a fase de pré-processamento é tida como semiautomática, pois é dependente da aptidão do operador em identificar problemas no conjunto de dados e a sua natureza, como expõe Batista (2003).

(3) Transformação (*Transformation*) dos dados possibilita a alteração da estrutura dos dados ou da base de dados, podendo ocorrer alterações significativas como a criação de novos subconjuntos de dados para atender a requisitos específicos de determinados softwares de mineração.

(4) Mineração de dados (*Data Mining*) é basicamente a aplicação de algoritmos escolhidos conforme o contexto, problema e dados para extrair padrões e relacionamentos. Essa etapa tem

grande diversidade de métodos disponíveis para uso. Na presente pesquisa há ênfase na técnica de análise de redes de informação que, segundo Porto e Ziviani (2014), é uma técnica de mineração de dados e é também um dos desafios de pesquisa em Ciência de Dados. Ela será abordada na subseção 3.1.

(5) interpretação e avaliação (*Interpretation / Evaluation*) dos dados resultantes obtidos na etapa da mineração de dados. Técnicas de formatação de dados na análise de redes acompanhada de inspeção visual são muito úteis nessa etapa. Além disso, o uso de softwares de visualização de dados e de informação contribuem de forma significativa para essa etapa que tem como principal objetivo a revelação de informações que permitam e impulsionam a criação de conhecimento.

3.1 Análise de Redes de Informação

A Ciência das Redes é uma área interdisciplinar que se compromete a compreender a estrutura topológica, dinâmica e funcionamento das redes complexas em diferentes áreas. Essa disciplina abrange teorias e métodos de diversos campos, como ciência da informação, sociologia, ciências humanas, física, matemática, estatística e ciência da computação. A Ciência das Redes traz como principal contribuição a percepção de que muitas redes emergem e se desenvolvem a partir de um conjunto comum de leis e mecanismos (BARABÁSI, 2013). A rede mundial de computadores, a Internet, é um exemplo importante dessa compreensão, onde alguns nós altamente conectados são responsáveis pela grande maioria das conexões da rede. É salutar entender que esta mesma estrutura é também observada em muitas outras redes complexas, tais como a rede de interações entre proteínas em sistemas biológicos e a rede de conexões entre células do sistema nervoso.

Uma rede complexa, ainda segundo Barabási (2003), é um conjunto de nós interligados por arestas que constitui uma estrutura topográfica sofisticada. O estudo das redes neste formato teve início com a solução do problema das pontes de Königsberg por Euler, em 1735, que derivou a teoria dos grafos (METZ, 2007). A análise de redes complexas, portanto, se destina a compreender a relação entre nós e as suas consequências, pouco se preocupando com as características particulares dos nós, e sim com a sua totalidade estrutural, tomando emprestados fundamentos da ARS, como sugerem Wasserman et al. (1994). Nesta etapa, os dados dispostos em nós e arestas fornecem possibilidades de inspeção visual, desde que adequadamente formatados sobre a topologia e as relações existentes entre os nós.

Uma rede bipartida (ou rede de afiliação ou rede de dois modos) é uma rede com dois ou mais conjuntos de nós onde as conexões só podem ocorrer com nós de outros conjuntos. Por exemplo, uma rede com um conjunto de nós representando as doenças e outro conjunto

representando medicamentos. Considerando essa rede como bipartida, um medicamento somente se relaciona com doenças e não com outro medicamento. De forma análoga, redes n-partidas, como a rede tripartida composta de três partes, tem o mesmo comportamento quanto às suas partes. Uma projeção bipartida é a geração de uma rede monopartida pela eliminação de um dos conjuntos de nós. São criadas ligações entre os nós do conjunto que permanece, a partir das conexões intermediárias com os nós eliminados.

Redes bipartidas podem levar à descoberta de conexões subjacentes (BORGATTI, HALGIN; 2014) desde que, para isso, sejam aplicadas projeções bipartidas que, apesar de eliminarem uma parte considerável da rede, são necessárias para viabilizar a análise por meio da inspeção visual, uma vez que as redes ficam menores e podem evidenciar relacionamentos escondidos. Em alguns casos ela é necessária devido às limitações de poder computacional que impedem a realização de determinadas operações em redes maiores. De uma maneira geral, as projeções bipartidas auxiliam a compreensão da dinâmica de interação entre grupos da rede.

Kadushin (2004) diz que as teorias que fundamentam a ARS é uma das poucas, senão a única teoria das Ciências Sociais, que não é reducionista, isto é, que necessita que o objeto seja analisado como um todo e não em suas partes. Nessa mesma linha, Newman (2010) recomenda que uma rede deve preferivelmente ser trabalhada em sua totalidade para que não haja perda de informações sobre os relacionamentos entre as entidades.

Por outro lado, a projeção bipartida, que de certa forma é uma redução da rede para uma topologia monopartida, em alguns casos pode evidenciar mais facilmente relações entre determinadas variáveis. Everett e Borgatti (2012) analisaram criticamente a suposição de que projeções bipartidas poderiam levar a perdas de informações estruturais da rede. Os autores concluíram que análises realizadas sobre as redes resultantes de projeções bipartidas não provocam perdas de informação e ainda podem ter vantagens conceituais sobre análises realizadas diretamente na rede bipartida original. Nessa mesma linha, Melamed (2014) também fez estudos comparativos e concluiu que os resultados de análises realizadas a partir de projeções bipartidas são, em muitos casos, melhores do que aqueles realizados diretamente na rede bipartida original.

Essas técnicas de análise e mineração de dados, baseadas em ARS e, em especial, por meio das projeções bipartidas, têm sido amplamente utilizadas em diversos trabalhos de pesquisa e ainda oferecem caminhos com resultados diferenciados. Por exemplo, Zhou *et al*. (2014) construíram uma rede de informação a partir de dados coletados sobre doenças acometidas por pacientes e seus respectivos sintomas e, após a aplicação de projeções bipartidas, obtiveram uma rede monopartida de doenças onde, por inspeção visual, foi possível identificar associações entre as doenças, chegando-se a conclusões próximas da determinação de comorbidades. Outra

aplicação, desenvolvida por Gao et al. (2017), relata resultados denominados de "previsão de links", onde conseguiu-se prever links futuros da estrutura topológica da rede como, por exemplo, na descoberta de grupos clandestinos de terroristas ou criminosos.

