

XX ENANCIB

21 a 25 Outubro/2019 – Florianópolis

A Ciência da Informação e a era da Ciência de Dados

ISSN 2177-3688

GT-1 – Estudos históricos e epistemológicos da Ciência da Informação

MINERAÇÃO DE DADOS NA PESQUISA EM CIÊNCIA DA INFORMAÇÃO: DESAFIOS E OPORTUNIDADES

DATA MINING IN INFORMATION SCIENCE RESEARCH: CHALLENGES AND OPPORTUNITIES

Josir Cardoso Gomes - Instituto Brasileiro de Informação em Ciência e Tecnologia – Universidade Federal Fluminense

Ricardo Medeiros Pimenta - Instituto Brasileiro de Informação em Ciência e Tecnologia – Universidade Federal Fluminense

Marco Schneider - Instituto Brasileiro de Informação em Ciência e Tecnologia – Universidade Federal Fluminense

Modalidade: Trabalho Completo

Resumo: Os termos mineração de dados, ciência de dados, aprendizado de máquina têm invadido não só as manchetes de jornais e as redes sociais como também têm aparecido com uma maior frequência em periódicos científicos. Os propagados avanços da Inteligência Artificial têm aberto o debate sobre questões éticas do uso destas tecnologias no cotidiano e também no próprio fazer científico. Assim, a questão inicial que se traz é: como as ciências sociais e as humanidades estão se apropriando destas novas técnicas? E, em especial, como a Ciência da Informação pode contribuir para o entendimento epistemológico destas novas áreas do conhecimento sem cair em uma narrativa meramente comercial ou em um discurso “revolucionário” de como a sociedade será “transformada” com o uso de tais tecnologias? Através de revisão bibliográfica e pesquisa exploratória, este ensaio procurou analisar como a Ciência da Informação tem utilizado técnicas de mineração de dados e se o campo pode trazer contribuições importantes para o uso de Mineração de Dados, tanto sob o ponto de vista epistemológico como também em seus usos práticos, especificamente na área de sistemas de organização do conhecimento.

Palavras-Chave: Mineração de Dados; Ciência de Dados; Epistemologia.

Abstract: *The terms data mining, data science, machine learning have invaded not only newspaper headlines and social media but have also appeared more frequently in scientific journals. The propagated advances of Artificial Intelligence have opened the debate on ethical issues of the use of these technologies in everyday life and also in scientific practice itself. Thus, the initial question that arises is: how are the social sciences and the humanities adopting these new techniques? And, in particular, how can Information Science contribute to the epistemological understanding of these new areas of knowledge without falling into a merely commercial narrative or a "revolutionary" discourse of how society will be "transformed" by the use of such technologies? Through literature review and exploratory research, this essay sought to analyze how the Information Science has used data mining techniques and whether the field can make important contributions to the use of Data Mining both from an epistemological point of view and also in its practical uses specifically in the area of knowledge organization systems.*

Keywords: *Data Mining; Data Science; Epistemology.*

1 INTRODUÇÃO

Os termos mineração de dados (MD), ciência de dados, aprendizado de máquina (*machine learning*), redes neurais (*neural networks*) e outros correlatos têm invadido não só as manchetes de jornais e as redes sociais, como também têm aparecido com maior frequência em periódicos científicos. Apesar de boa parte destes conceitos terem sido criados pela Ciência da Computação desde a década de 60, somente agora seus usos passam a ser popularizados em função do maior processamento dos computadores pessoais, da disponibilização de serviços na nuvem e da maior oferta de produtos e serviços que fazem uso dessas tecnologias e processos.

A discussão sobre Inteligência Artificial (IA) também tem aparecido constantemente no noticiário, seja para relatar os últimos testes (e acidentes) de carros autônomos ou para anunciar os novos dispositivos domésticos como o Alexa e o Google Home, que já podem ser adquiridos a preços módicos. Entusiastas na área alardeiam que o uso de IA será um tsunami (MANNING, 2015) ou uma revolução (SEJNOWSKI, 2018) que irá alterar profundamente a sociedade, incluindo aí o próprio fazer científico. E, de fato, já é possível encontrar sinais de que tal situação é cada vez mais concreta. Filip Vostal já aponta que a digitização (GOBBLE, 2018) das práticas de pesquisa - em uma medida anteriormente impossível - pelo menos parcialmente -move processos analíticos de pesquisa social cada vez mais para a direção e domínio da computação (VOSTAL, 2016, p. 178). Neste escopo, seria possível, ao menos desejável, que refletíssemos se as universidades estariam sendo (re)sincronizadas no sistema capitalista através de seus imperativos materiais e culturais que, nesse caso, operariam pelo apelo aos recursos computacionais relacionados ao estudo e uso do Big Data (VOSTAL, 2016, p. 62). Recursos estes que estariam ligados a um mercado em ebulição e, portanto, sob o paradigma do capital e de sua aceleração compulsória.

Mesmo fazendo um filtro de todo o discurso mercadológico e de marketing de empresas interessadas em fazer novos negócios, percebe-se que o próprio meio científico já aderiu a este discurso. Refletir sobre qual cenário influencia o outro é tarefa árdua. Ao passo que a cena científica produz avanços relacionados à temática, é esperado que o mercado se aproprie dos respectivos avanços na intenção de reificá-los (LUKACS, 1989) ao passo que este mesmo também acaba por influenciar nas escolhas de temas, objetos de pesquisa e problemáticas que reincidem na academia buscando maiores esclarecimentos. Cursos de

“ciência de dados” começam a surgir em várias universidades, tanto na área da Estatística e de Computação que seriam o território natural da chamada Ciência de Dados, mas também em programas de Ciência da Informação.

Especificamente na CI, já existem inúmeros estudos de como os algoritmos têm impactado os aspectos sociais, econômicos e culturais na sociedade contemporânea (PIMENTA, 2013; BEZERRA, 2015; CALDAS e CALDAS, 2019). Questões sobre a invasão de privacidade, vigilância, desinformação e os fluxos informacionais e econômicos das redes sociais privadas já fazem parte dos anais e revistas científicas do campo. Entretanto, o campo está se aproveitando destas tecnologias para criar novos métodos de pesquisa que consiga analisar tal volume e variedade de dados?

Assim, a motivação inicial para esta pesquisa partiu do seguinte questionamento: em que medida pesquisadores da CI estão fazendo uso de técnicas de mineração de dados como método de pesquisa? E de forma inversa, pode-se perguntar também: existe interdisciplinaridade entre a mineração de dados e a CI? Afinal, o estudo crítico sobre a influência algorítmica e como as novas tecnologias de informação impactam a sociedade são uma das principais vocações do campo da Ciência da Informação apesar desse mesmo não ainda ter contato com contribuições importantes feitas *dehors* de suas próprias fronteiras. Exemplo disso é o trabalho de Antoinette Rouvroy e Thomas Berns, *Gouvernementalité algorithmique et perspectives d’émancipation: Le disparate comme condition d’individuation par la relation?* (2013). Nele a discussão em torno de um pensamento crítico sobre como a sociedade, as políticas públicas, seus atores e a própria legislação precisarão ser reavaliados à luz de uma perspectiva crítica.

Como metodologia, utilizou-se a pesquisa bibliográfica e exploratória. A pesquisa bibliográfica faz uso de documentos de fontes reconhecidas, sobretudo de livros e artigos científicos (GIL, 2006) enquanto que a pesquisa exploratória procura examinar o conhecimento sobre o tema pesquisado com perspectivas distintas (SAMPIERI; COLLADO; LUCIO, 2013).

Para tentar mapear as questões abordadas, o estudo segue dividido em três partes: se inicia a partir da montagem dos marcos teóricos sobre os conceitos do termo “dado” no âmbito da Ciência da Informação. Esta introdução se mostra essencial em função das inúmeras definições encontradas na literatura não só do termo dado, mas também dos

termos informação e conhecimento. Como diversos pesquisadores utilizam tais termos de forma distinta (CAPURRO; HJØRLAND, 2007), é importante que se delimite quais autores serão utilizados como referência para as reflexões que serão apresentadas. Em seguida, traremos as definições para Mineração de Dados, aprendizado de máquina e Ciência de Dados, a correlação entre eles e, principalmente, quais as relações interdisciplinares com a CI dentro do referencial teórico explicado na primeira seção do artigo.

Por fim, com o objetivo de analisar como tais técnicas estão sendo utilizadas no campo, realizou-se uma busca de artigos em periódicos e teses no Brasil que utilizaram técnicas de MD na CI. E com base nos resultados encontrados, busca-se uma reflexão sobre as possibilidades do uso de mineração de dados na área e se a integração do ferramental teórico da CI em conjunto com a MD pode trazer contribuições para outras áreas de conhecimento.

2 DADO, INFORMAÇÃO E CONHECIMENTO

Para que se possa entender o conceito de Mineração de Dados (MD), é importante que se defina primeiramente o conceito do que é um Dado. Tradicionalmente, o termo “dado” geralmente é considerado como a base da hierarquia DIKW (*Data, Information, Knowledge and Wisdom*) onde informação é entendida como dado processado, conhecimento como informação processada ou interpretada e sabedoria como conhecimento interpretado (HJØRLAND, 2018).

Já a tríade Dado, Informação e Conhecimento também já foi amplamente discutida na CI apesar de não existir um consenso entre os principais autores da área. Buckland (1991) procurou trazer o aspecto da materialidade da Informação (Information as Thing), bem como da informação como processo e como fenômeno cognitivo. Capurro e Hjørland (2007) discutiram sobre o conceito de Informação e seu caráter polissêmico e Furner (2016) realizou uma extensa pesquisa não só sobre a história e etimologia do termo dado (*data* em inglês), mas trouxe também os vários usos do termo em diferentes áreas do conhecimento.

Desta forma, um dos objetivos iniciais do artigo é propor uma estrutura teórica que traga uma definição o mais objetiva possível do que é um “dado” para, em seguida, se possa adentrar nos aspectos da MD e da sua interligação com a Ciência da Informação. Ou

seja, não se busca aqui listar todas as definições existentes, mas, ao contrário, delimitar o termo para que se possa construir um raciocínio sobre o conceito de mineração de dados.

Diferentemente dos dois textos de Furner (2016) e de Capurro e Hjørland (2007) onde os autores detalham a formação dos termos “dado” e “informação” através da história e de sua etimologia e, tal como em um livro de suspense, constroem uma narrativa onde somente no final da “trama”, no clímax dos respectivos artigos, trazem algumas possibilidades para a definição dos dois conceitos, aqui se fará o caminho inverso. Uma definição será proposta e, em seguida, o raciocínio utilizado para a escolha da mesma será detalhado. Obviamente esta definição não tem a pretensão de ser definitiva. Pelo contrário, o objetivo é justamente trazer novas reflexões que possam enriquecer o entendimento do assunto e delimitar o termo para o objetivo central do artigo. Isso posto, a definição candidata dentro do modelo teórico aqui proposto para o termo dado é: **Dado é o registro do atributo de um ente, objeto ou fenômeno.**

Registro é o ato de registrar, ou seja, é a gravação ou a impressão de caracteres ou símbolos que tenham um significado em algum **documento ou suporte físico**. A temperatura de uma lâmina de aço só se transforma em dado, quando ela é medida e registrada, a cor de uma planta só passa a ser um dado quando ela é registrada seja através de uma fotografia ou por uma pintura à mão. Assim, se a temperatura ou a cor forem apenas observadas, mas não houver o registro, esses atributos não podem ser considerados um dado.

Propositalmente, o termo “registro” foi escolhido justamente para trazer a ideia de documento, ou seja, não há dado sem documento. Interessante notar que à informação também se preconiza a existência do registro, do documento. Daí encontramos um ponto de grande dúvida no senso comum sobre estas duas categorias: dado e informação. Se não houver o registro, se não houver a medição, se não houver a observação, não existe dado. Desta forma, o termo registro está intimamente relacionado ao ato de documentar e organizar pois para que haja o registro, deve-se criar minimamente um conjunto de regras para que tal registro fique ordenado para que futuramente possa ser recuperado.