3.2 Visualização de Dados e de Informação

As pessoas têm normalmente maior compreensão de conteúdos veiculados por meio de informação visual (imagens) do que informação puramente textual. A visualização é também considerada um processo de transformação de dados, informação e conhecimento em representações gráficas para apoiar tarefas, tais como análise de dados, exploração de informação, explicação da informação, previsão de tendências, detecção de padrões, descobertas etc. (ZHANG, 2008).

A visualização de dados, ou *data visualization*, tem como objetivo a representação de dados abstratos em formatos de gráficos, tabelas, desenhos que facilitam o leitor a obter significado desses dados. Nesse mesmo contexto, a visualização de informação, ou *information visualization*, tem como objetivos a revelação de padrões invisíveis a partir de dados abstratos e trazer novas percepções para as pessoas, e não apenas imagens bonitas (CHEN, 2013). Com o surgimento das aplicações auxiliadas por computador, a visualização da informação tornou-se uma ferramenta essencial para a exploração de grandes quantidades de dados em diversas áreas, proporcionando vantagens como a obtenção de novos insights e a melhoria do acesso, recuperação e exploração de grandes bancos de dados tendo, portanto, uma forte influência em diversas áreas, como a ciência da informação, mineração e análise de dados, além de fornecer novas pistas para a exploração e recuperação de informação, interação humano-computador e design de interface. É importante ressaltar que a visualização da informação é um campo que tem crescido significativamente com as inovações tecnológicas, contribuindo cada vez mais para a solução de problemas complexos e para a tomada de decisões em diversas áreas de atuação (BURKHARD, 2005).

A inspeção visual é uma das tarefas mais importantes na análise de redes de informação, tendo um suporte forte da área de visualização de informação. Ela permite observar, tornar padrões despercebidos evidentes (CHEN, 2013). O processo de visualização pode, além do mais, ser constituinte da fase de pré processamento ou de transformação, uma vez que os limites entre as fases delimitadas pelo KDD não são bem estabelecidos, segundo HAND *et al*. (2001), sendo de toda forma um artifício da mineração sobretudo da interpretação dos dados.

Na inspeção visual, a visualização de informação é a principal ação que direciona o cientista de dados em novas buscas. O termo "*what you see drives your search*" (o que você vê direciona sua pesquisa) representa bem essa ideia investigativa para extração de informações (PENNA;

MAGAZZENI; OREFICE, 2012). A exploração da estrutura de uma rede por cálculo é muito mais concisa e precisa do que uma inspeção visual. No entanto, em alguns casos, a exploração por cálculo pode ser abstrata e de difícil interpretação (NOOY; MRVAR; BATAGELJ, 2018). A visualização é uma ferramenta útil na análise de dados de rede, pois permite ver instantaneamente características estruturais importantes, considerando que o [...] olho humano tem um talento enorme para identificar padrões, e as visualizações nos permitem colocar esse dom para trabalhar em nossos problemas de rede" (NEWMAN, 2010, tradução nossa).

Outra ferramenta que contribui com a visualização da informação é o*dashboard* (painel), desenvolvido por ferramentas de Business Intelligence (BI), como o Power BI Desktop[1], é bastante utilizado na área de ciência de dados uma vez que oferece recursos avançados de visualização de dados, permitindo a criação de gráficos, tabelas e outras representações visuais para análise de grandes quantidades de dados. Contudo, para que as informações representadas nesses *dashboards* sejam acessíveis e compreensíveis a um público mais amplo, pode ser necessário se valer de técnicas como *data storytelling.*

O *data storytelling* é uma estrutura de comunicação que se esforça para contar uma história a partir dos dados coletados e analisados. Por meio da criação de uma narrativa, é possível explicar o significado dos dados e seu impacto no contexto em que estão inseridos. Isso pode ser feito por meio de gráficos, infográficos, mapas e outras visualizações que ajudem a transmitir as informações de forma clara e objetiva.

A união destas duas ferramentas - *dashboards* e *data storytelling* - pode aperfeiçoar a maneira como o conhecimento descoberto será comunicado, facilitando o acesso e a compreensão dos dados por um público não especializado. Essa combinação também pode ser uma ferramenta valiosa para persuadir e inspirar ações com base nos insights fornecidos pelos dados. Dessa forma, a utilização dessas ferramentas pode impulsionar a tomada de decisões mais informadas e efetivas, ajudando empresas e organizações a alcançarem seus objetivos.

3 Procedimentos Metodológicos

A pesquisa tem abordagem qualitativa e natureza aplicada, e usa técnicas da ARS enquanto procedimentos metodológicos para a descoberta de conhecimento. Na área de Ciências Sociais Aplicadas, entre outras como a Saúde Coletiva, a ARS é uma ferramenta útil e complementar aos processos de análise de dados. Ela é considerada uma ferramenta metodológica que permite enxergar o que outras abordagens não permitem (WASSERMAN; FAUST, 1994; HIGGINS, RIBEIRO; 2018) tendo "[...] origem multidisciplinar (psicologia, sociologia, antropologia,

[1] O Microsoft Power BI Desktop é um conjunto de serviços de software de uso gratuito destinados à manipulação de dados com o propósito de gerar, por exemplo, painéis interativos. Disponível em https://powerbi.microsoft.com/.

matemática, estatística) cuja principal vantagem é a possibilidade de formalização gráfica e quantitativa de conceitos abstraídos a partir de propriedades e processos característicos da realidade social" (SOUZA; QUANDT, 2008) além de ser considerada uma estratégia para investigar estruturas sociais e sendo fortemente vinculada à área da Ciência da Informação (OTTE, ROUSSEAU, 2002).