Já o termo **atributo** se refere a uma característica (uma propriedade) de um ente, objeto ou fenômeno. O termo “ente” foi uma opção preferencial pois estaria englobando tanto os seres vivos como os inanimados e também o que se encontra no universo da imaginação. Assim, recairia no conjunto dos entes, uma estrela, um elefante, um micróbio,

um martelo, uma poesia ou um personagem de ficção. Ou a lista de atributos de um personagem de Jorge Amado não podem ser considerados como dados de tal personagem?

O termo “objeto” foi também incluído pois é utilizado por vários autores de referência na CI (DAHLBERG, 1978; BRÄSCHER, 2014) e, por fim, optou-se também por incluir o termo fenômeno para incluir o registro de um fato, de algo que ocorreu em um determinado intervalo de tempo, e que efetivamente não está ligado a um objeto ou ente específico. Como exemplo, pode-se citar os fenômenos meteorológicos ou a fala de um indivíduo.

A definição de “dados” utilizada aqui tem semelhança com definições propostas por alguns autores. Christine Borgman (2016) reforça que não existe uma definição única para dados, mas também propõe uma definição para o termo. Para a autora, um dado é uma “entidade utilizada como evidência de um fenômeno para propósitos de pesquisa ou ensino” (BORGMAN, 2016, p.29, tradução nossa)¹. Esta definição fica bem próxima da nossa pois reforça o conceito de registro através do uso do termo “evidência”. Entretanto, o uso do termo “entidade” não consegue retratar bem o que efetivamente está sendo registrado.

Já Semeler e Pinto definem dado como “todo objeto criado em formato digital ou convertido para o formato digital que possa ser usado para geração de insights de informação e conhecimento” (2019, p.115). Aqui a delimitação de que o dado só existe digitalmente diminui a universalidade do termo já que a sociedade continua criando dados não digitais. Fica também em aberto o que seria efetivamente um “insight”.

Dessa forma, tendo uma definição candidata para “dado”, passa-se a uma outra fase tão ou mais importante. Ora, se houve o registro do atributo de um objeto, como ele se deu? Quem fez o registro? O que significa esse atributo que está sendo registrado? Ou seja, por trás de todo o dado, existe sempre um conjunto de **metadados**. Nesta construção conceitual, os metadados são os dados que detalham como o dado foi registrado. É o dado sobre o dado.

O primeiro uso do termo "metadados" ocorreu em janeiro de 1967 por Stuart McIntosh e David Griffel que utilizaram o prefixo “meta” (do grego τὰ μετὰ) para tentar sintetizar o conceito "sobre os dados" e provavelmente a coleção de ensaios sobre metafísica de Aristóteles serviu de inspiração para a confecção do termo (STEINER, 2017). Entretanto, o conceito por trás da definição do termo pode ser rastreado desde a antiguidade: uma das

¹ Do original: “data refers to entities used as evidence of phenomena for the purposes of research or scholarship”

primeiras formas de metadados é a obra de Pinakes, criada em 245 a.C. por Kallimachos de Cirene para catalogar a Biblioteca de Alexandria de forma sistematizada (POMERANTZ, 2015, p.6), ou seja, o termo metadado está intimamente associado à Biblioteconomia e por consequência à Ciência da Informação.

Nota-se que o uso do termo metadado associado ao documento bibliográfico difere da definição inicial já que, nesse contexto, a definição seria “os dados sobre o documento” e não a definição inicial “os dados sobre o dado”. Entretanto, se o dado (ou o conjunto de dados) for considerado um tipo de documento, as duas definições não se contradizem. E realmente, a partir da ideia inicial de que o dado pressupõe um registro, não seria estranho pensar que o dado pode ser considerado um tipo de documento. Furner corrobora essa relação entre documento e os dados:

De fato, não é verdade que os documentos sejam constituídos por dados, nem que o documento seja uma espécie de conjunto de dados: é o contrário, em ambos os aspectos. Um conjunto de dados é constituído por documentos; e o conjunto de dados é uma espécie de documento (p. 303, tradução nossa)²

Outras definições de metadado já incorporam o termo “informação” criando ainda mais ruído na tentativa de se encontrar uma definição mais precisa. Por exemplo, a *National Information Standards Association* (associação que organiza padrões informacionais nos EUA) define metadado como “informação estruturada que descreve, explica, localiza ou facilita a recuperação, utilização ou gestão de um recurso de informação”³ (NISA, 2004).

Henning Spang-Hanssen (2001) corrobora a questão da necessidade de como o dado foi coletado, apesar de não utilizar o termo “metadado”:

A informação sobre as propriedades físicas de um material está realmente incompleta sem informação sobre a precisão dos dados e sobre as condições em que esses dados foram obtidos. Além disso, várias investigações de uma propriedade conduziram frequentemente a resultados diferentes que não podem ser comparados e avaliados para além das informações sobre os seus antecedentes. (p.3, tradução nossa)⁴

² No original “it is not in fact the case that documents are made up of data, nor that the document is a species of dataset: rather it is the other way around, in both respects. A dataset is made up of documents; and the dataset is a species of document”.

³ No original “structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use or manage an information resource”

⁴ No original “Information about some physical property of a material is actually incomplete without information about the precision of the data and about the conditions under which these data were obtained. Moreover, various

A complexidade deste modelo teórico pode ainda aumentar um pouco com a seguinte reflexão: para todo dado, pode-se considerar que existem dois tipos de metadados que o acompanham: metadados **explícitos** e **implícitos**. Metadados explícitos são os metadados registrados em conjunto com o dado e os metadados implícitos são sabidos a partir de sua descrição ou de outros metadados explícitos.

Voltando ao exemplo da temperatura da chapa de aço, o dado é o número medido e os metadados explícitos são a indicação do instante em que a medição foi feita, do tipo de material (no caso o aço), do equipamento que realizou a medição e da unidade de medida utilizada. O próprio descritor “temperatura da chapa de aço” é um metadado. Entretanto, não se registra o conceito do que é uma temperatura, pois o produtor deste dado parte do pressuposto que o usuário ou leitor já conheça esse conceito. Ou seja, alguns metadados deste documento já são implícitos ao documento. São implícitos inclusive os metadados registrados: a unidade de medida de temperatura, as propriedades do instrumento de medição e uma infinidade de conceitos que não foram registrados pois se pressupõe que o leitor não necessite saber pois é um dado (ou melhor, um metadado) conhecido.

Desta forma, pode-se estender um pouco mais esse modelo teórico: quando se tem o dado e seus metadados, o usuário consegue ter uma compreensão maior do que aquele dado representa e do **contexto** em que ele foi registrado, ou seja, o conjunto (dado + metadados explícitos e implícitos) pode ser uma possível definição para **informação**. Contexto no qual se identificam conjuntos, como propusemos, nos remete à perspectiva de **relevância** em Tefko Saracevic (1970; 1975), onde tal perspectiva aponta para a eficácia do contato entre fonte e destinatário. Logicamente, se pensarmos no contexto atual da intermediação eletrônica, a relevância se torna categoria mais ainda necessária e, talvez, uma das mais críticas para compreendermos o processo existente entre dados, informação e produção do conhecimento. Tal definição busca resgatar o aspecto cognitivo da informação que é contemplado pelos metadados implícitos que compõem a “fórmula” (dado + metadado), da relevância do contexto e da percepção do usuário que estará recuperando esse dado. A análise ou percepção do conjunto de metadados implícitos é tão ou mais marcante que o aspecto físico ou documental.

Voltando ao exemplo da lâmina de aço, um leitor só conseguirá interpretar o dado da temperatura se ele souber o que significa uma temperatura. Ou seja, o dado foi registrado corretamente, o contexto em que o dado foi coletado foi fornecido mas se o leitor não tiver o **conhecimento** prévio dos termos empregados na definição dos metadados apresentados, o dado não se transforma em conhecimento. Conforme atestam Marteleto, Nóbrega e Morado, o “conhecimento que resulta da ação e das relações dos sujeitos em seu fazer na história, em suas práticas culturais, é sempre um conhecimento contextual” (2013, p. 100).

O que está sendo chamado aqui de metadado implícito é efetivamente o que Dahlberg define como conceito em sua Teoria do Conceito (DAHLBERG, 1978). E esta teoria procura tratar toda a problemática do que é um conceito, o que é um termo, o que é uma definição (a definição do que é uma definição), do que é a representação do conhecimento e dos aspectos semióticos que são intrínsecos a uma visão cognitivista e conceitualista.

Ainda segundo Dahlberg, o conceito é um dos elementos centrais para a informação e para a geração do conhecimento e reforça que todo conceito tem uma dimensão física e uma dimensão cognitiva. Capurro joga com as palavras quando afirma que, enquanto “*information is a thing*” (citando Buckland), “*knowledge is no-thing*”, ou seja, não existe materialidade no conhecimento (CAPURRO, 2017). E, por não ser material, a definição do que é conhecimento enfrenta enorme resistência em áreas como a Ciência da Computação:

[...] a visão “conceitualista”, apesar de servir a diversos propósitos na CI como, por exemplo, em instrumentos como os vocabulários controlados, não é adequada à realidade atual, caracterizada pelo uso de computadores. Sistemas automatizados e computadores se tornam elementos obrigatórios ante o grande volume de informação disponível. E computadores necessitam de representações formais (SOUZA; ALMEIDA; BARRACHO, 2013 p.164)

A construção destas representações formais se traduz justamente na transformação de conceitos (unidades de conhecimento) em metadados explícitos. Esta é a grande tarefa da elaboração de instrumentos de construção semântica e de ontologias para que as tecnologias de aprendizado de máquina e IA possam funcionar. Ou seja, quanto mais dados, atributos e metadados estiverem disponíveis para um objeto, melhor o algoritmo poderá “entender” e “aprender” de forma autônoma.

O grande esforço de construir uma Web Semântica (BERNERS-LEE et al, 2001) ou Web 3.0 é um exemplo de como a indústria e a ciência buscam insumos para que os algoritmos

tenham a possibilidade de aprender. Viana (2018) reforça este argumento quando ressalta que o termo “semântica” visa “estabelecer uma interligação dos significados das palavras, o que as possibilita de serem compreendidas por computadores” e complementa que a Web 3.0 pretende ser um espaço digital “desambiguado” pois as máquinas poderão “entender” o significado do conteúdo de um site ou texto sem margem para dúvida na interpretação do que está sendo apresentado.

Apesar do termo Web 3.0 ter sido criado há quase duas décadas, percebe-se que o desenvolvimento da Web Semântica não atingiu seus objetivos com a mesma velocidade que as demais tecnologias da Internet (MICHALAKIDIS, 2016). Mesmo com os inúmeros avanços tecnológicos como a criação de padrões e boas práticas e com o uso de Linked Data, RDF e SPARQL, o ser humano continua sendo peça essencial para a criação da camada extra de dados (e metadados) que devem ser inseridas nas páginas web para que a máquina consiga interpretá-las. A inclusão automatizada de artefatos semânticos nas páginas web ainda não é possível e um dos motivos é a dificuldade em representar o conhecimento.

A Representação do Conhecimento (*Knowledge Representation*) é a disciplina que tem como objetivo a captura e modelagem de representações sobre o mundo que possam ser interpretadas por algoritmos para que estes possam resolver problemas complexos. E esta área tem uma extensa fronteira interdisciplinar com a CI, Linguística, Filosofia, Psicologia e Inteligência Artificial (BRÄSCHER, 2014).

A pesquisadora Marisa Bräscher (2014) ressalta ainda que o conhecimento acumulado na construção de sistemas de indexação e de organização do conhecimento como os tesouros, ontologias e taxonomias podem contribuir no campo teórico e prático para a superação das dificuldades enfrentadas na área de representação do conhecimento e, conseqüentemente para áreas como a Inteligência Artificial e a mineração de dados. Na seção seguinte, a interdisciplinaridade das duas ciências será ainda mais ressaltada.

3 MINERAÇÃO DE DADOS

Se buscarmos na história da ciência da computação, a MD vem sendo utilizadas desde 1960, mas somente a partir da década de 80 que o termo passou a ser empregado para consolidar um conjunto de processos e algoritmos que tinham um mesmo objetivo:

[...] a extração automatizada de padrões que representem algum conhecimento implícito que esteja armazenado em grandes volumes de dados, sejam eles armazenados em banco de dados, data warehouses, na web ou qualquer outro tipo de repositório massivo ou fluxo de dados. (HAN; KAMBER; PEI, 2011, p. 23 do prefácio, tradução nossa).