A base de dados analisada é mantida e alimentada pela ANVISA, disponível no portal de dados abertos do governo[2] de maneira estruturada e é composta por 26310 registros correspondentes ao período de 2017 a 2021, e com tamanho aproximado de 10Mb.

Após a obtenção dos dados, as etapas do KDD foram executadas de forma compatível com as etapas mostradas na Figura 1, e são descritas nas próximas subseções

3.1 Pré-processamento

Nesta etapa foram removidos ruídos da base de dados e selecionados os atributos mais relevantes para a análise, bem como realizada a formatação adequada dos dados.

Fazendo uso da ferramenta OpenRefine[3], foram excluídas as colunas CNPJ, Registro, EAN, PF 0%, PF 12%, PF 17,5%, PF 17,5% ALC, PF 18%, PF 18% ALC, PF 20%, PMC 12%, PMC 17,5%, PMC 17,5% ALC, PMC 18%, PMC 18% ALC, PMC 20%, Código GGREM, por serem irrelevantes para o estudo, do ponto de vista dos autores e objetivos da pesquisa. O nome das colunas foi alterado para o padrão camelCase[4] para melhor integração entre as ferramentas de análise utilizadas. As colunas contendo valores numéricos foram convertidas para a classe de números (variável quantitativa contínua) e os valores coluna de *tarja,* que continham irregularidades derivadas da inserção dos registros foram padronizados em 4 categorias: *Tarja Preta, Tarja Vermelha, Venda Livre e Sem Tarja*.

3.2 Transformação

Neste ponto, alguns dados são transcritos para formatos mais adequados para a análise, considerando-se peculiaridades do método de análise de redes de informação. Utilizando a linguagem GREL[5], os valores numéricos de preço (quantitativos contínuos) foram transformados em categorias de preço (qualitativos ordinais) seguindo a seguinte organização para as categorias:

---

[2] Disponível em: *https://dados.gov.br/dataset/preco-de-medicamentos-no-brasil-consumidor*

[3] O software OpenRefine é uma ferramenta de código aberto utilizada para a limpeza e transformação de dados. Disponível em https://openrefine.org/.

[4] O formato camelCase integra o uso de vários softwares e linguagens por meio da padronização sugerida na terminologia de nomes de variáveis. Disponível em: https://pt.wikipedia.org/wiki/CamelCase.

[5] *General Refine Expression Language* (GREL). É uma linguagem de script similar ao javascript utilizada no ambiente do OpenRefine.

- Abaixo de R$100
- Entre R$100 e R$1000
- Entre R$1000 e R$50000
- Entre R$50000 e 1 milhão
- Acima de 1 milhão de reais

Em seguida, foram geradas redes de informação por meio de mapeamento realizado pelo software OpenRefine para o formato de rede GML[6] reconhecido pelo software Gephi[7]. A Figura 2 mostra um dos mapeamentos realizados para a criação da rede tripartida citada na seção 3.3. Esse código da Figura 2 é dividido em duas partes: a geração dos nós, identificados pela cláusula **'node'**, e a geração de arestas, identificadas pela cláusula **'edge'**. Existe uma configuração inicial para determinar que a rede não é direcionada, feita com o comando '**directed 0**'.

O comando '**{{jsonize(cell.classeTerapeutica.value)}}**', aplicado ao primeiro grupo de nós, extrai o valor da variável (coluna da base) '**classeTerapeutica**'. A parte denominada de **'variável'**, que é associada a cada nó gerado, classifica os nós em grupos, nesse caso, '**grupo A'** e '**grupo B'**. A criação desses grupos de nós, ainda nessa fase de mapeamento, facilita depois o processo de aplicação de projeção bipartida e técnicas de ARS no software de análise de redes.

Figura 2 - Código de mapeamento GML pelo software OpenRefine para geração de rede tripartida.

[6] GML (Graph Modelling Language) é um formato para representação de grafos de fácil leitura por humanos e com uma capacidade semântica razoável para configurar as características da rede, dos nós e das arestas. Disponível em: https://en.wikipedia.org/wiki/Graph_Modelling_Language/.

[7] GEPHI é um software de código aberto utilizado para visualização, análise e manipulação de redes e grafos. Disponível em https://gephi.org/.

```
graph [
directed 0


node [ id {{jsonize(cells.classeTerapeutica.value)}}
variavel "Classe Terapeutica" agrupamento "grupo A" ]

node [ id {{jsonize(cells.faixaPrecoFabrica17.value)}}
variavel "Faixa de Preço de Fabrica" agrupamento "grupo
B" ]

node [ id {{jsonize(cells.regimeDePreco.value)}}
variavel "Regime de Preço" agrupamento "grupo B" ]


edge [ source {{jsonize(cells.classeTerapeutica.value)}}
target {{jsonize(cells.faixaPrecoFabrica17.value)}}]

edge [ source {{jsonize(cells.classeTerapeutica.value)}}
target {{jsonize(cells.regimeDePreco.value)}}]

]
```

Fonte: autoria própria.

3.3 Mineração de dados, interpretação e avaliação

Na fase de mineração utilizou-se a metodologia de análise de redes complexas que, no caso da presente pesquisa, as redes são classificadas como redes de informação. Também foi criado um *dashboard* para a visualização sintetizada de alguns dados.

Os objetivos da análise de redes resumiram-se a analisar a relação entre laboratórios (de produção dos medicamentos) e tarja, a relação entre classes terapêuticas, faixa de preço de revenda e a relação entre laboratórios e faixa de preço de fábrica, classe terapêutica e regime de preço via projeção bipartida com geração de uma rede monopartida de classe terapêutica.