Fayyad, Piatetsky-Shapiro e Smyth (1996) consideram que a mineração de dados é parte de um processo maior denominado *Knowledge Discovery on Databases* (KDD), que envolve a coleta, a limpeza dos dados, o processamento em si e a análise e visualização dos resultados encontrados. Entretanto considera-se aqui que a escolha do termo KDD é equivocada por diversas razões. Como primeiro ponto, se utiliza mineração de dados em repositórios que não estão necessariamente em banco de dados (*databases*), podendo estar em coleção de textos, em imagens e até mesmo em fluxos de dados enviados em tempo real.

Mesmo se o termo *databases* fosse substituído por *data*, alguns autores ressaltam que o processo de pré-processamento e limpeza de dados é intrínseco ao processo de mineração, ou seja, não existe mineração sem modelagem e pré-processamento dos dados (HAND; MANNILA; SMYTH, 2001). E, por fim, sob o ponto de vista epistemológico, o termo KDD é muito amplo e pode se confundir com a própria definição do que é Ciência. Afinal, o que é a ciência senão a busca de conhecimento através de dados empíricos coletados?

Vale notar que a obra de Han e Kamber vai mudando a sua percepção sobre os termos KDD e MD ao longo de suas três edições. Em sua primeira edição (2001), os autores seguem a definição de Fayyad. Já na segunda edição (2006), eles já concordam que o termo *Data Mining* é considerado pela academia e pelo mercado como o processo integral e por fim, na terceira edição (2011), citam apenas a perspectiva de Fayyad como uma forma de ver o processo, mas que a grande maioria dos autores já colocam os dois termos como sinônimos, sendo que a preferência é pelo termo *Data Mining*.

O conceito de KDD tampouco é compatível com o modelo teórico proposto no início deste ensaio já que conhecimento não pode ser “descoberto”, ele é transmitido e adquirido pelo ser humano através de processos de aprendizado que são alheios a técnicas de MD. Assim, o presente texto segue o entendimento de outros autores (COENEN, 2011; HAN; KAMBER; PEI, 2011; TRYBULA, 1999) e utiliza a expressão “Mineração de Dados” como mais apropriada que KDD.

O estudo dos algoritmos e usos de mineração de dados se originou na Ciência da Computação e recentemente passou a fazer parte de um novo campo se convencionou chamar de Ciência de Dados (*Data Science*). Segundo Carvalho (2016), a Ciência de Dados

[...] estuda princípios, métodos e sistemas computacionais capazes de extrair de forma eficiente conhecimento novo, útil e relevante presente em conjuntos de dados. Para isso, ela faz uso de técnicas de mineração de dados, particularmente de construção automática de modelos, capazes de extrair esse conhecimento. A construção automática de modelos permite que funções, hipóteses e regras sejam extraídas a partir de experiências passadas, representadas no conjunto de dados (CARVALHO, 2016, p.63)

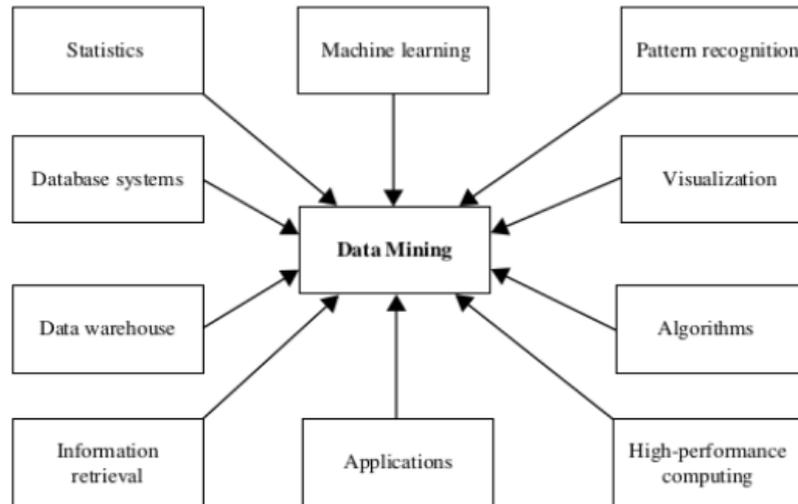
Na prática, a chamada Ciência de Dados vem sendo disputada também pelos cursos de CI, Ciências Estatísticas e Administração. No exterior, a maior parte dos cursos que tem no título a expressão “*Data Science*” estão sendo organizados pelos departamentos de Ciência da Computação⁵. Já no Brasil, o primeiro curso de graduação de Ciência de Dados estará sob a égide do Departamento de Matemática Aplicada e Estatística da USP/São Carlos⁶. Assim, seria de bom tom que o termo Ciência de Dados ainda fosse utilizado com cautela já que existe muito ruído entre o que está no âmbito da ciência e o que é apenas marketing ou *hype*.

Outro termo que também merece desambiguação é o “Aprendizado por máquina” (*Machine Learning*) que algumas vezes é confundido com o processo de MD (LESKOVEC; RAJARAMAN; ULLMAN, 2014, p.20). De fato, a mineração de dados frequentemente utiliza algoritmos de aprendizado por máquina para resolver determinadas etapas do processo e nem todo algoritmo de aprendizado por máquina é utilizado no processo de mineração de dados. Resumidamente, o aprendizado por máquina é uma das tecnologias utilizadas na mineração de dados (HAN; KAMBER; PEI, 2011). A figura abaixo mostra melhor essa relação:

Figura 1: Tecnologias utilizadas pela Mineração de Dados

⁵ A lista dos cursos pode ser vista em <http://datascience.community/colleges>.

⁶ A ementa do curso pode ser vista em <https://www.icmc.usp.br/graduacao/estatistica-bacharelado>



Fonte: Han, Kamber e Pei, 2011

As diversas estratégias de mineração de dados podem ser categorizadas em dois grandes tipos em função dos processos que são empregados: descoberta supervisionada e não supervisionada.

A descoberta supervisionada requer inicialmente que seja realizada uma etapa de treinamento, na qual parte dos objetos será pré-classificados a priori. Seleciona-se um subconjunto de objetos que serão estudados e associa-se um termo único que identifique a classe à qual o item pertence. Dessa forma, o algoritmo terá uma base inicial de “conhecimento” e a partir daí o algoritmo “constrói” um modelo computacional se baseando no conjunto de objetos iniciais. Por fim, quando um novo elemento (do qual não se sabe a qual classe pertence) for inserido no sistema, o algoritmo irá indicar a qual grupo ele provavelmente irá pertencer.

Para tornar o método mais claro, vale um exemplo na ciência política. Seleciona-se um grupo aleatório de parlamentares e cada um recebe um rótulo indicando se é de esquerda, de direita ou de centro. Ou seja, o ser humano informa ao algoritmo qual o grupo que cada parlamentar faz parte e indica também suas inúmeras características, como sexo, profissão, idade, partido político, votações anteriores etc. A partir destes dados, o algoritmo constrói um modelo e quando um novo candidato for inserido no sistema, o algoritmo indicará a qual grupo ele pertence.

Modelos supervisionados geralmente têm o objetivo de previsão. E no caso do uso de mineração de dados na pesquisa científica, estes algoritmos são muito úteis na fase de geração de hipóteses. Um estudo que tenha como objetivo prever ao invés de explicar pode

revelar associações entre variáveis que formam a base para o desenvolvimento de teorias e que podem ser posteriormente submetidas a outros tipos de testes empíricos mais rigorosos (SHMUELI, 2010).

Já as técnicas ditas não supervisionadas procuram categorizar os objetos sem que se saiba de antemão quantas classes existem. Retomando o exemplo dos políticos, imagine que temos apenas as características dos políticos (dados de entrada) e nenhuma categorização aparente (dado de saída), ou seja, não temos um modelo ou categorização prévia para aplicar ao nosso conjunto de dados. Como podemos aprender mais sobre esse conjunto de dados? No nosso exemplo, qual seria a melhor forma de dividir os políticos em categorias? Será que existe apenas esquerda, direita e centro, ou há alguma outra categorização que permita distinguir os políticos, mas que não esteja aparente? O nome desse processo de classificação é chamado de segmentação (*clustering*) e o modelo teórico que fundamenta esse processamento é baseado no princípio de maximizar as similaridades intra-classes e minimizar a similaridade inter-classes (HAN; KAMBER, 2001).

Ainda dentro dos algoritmos não supervisionados, destaca-se a detecção de anomalias (*outliers*) ou, em jargão popular, os pontos fora da curva. A detecção de anomalias e os processos de segmentação são tarefas altamente relacionadas. O agrupamento encontra os padrões em um conjunto de dados e os organiza em função da similaridade de suas características, enquanto a detecção de *outliers* tenta capturar os casos excepcionais que se desviam substancialmente dos padrões da maioria (HAN; KAMBER; PEI, 2011).

É importante também ressaltar um outro recurso que pode ser utilizado tanto nas estratégias supervisionadas quanto nas não-supervisionadas: a visualização de dados. A construção de visualizações planas ou espaciais dos dados procura aproveitar as capacidades cognitivas dos seres humanos para favorecer percepções visuais sobre os dados que muitas vezes os algoritmos de mineração de dados não conseguem fornecer.

A tecnologia digital implementou novas intermediações da informação por meio de formas de ver até o momento incomuns ao sujeito informacional ordinário. Tais práticas têm ressignificado a pesquisa em áreas como as ciências humanas, sociais e sociais aplicadas. A forma como a informação é produzida e divulgada, compartilhada, consumida e sobretudo apreendida pelas subjetividades de quem lhe acessa ou que por ela é (re)classificado, precisa

ser indagado à luz de seu regime sociotécnico de informação. Este é, sobretudo, um regime marcado nos dias atuais por uma “visibilidade informacional” (PIMENTA, 2017).

Entretanto, o processo de montar visualizações em grandes bases não é trivial, pois é necessário encontrar uma visão ou projeção que reduza a complexidade (o número de atributos de cada objeto) enquanto captura informações importantes. O objetivo é reduzir a complexidade perdendo a menor quantidade de informação possível (FAYYAD, 2001).

Também é possível categorizar os algoritmos de mineração de dados em função dos tipos de suportes onde o dado está representado. Além de minerações em tabelas relacionais, utiliza-se a mineração de dados em textos, na web, em imagens e vídeos e em grafos.

A Mineração de Textos (*Text Mining*) procura categorizar ou encontrar padrões em grandes coleções de documentos como notícias de jornal, páginas na Web, artigos de revistas científicas e obras de um determinado autor ou período (PINTO, 2007). Também é muito utilizado na mineração de opiniões ou análise de sentimentos (SOUZA; CAFÉ, 2018), quando se procura encontrar padrões em conversações realizadas nas redes sociais ou formulários online sobre algum tema específico.

A mineração na Web é uma categoria especial de mineração de texto pois necessita de um passo extra (denominado *Web Scraping* ou *Crawling*) para transformar os dados não estruturados de uma página web em um dado estruturado que possa ser minerado. Esse passo extra é quase sempre manual (realizado por humanos) se a página web não contém recursos semânticos como RDF⁷ ou outros tipos de metadados. Ressalta-se que o simples processo de captura de dados na web não pode ser chamado de MD, ou seja, se a captura não vem acompanhada de um processo de mineração em si, ela é apenas uma captura.

Ainda na área de mineração de texto, uma técnica que vem sendo muito estudada na CI é a Modelagem de Tópicos (*Topic Modeling*). Muitas vezes há a necessidade de categorizar coleções de documentos por assunto e para tal, existem método de classificação não-supervisionada, semelhante ao agrupamento em dados numéricos, que procura descobrir em qual categoria (tópico) cada um dos elementos pertence (NAVARRO; CONEGLIAN; SEGUNDO, 2018).

⁷ RDF ou *Resource Description Framework* é uma especificação utilizada para modelagem e descrição formal de metadados para que possa ser lida por um sistema computacional.