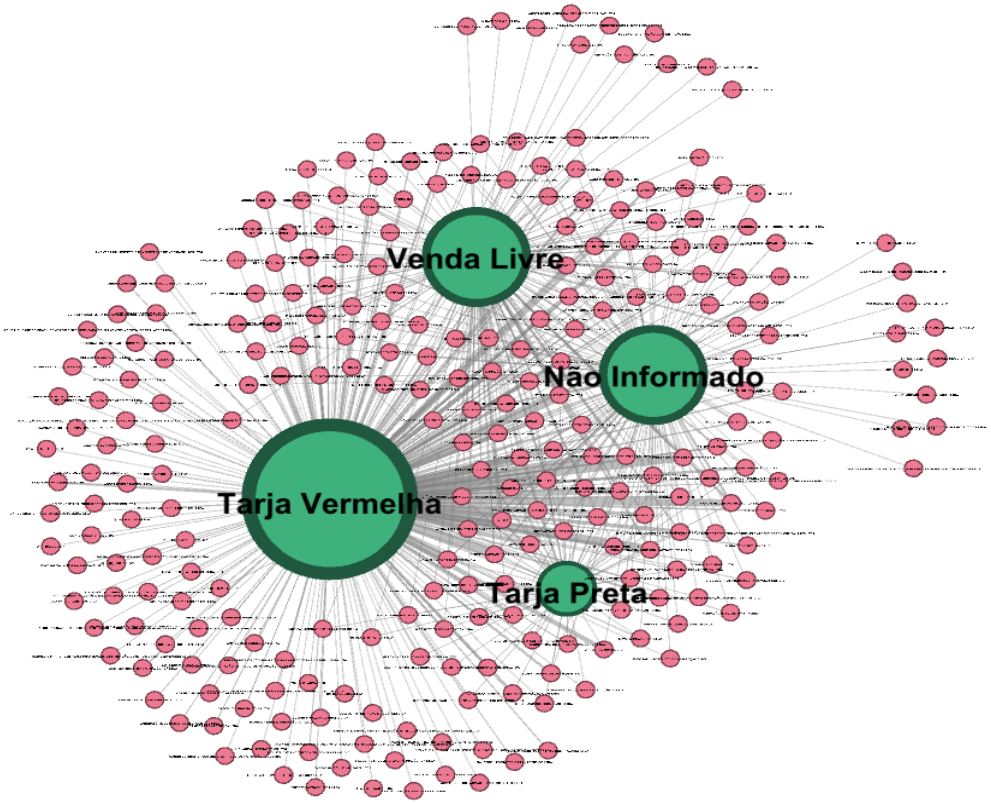
As redes brutas geradas foram organizadas utilizando os métodos de distribuição Yifan Hu e Fruchterman Reingold, cujos algoritmos estão disponíveis no software Gephi. A visualização dos dados, utilizando *dashboard*, foi desenvolvida no software Microsoft Power BI Desktop a partir da base de dados em extensão '.xlsx' exportada pela ferramenta OpenRefine, após as etapas de pré-processamento e transformação. O *dashboard* foi construído a partir da seleção de laboratórios, tarja e regime de preço, exibindo a comparação entre preço de fábrica e preço de revenda para cada produto, a distribuição de todos os produtos em faixas de preço de fábrica e o número de medicamentos por tarja.

4 Resultados e Discussão

Na etapa de mineração, utilizando-se de técnicas de análise de redes de informação, destacaram-se algumas redes envolvendo os seguintes nós: classes de preço de fábrica, laboratórios, e tarja. O relacionamento entre as variáveis tarja e laboratórios, Figura 3, gerou

uma rede que sugere uma predileção dos laboratórios pela produção de medicamentos de tarja vermelha, que são vendidos apenas sob a prescrição de médicos ou dentistas. Entre os principais medicamentos de tarja vermelha comercializados no Brasil estão os fármacos para o tratamento de diabetes, hipertensão e medicamentos psicotrópicos.

Figura 3 - Rede informacional bipartida relacionando as variáveis tarja e laboratório.



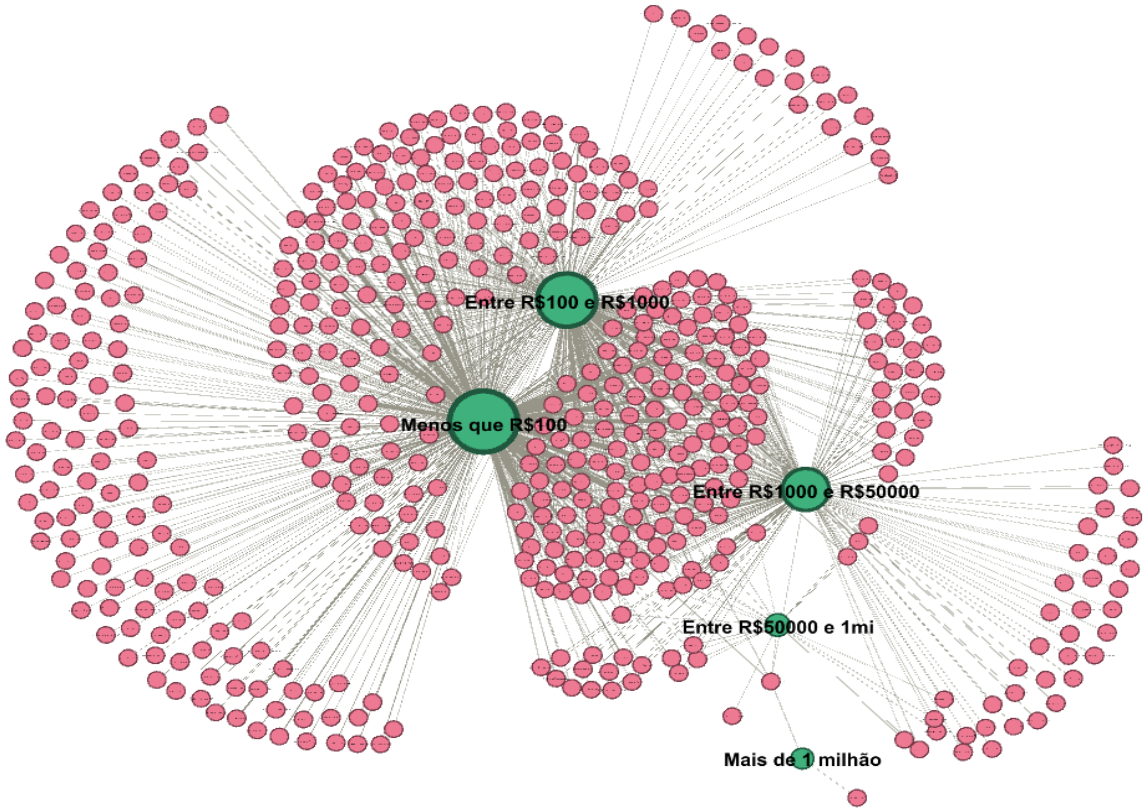
Fonte: autoria própria, com apoio do software Gephi.

Na mesma rede nota-se um alto número de registros onde a tarja do produto não é informada, além de observar 11 laboratórios que não informaram a tarja de nenhum dos seus medicamentos produzidos. Além disso, nenhum dos laboratórios produz apenas medicamentos de tarja preta, diferente do que ocorre com as tarjas vermelha e venda livre.

Quando relacionamos as variáveis faixas de preço de fábrica e laboratórios, Figura 4, observou-se que a maioria dos laboratórios tem em seu catálogo produtos cujo preço para os varejistas custa menos do que R$1.000,00, sendo a faixa '*Menos que R$100*' a maior parcela entre todas. O número de laboratórios vai diminuindo conforme aumenta-se o valor dos produtos, chegando ao extremo onde apenas um laboratório produz medicamentos com valor de fábrica acima de 1 milhão de reais. Nesta mesma rede, também foi possível observar uma significativa quantidade de laboratórios que produzem apenas medicamentos com valor de

fábrica abaixo de R$100. Ainda, dois dos 268 laboratórios produzem apenas medicamentos com custo acima de R$50.000,00 e abaixo de 1 milhão de reais.

Figura 4 - Rede informacional bipartida relacionando as variáveis faixa de preço de fábrica e classe terapêutica.



Fonte: autoria própria, com apoio do software Gephi.

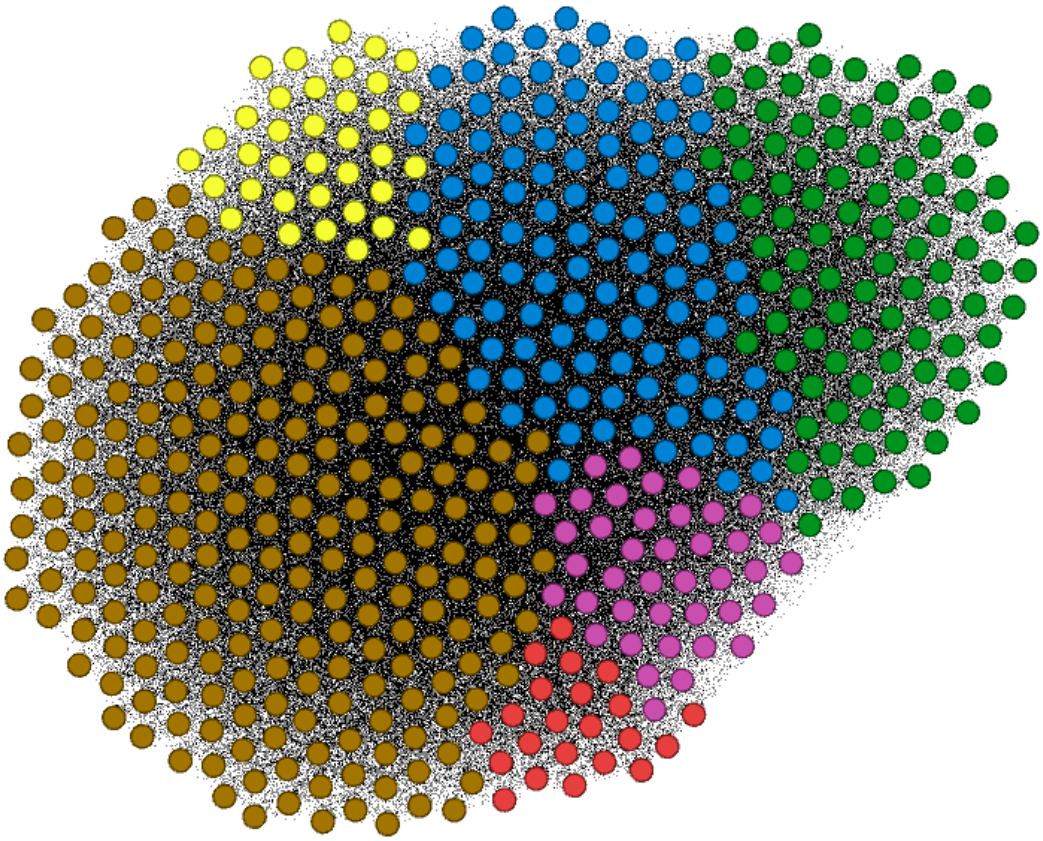
A rede monopartida, Figura 5, formada por nós da Classe Terapêutica por meio de projeção bipartida com as variáveis Faixa de Preço de Fábrica e Regime de Preço possui 534 nós, equivalentes às diferentes classes terapêuticas. Ao realizar o cálculo de modularidade, foram gerados cinco agrupamentos de classes terapêuticas identificados na rede por cores:

- Marrom: 224 nós
- Azul: 115 nós
- Verde: 97 nós
- Roxo: 40 nós
- Amarelo: 36 nós
- Vermelho: 22 nós

Destaca-se, por exemplo, que o grupo verde é composto por 60% de classes terapêuticas de venda livre (tarja), enquanto que o segundo grupo com maior proporção desta categoria de

tarja conta com apenas 10%. Com relação ao tipo de produto, o agrupamento verde também é o que possui o maior número de fitoterápicos, com 8,32%, enquanto que em todos os outros grupos esta porcentagem não chega a 1%. Ainda, no grupo marrom há aproximadamente o dobro ou mais de classes terapêuticas com restrição hospitalar quando comparado aos outros agrupamentos.