Já a Mineração de Imagens analisa imagens ou vídeos basicamente para categorizá-las a partir de critérios pré-definidos ou para se encontrar padrões específicos em um grande volume de imagens, como, por exemplo, imagens de satélite. Aqui entram técnicas complexas de aprendizado de máquina, redes neurais (*neural networks*).

Por fim, a Mineração de Grafos é o estudo de grafos gerados a partir do relacionamento entre entidades ou pessoas, como o relacionamento de pessoas nas redes sociais, operações financeiras ou citações bibliográficas. Apesar do uso de grafos na análise bibliométrica já ser uma prática comum na CI, novas técnicas vêm sendo empregadas na área da Economia Política, para encontrar, por exemplo, relações entre financiadores de campanhas eleitorais e empresas envolvidas em processos criminais.

Em síntese, ao analisar cada uma das tipologias e estratégias de mineração de dados, percebe-se um elemento comum: a classificação. O objetivo final recai sempre no classificar, ou seja, na busca dentro do universo pesquisado de um padrão que permita unir (ou segregar) os entes, objetos ou fenômenos analisados. Pode-se dizer que ao final de cada processo de mineração, emerge pelo menos um novo dado, um novo atributo para cada um dos entes do conjunto. A metáfora que pode ser feita é comparar o processo de mineração tradicional (no mundo material) que objetiva encontrar um metal oculto em um grande volume de rocha bruta e, no caso da mineração de dados (no mundo informacional), o objetivo é encontrar um dado oculto em um grande volume de dados.

E as classificações que emergem do processo de mineração nem sempre são triviais e quanto mais complexas forem, mais necessitarão de sistemas que organizem os resultados obtidos no processo. A partir deste raciocínio, pode-se intuir que o uso do conhecimento teórico acumulado na área de Organização do Conhecimento pode trazer benefícios também para essa etapa de análise dos resultados obtidos no processo de MD.

Em 1995, ao discorrer sobre a interdisciplinaridade entre Ciência da Informação e Ciência da Computação, Saracevic afirmou que a Ciência da Computação era o braço tecnológico da Ciência da Informação. Seria a Ciência da Informação o braço de construção semântica para a Ciência da Computação e para a Ciência de Dados?

4 PESQUISA BIBLIOGRÁFICA

O uso de técnicas de mineração de dados tem aparecido com maior frequência em artigos nas ciências sociais e nas chamadas humanidades. Só a revista científica especializada em Mineração de Dados para Humanidades Digitais *The Journal of Data Mining & Digital Humanities* já conta com mais de 58 artigos publicados. Em pesquisa bibliográfica preliminar na base *Web of Science* utilizando o termo “*Data Mining*”, foram encontrados 79 artigos só na área de Sociologia e Ciência Política.

Para tentar mapear de forma ampla o que se tem estudado no Brasil sobre MD na CI, foram realizadas buscas com os termos “*mineração*”, “*mining*”, “*machine*” na BDTD (Biblioteca Digital de Teses e Dissertações), na BRAPCI (que agrega parte das revistas específicas de Ciência da Informação) e na BENANCIB (que agrega todos os trabalhos apresentados no ENANCIB) sendo que na BDTD foi aplicado um filtro extra para trazer registros apenas na área das Ciências Sociais e nas 3 bases, aplicou-se um filtro extra para trazer documentos somente a partir de 2010 (inclusive).

A partir dos resultados iniciais, fez-se uma análise de cada um dos artigos para verificar qual tipo de técnica ou algoritmo de MD foi aplicado ou se o artigo relatava uma análise teórica ou descritiva sobre o tema. O objetivo desta análise era entender quais as técnicas que eram mais utilizadas na CI. A busca pelo banco de teses e dissertações trouxe 69 teses na Grande Área de Ciências Sociais Aplicadas sendo 24 na área de Ciência da Informação (8 teses e 13 dissertações). Deste total, 10 trabalhos foram sobre Mineração de Textos, ou seja, quase a metade. A segunda maior frequência foi a de análises descritivas sobre o processo de mineração de dados e a sua interdisciplinaridade com a CI.

Já na base BRAPCI, até Maio de 2019, foram encontrados 60 artigos⁸ entretanto 20 foram descartados pois continham os termos de busca no assunto, título ou palavra-chave mas efetivamente não faziam uso de nenhuma técnica de MD ou tampouco faziam alguma análise crítica sobre o uso de tais técnicas. O assunto que predominou foi também o de Mineração de Texto tal como podemos observar na tabela abaixo:

⁸ A relação completa dos artigos pode ser acessada em: <https://doi.org/10.5281/zenodo.3470020>

Tabela 1: Total de Artigos na BRAPCI de 2010 a 2019/05 sobre Mineração de Dados

Categoria	Número de Artigos
Análise de Sentimentos	2
Árvores de decisão	1
Classificação	2
Descritivo	11
Regras de associação	4
Segmentação	2
Mineração de Texto	12
Modelagem de Tópicos	3
Visualização	3
Total de Artigos	40

Fonte: Os autores

Por fim, na base do BENANCIB foram encontrados apenas 15 artigos sendo que em 2018 foram publicados 7 artigos, ou seja, quase metade do total de artigos analisados dos 8 anos anteriores. Da mesma forma que nas outras bases a preponderância se deu nas técnicas de Mineração de Texto: 7 dos 15 artigos utilizaram tais técnicas em suas pesquisas.

5 CONSIDERAÇÕES FINAIS

A partir de uma perspectiva epistemológica para a tríade dado - informação - conhecimento, buscou-se construir uma proposta de compreensão da inter-relação possível entre MD e a CI, em especial na área de organização do conhecimento. Sabemos que há certamente a necessidade de uma agenda ou programa de pesquisa que possa dar conta de todas as singularidades possíveis dessa articulação, mas para o momento nossa proposição se baseia em uma perspectiva inicial, de construção de um olhar sobre a temática.