Figura 5 - Rede informacional monopartida relacionando as variáveis faixa de preço de fábrica, classe terapêutica e regime de preço colorida após cálculo de modularidade.

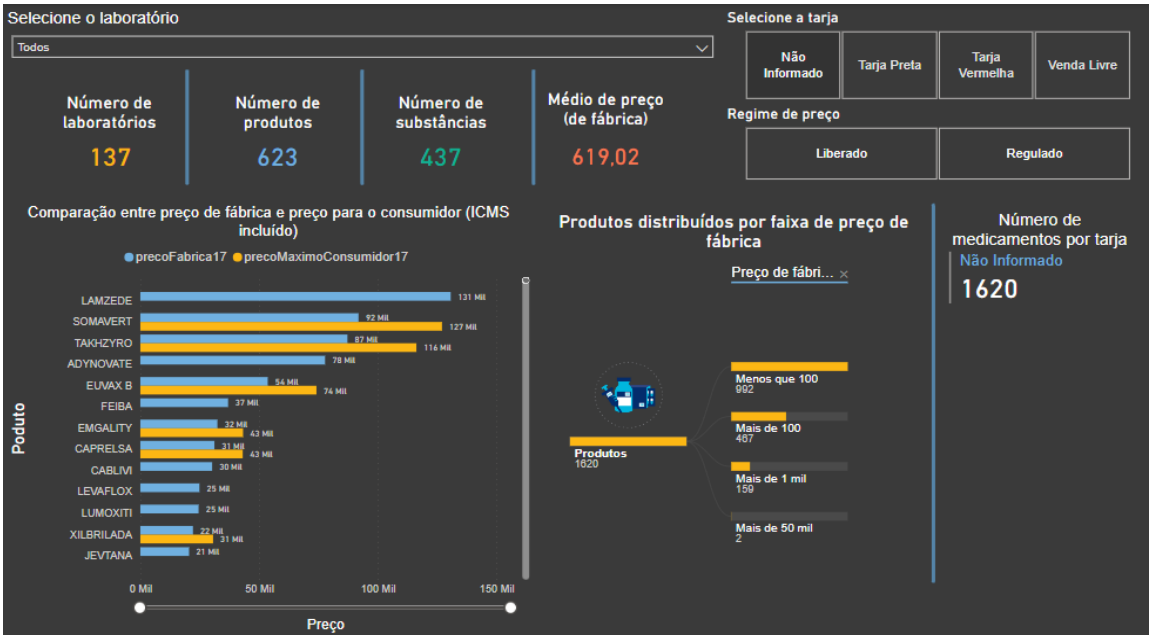


Fonte: autoria própria, com apoio do software Gephi.

O *dashboard* gerado com o *software* Power BI Desktop, Figura 6, permitiu observações amplas da base de dados como, por exemplo, o número total de laboratórios, produtos e substâncias e o cálculo, por exemplo, do valor médio do preço de fábrica de um subconjunto específico de dados. O painel possibilitou a segmentação precisa da base de dados, como a escolha de um ou vários laboratórios e o filtro de tarja e de regime de preço. Ainda, foi possível visualizar em gráfico de barras a comparação entre o preço de fábrica e o preço de revenda dos produtos, a distribuição de produtos por faixa de preço e a quantidade de produtos em cada categoria de tarja. No total, observou-se que a base de dados com 268 laboratórios tem registro

de 6242 produtos e 2307 substâncias, sendo que a média do preço de fábrica entre todos os produtos é de aproximadamente R$3.000,00. Nota-se também que a proporção de produtos entre as faixas de preço se manteve independentemente da tarja.

Figura 6 - Dashboard de variáveis da base de dados



Fonte: autoria própria, com apoio do software Power BI Desktop

5 Considerações Finais

É interessante observar que o processo de descoberta de conhecimento, conforme formulado por Fayyad *et al*. (1996), passando pelas fases de seleção, pré processamento, transformação, mineração de dados por meio de técnicas de análise de redes de informação, intepretação/avaliação por meio de visualização de informação proporcionadas pelas redes de informação, Figuras 3, 4 e 5, e o *dashboard* da Figura 6, culminou, como era de se esperar, em uma descoberta de conhecimento, ainda que simples, interessante do ponto de vista de que essas informações seriam difíceis de serem obtidas sem as técnicas usadas. Observa-se também que elementos de ambas as áreas, Ciência de Dados e Ciência da Informação, como o aspecto interdisciplinar, as técnicas de ARS e a vocação do empenho de esforços direcionados a resolução de problemas reais da sociedade, conduziram a pesquisa de uma forma fluida aproveitando-se de potencialidades das duas áreas de conhecimento como a preocupação na organização e recuperação da informação da Ciência da Informação e a mineração de dados da Ciência de Dados.

Dentre as categorias de tarja, a tarja vermelha é a que é produzida pelo maior número de laboratórios, o que pode se relacionar com um cenário em que a hipertensão arterial é uma das

principais causas de morte no Brasil, que acomete 24% das pessoas com mais de 18 anos, e em que a diabetes ocorre em 7% da população. Ambas as condições clínicas são tratadas por remédios controlados de tarja vermelha, em que é obrigatória a prescrição médica. Outras condições de saúde com diferentes prevalências também podem contribuir para o aumento da demanda de medicamentos de tarja vermelha.

Muitos registros de produtos não incluíram a informação de tarja, o que prejudica a análise assertiva da base de dados. Imagina-se que a não inclusão desse tipo de informação é devida à falhas humanas no processo de alimentação da base de dados. Essa situação ilustra um problema grave da gestão de dados e, se tratando de dados públicos, mantidos por uma instituição pública e relacionados com uma temática tão cara à população, calcula-se que o prejuízo gerado pela má gestão dos dados seja potencialmente alto.

Alguns laboratórios se reservam a produzir apenas um tipo específico de tarja, embora nenhum dos 268 laboratórios produza exclusivamente medicamentos de tarja preta. Os medicamentos de tarja preta são os psicoativos que têm um alto potencial de causar dependência, como as morfinas e anfetaminas e, por isso, deduz-se que seu uso restrito limite a demanda destes no mercado, talvez impossibilitando um modelo de negócios farmacêutico com produção exclusiva.

A maioria das classes terapêuticas registradas na base de dados tem o preço de fábrica girando em torno de R$100, enquanto que apenas duas classes terapêuticas tem custo acima de 1 milhão de reais. São elas *M5X - TODOS OS OUTROS FÁRMACOS COM AÇÃO MÚSCULO-ESQUELÉTICA* e *S1X1 - OUTROS PRODUTOS OFTALMOLÓGICOS SISTÊMICOS.* Ambas as classes terapêuticas são produzidas por um único laboratório, Novartis Biociências S.A., que produz todos os produtos com valor acima de um milhão de reais.

As relações intra-cluster da rede monopartida ainda precisam ser investigadas de maneira mais minuciosa a fim de entender de forma mais plena quais as características fundamentais de cada um dos agrupamentos e a sua relevância para esta pesquisa.

Portanto, o objetivo da pesquisa foi parcialmente alcançado uma vez que algumas relações entre as variáveis foram descobertas e investigadas. Contudo, é necessária a continuação do esforço de trabalho para analisar possíveis relações ainda não descobertas na base de dados do estudo.

Referências

ALBERT, Réka; BARABÁSI, Albert-László. Statistical mechanics of complex networks. **Reviews of modern physics**, v. 74, n. 1, p. 47, 2002.

BALDONI, André de Oliveira et al. Elderly and drugs: risks and necessity of rational use. **Brazilian Journal of Pharmaceutical Sciences**, v. 46, p. 617-632, 2010.

BARABÁSI, Albert-László. Network science. **Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences**, v. 371, n. 1987, p. 375, 2013.

BARABÁSI, Albert-László. **Linked: The new science of networks**. Cambridge, MA: PerseusPublishing,2003. 280 p.

BÁRRIOS, Maria João; MARQUES, Rita; FERNANDES, Ana Alexandre. Aging with health: aging in place strategies of a Portuguese population aged 65 years or older. **Revista de Saúde Pública**, v. 54, p. 103 - 138, 2020.

BATISTA, Gustavo Enrique de Almeida Prado. **Pré-processamento de dados em aprendizado de máquina supervisionado**. 2003. Tese (Doutorado) - Universidade de São Paulo.

BORGATTI, Stephen P.; HALGIN, Daniel S. Analyzing affiliation networks. *Em*: SCOTT, John; CARRINGTON, Peter (eds.). **The SAGE Handbook of Social Network Analysis**. London: SAGE, 2014. p. 417–433. DOI:10.4135/9781446294413.n28. Disponível em:https://methods.sagepub.com/book/the-sage-handbook-of-social-network-analysis/n28.xml. Acesso em: 2 mar. 2023.

BLASCO PATIÑO, F. *et al.* Estudio del consumo de fármacos inadecuados o no indicados en el anciano que ingresa en un Servicio de Medicina Interna. In: **Anales de Medicina Interna**, Madrid, v. 25, n. 6, p. 269-274, 2008.

BURKHARD, Remo Aslak. Towards a framework and a model for knowledge visualization: Synergies between information and knowledge visualization. Knowledge and information visualization: Searching for synergies**. Lecture Notes in Computer Science**, 3426, p. 238-255, 2005.ISBN-13: 978-3540269212

CAPURRO, Rafael; HJØRLAND, Birger. The concept of information. **Annual Review of Information Science and Technology**, *[S. l.]*, v. 37, p. 343–411, 2003. DOI:10.1590/S1413-99362007000100012. Disponível em: http://fiz1.fh-potsdam.de/volltext/stuttgart/04058.html. Acesso em: 2 mar. 2023.

CHEN, Chaomei. **Mapping scientific frontiers: the quest for knowledge visualization**. 2. ed. London: Springer Science & Business Media, 2013. 215p.

COSTA, Claudio Napolis *et al*. Descoberta de conhecimento em bases de dados. **Revista Eletrônica: Faculdade Santos Dumont**, v. 2, p. 20, 2019. Disponível em: https://www.fsd.edu.br/wp-content/uploads/2019/12/artigo9.pdf. Acesso em: 05 mar. 2023.

CHAIMOWICZ, Flávio. A saúde dos idosos brasileiros às vésperas do século XXI: problemas, projeções e alternativas. **Revista de Saúde Pública**, v. 31, p. 184-200, 1997.

EVERETT, M. G.; BORGATTI, S. P. The dual-projection approach for two-mode networks. **Social Networks**, *[S. l.]*, v. 35, n. 2, p. 204–210, 2012. DOI:10.1016/j.socnet.2012.05.004. Disponível em:https://linkinghub.elsevier.com/retrieve/pii/S0378873312000354. Acesso em: 2 mar. 2023.

FAYYAD, Usama; PIATETSKY-SHAPIRO, Gregory; SMYTH, Padhraic. The KDD process for extracting useful knowledge from volumes of data. **Communications of the ACM**, v. 39, n. 11, p. 27-34, 1996.

GAO, Man; CHEN, Ling; LI, Bin; LI, Yun; LIU, Wei; XU, Yong-cheng. Projection-based link prediction in a bipartite network. **Information Sciences**, *[S. l.]*, v. 376, p. 158–171, 2017. DOI:10.1016/j.ins.2016.10.015. Disponível em: https://doi.org/10.1016/j.ins.2016.10.015. Acesso em: 2 mar. 2023.

HAND, David J. Principles of data mining. **Drug safety**, v. 30, n. 7, p. 621-622, 2007.

HAND, David J.; MANNILA, Heikki; SMYTH, Padhraic. **Principles of data mining**. Cambridge, MA: MIT Press, 2001. 556p.

HIGGINS, Silvio Salej; RIBEIRO, Antonio Carlos Andrade. **Análise de redes em Ciências Sociais**. Brasília: Enap, 2018. Disponível em:https://repositorio.enap.gov.br/bitstream/1/3337/1/Livro_Analise%20de%20Redes%20em%20Ci%C3%AAncias%20Sociais.pdf. Acesso em: 2 mar. 2023.

KADUSHIN, Charles. **Introduction to social network theory**. Boston, MA, *[S. l.]*, 2004. Disponível em: http://melander335.wdfiles.com/local--files/reading-history/kadushin.pdf. Acesso em: 2 mar. 2023.

KAUFMAN, David W. *et al.* Recent patterns of medication use in the ambulatory adult population of the United States: the Slone survey. **Jama**, v. 287, n. 3, p. 337-344, 2002.

MACEDO, Giani Rambaldi *et al*. O poder do marketing no consumo excessivo de medicamentos no Brasil. **Revista Transformar**, v. 9, p. 114-128, 2016.

MARTINS, Dalton Lopes. Data science teaching and learning models: focus on the Information Science area. In: RODRIGUES DIAS, Thiago Magela (org.). **Advanced Notes in Information Science**. [s.l.] : ColNes Publishing, 2022. v. 2. DOI:10.47909/anis.978-9916-9760-3-6.100. Disponível em: https://pub.colnes.org/index.php/anis/article/view/100. Acesso em: 2 mar. 2023.

MELAMED, David. Community Structures in Bipartite Networks: A Dual-Projection Approach. **PLOS ONE**, *[S. l.]*, v. 9, n. 5, p. e97823, 2014. DOI:10.1371/journal.pone.0097823. Disponível em: https://doi.org/10.1371/journal.pone.0097823. Acesso em: 2 mar. 2023.

METZ, Jean et al. **Redes complexas: conceitos e aplicações**: Relatório Técnico do ICMC. São Carlos: Universidade de São Paulo, 2007. 45p. Disponível em: https://repositorio.usp.br/bitstreams/30f00c12-d53f-4c46-911f-a84b360575a3&hl=pt-BR&sa=T&oi=gsb-gga&ct=res&cd=0&d=13967733084527102919&ei=jr0EZPPWGoKLmwGN2aygDQ&scisig=AAGBfm29Fft7twysMBY8kaz_LZnL0XPukw. Acesso em: 05 mar. 2023.

MOHAMMED, Mohammed A.; MOLES, Rebekah J.; CHEN, Timothy F. Impact of pharmaceutical care interventions on health-related quality-of-life outcomes: a systematic review and meta-analysis. **Annals of Pharmacotherapy**, v. 50, n. 10, p. 862-881, 2016.

NEWMAN, M. E. J. **Networks: an introduction**. Oxford ; New York: Oxford University Press, 2010. 772p.

NOOY, Wouter De; MRVAR, Andrej; BATAGELJ, Vladimir. **Exploratory social network analysis with Pajek: revised and expanded edition for updated software**. 3rd ed. New York: Cambridge University Press, 2018. Acesso em: 2 mar. 2023.

OTTE, Evelien; ROUSSEAU, Ronald. Social network analysis: a powerful strategy, also for the information sciences. **Journal of Information Science**, *[S. l.]*, v. 28, n. 6, p. 441–453, 2002. DOI:10.1177/016555150202800601. Disponível em: https://doi.org/10.1177/016555150202800601. Acesso em: 2 mar. 2023.

PENNA, Giuseppe Della; MAGAZZENI, Daniele; OREFICE, Sergio. A spatial relation-based framework to perform visual information extraction. **Knowledge and Information Systems**, *[S. l.]*, v. 30, n. 3, p. 667–692, 2012. DOI:10.1007/s10115-011-0394-4. Disponível em: https://link.springer.com/article/10.1007/s10115-011-0394-4. Acesso em: 2 mar. 2023.

PORTO, Fábio; ZIVIANI, Arthur. Ciência de Dados. *Em*: 2014, Rio de Janeiro. **Anais** [...]. . In: SEMINÁRIO DE GRANDES DESAFIOS DA COMPUTAÇÃO NO BRASIL. Rio de Janeiro: SBC, 2014. Disponível em: https://www.lncc.br/~ziviani/papers/III-Desafios-SBC2014-CiD.pdf. Acesso em: 2 mar. 2023.

SARACEVIC, Tefko. Interdisciplinary nature of information science. **Ciência da informação**, Brasília, v. 24, n. 1, p. 36–41, 1995. Disponível em: http://www.brapci.inf.br/_repositorio/2010/03/pdf_dd085d2c4b_0008887.pdf. Acesso em: 2 mar. 2023.

SOUZA, Queila; QUANDT, Carlos. Metodologia de análise de redes sociais. *Em*: DUARTE, F.; QUANDT, Carlos; SOUZA, Queila (eds.). **O Tempo das redes**. São Paulo: Perspectiva, 2008. p. 31–63. Disponível em:https://www.academia.edu/257818/Metodologia_De_An%C3%A1lise_De_Redes_Sociais. Acesso em: 2 mar. 2023.

VIRKUS, Sirje; GAROUFALLOU, Emmanouel. Data science from a library and information science perspective. **Data Technologies and Applications**, v. 53, n. 4, p. 422–441, 2019. DOI:10.1108/DTA-05-2019-0076. Disponível em: https://doi.org/10.1108/DTA-05-2019-0076. Acesso em: 2 mar. 2023.

WASSERMAN, Stanley; FAUST, Katherine. **Social network analysis: methods and applications**. Cambridge, England; New York: Cambridge University Press, 1994.

ZHANG, Jinson. **Visualization for information retrieval**. Berlin: Springer, 2008.

ZHOU, Xue Zhong; MENCHE, Jörg; BARABÁSI, Albert-László; SHARMA, Amitabh. Human symptoms–disease network. **Nature Communications**, v. 5, n. 1, p. 4212, 2014. DOI:10.1038/ncomms5212. Disponível em: https://www.nature.com/articles/ncomms5212. Acesso em: 2 mar. 2023.