As reflexões trazidas neste ensaio mostram que o uso de sistemas de organização do conhecimento pode contribuir em grande medida no desenvolvimento das chamadas Ciências de Dados e nos usos de MD e IA. A utilização de vocabulários controlados, taxonomia e ontologias é peça essencial para toda a infraestrutura semântica necessária para a construção de sistemas autônomos e de algoritmos de aprendizado de máquina. Quanto mais organizados os metadados estiverem, quando maior for a capacidade de transformar metadados implícitos (contexto) em metadados explícitos, maior será a capacidade de tais sistemas aprenderem e maior será a capacidade e acurácia destes mesmos sistemas.

Percebe-se também que a infraestrutura para a construção da Ciência de Dados e de processos aptos a minerar e trabalhar com o chamado *Big Data* depende de materializar (ou digitalizar) o contexto dos problemas por eles enfrentado pois, efetivamente, as máquinas e os algoritmos só podem tratar dados digitalizados. Desta forma, um investimento maior na interdisciplinaridade entre a CI e Ciência de Dados poderá trazer benefício para a solução de questões de pesquisa nas duas áreas.

Este estudo também mostrou que existem poucas pesquisas em CI no Brasil que fazem uso de MD e de aprendizado de máquina tanto para resolver questões metodológicas da própria CI como também de pesquisas que estejam avaliando o impacto destas tecnologias na sociedade. Uma hipótese a ser estudada em trabalhos futuros recai sobre a escassez de disciplinas de computação nos programas de pós-graduação de CI no Brasil e se existe um real interesse dos estudantes e pesquisadores do campo em trabalhar nestes novos campos científicos. O crescimento do número de artigos no Enancib 2018 e a escolha do tema do encontro em 2019 podem ser indicativos que a comunidade acadêmica passou a se interessar mais pelo assunto.

REFERÊNCIAS

BERNERS-LEE, T.; HENDLER, J.; LASSILA, O. The Semantic Web. **Scientific American**, v. 284, n. 5, p. 34–43, 2001.

BEZERRA, Arthur Coelho. Vigilância e filtragem de conteúdo nas redes digitais: desafios para a competência crítica em informação. In: **Encontro Nacional de Pesquisa em Ciência da Informação**, 16, 2015, João Pessoa. Anais... João Pessoa: UFPB, 2015.

BORGMAN, C. L. **Big data, little data, no data: scholarship in the networked world**. MIT Press. Cambridge. 2016.

BRÄSCHER, M. Semantic Relations in Knowledge Organization Systems. **Knowledge Organization** v. 41, n. 412, 2014.

BUCKLAND, M. K. Information as Thing. **Journal of the American Society of Information Science** v.42, n.5, p.351-360. 1991.

CAMPOS, M.L.A. Teoria do conceito. In: _____. **Linguagem documentária: teorias que fundamentam sua elaboração**. Niterói, Editora da Universidade Federal Fluminense, 2001.

CALDAS, C. O. L.; CALDAS, P. N. L. Estado, democracia e tecnologia: conflitos políticos no contexto do big-data, das fake news e das shit storms. **Perspectivas em Ciência da Informação**, v. 24, n. 2, p. 196–220, 2019.

CAPURRO, R.; HJØRLAND, B. O conceito de informação. **Perspectivas em Ciência da Informação**, v.12, n.1, p.148-207, 2007.

CAPURRO, R. **Knowledge Map of Information Science: Rafael Capurro's Responses to Chaim Zins**. Disponível em: <<http://www.capurro.de/zins.html>>. Acesso em: 5 ago. 2017.

CARVALHO, A.C.F. Interdisciplinaridade da Ciência de Dados. **Computação Brasil**. v.31, p.62-65. 2016.

CLEARY, J., Holmes, G., Cunningham, S.J., and Witten, I.H. MetaData for database mining. **Proceedings of the First IEEE Metadata Conference**. Silver Spring, MD, USA. 1996.

DAHLBERG, Ingetraut. Teoria do conceito. **Ciência da Informação** v. 7, n. 2, p. 101-07, 1978.

FROHMANN, B. **Deflating information: From science studies to documentation**. Toronto, Canada: University of Toronto Press, 2004.

FURNER, Jonathan. "Data": The data. In: KELLY, Matthew; BIELBY, Jared (Orgs.). **Information Cultures in the Digital Age: A Festschrift in Honor of Rafael Capurro**. Wiesbaden: Springer Fachmedien Wiesbaden, 2016, p. 287–306. Disponível em: <https://doi.org/10.1007/978-3-658-14681-8_17>. Acesso em: 14 mar. 2019.

GIL, A. C. **Métodos e Técnicas de Pesquisa Social**.6. ed. São Paulo: Atlas, 2008.

GOBBLE, M. M. Digitalization, Digitization, and Innovation. **Research-Technology Management**, v. 61, n. 4, p. 56–59, 4 jul. 2018.

HAN, J.; KAMBER, M. **Data Mining: Concepts and Technics**. Londres: Academic Press, 2001.

HAN, J.; KAMBER, M. **Data Mining: Concepts and Techniques**. San Francisco: Morgan Kaufmann Publishers. 2nd Edition. 2006.

HAN, J.; KAMBER, M.; PEI, J. **Data Mining: Concepts and Techniques**. San Francisco: Morgan Kaufmann Publishers. 3rd edition. 2011.

HAND, D. J.; Mannila H.; Smyth P. **Principles of Data Mining (Adaptive Computation and Machine Learning)**. Cambridge, MA: MIT Press, 2001.

HJØRLAND, Birger. Data (with big data and database semantics). **Knowledge Organization**45, no. 8: 685-708. ISKO Encyclopedia of Knowledge Organization, ed. Birger Hjørland, coed. Claudio Gnoli. Disponível em <<http://www.isko.org/cyclo/data>>. Acesso em 05 jul. 2018.

LESKOVEC, J., RAJARAMAN, A., ULLMAN, J. D. **Mining of massive datasets**. Cambridge University Press. 2014.

LUKÁCS, G. História e consciência de classe: estudos de dialética marxista. Trad. Telma Costa; Revisão Manuel A. Resende e Carlos Cruz – 2º Edição, Rio de Janeiro: Elfos Ed.; Porto, Portugal, Publicações Escorpião, 1989.

MANNING, C. D. Computational Linguistics and Deep Learning. **Computational Linguistics**, v. 41, n. 4, p. 701–707, dez. 2015.

MICHALAKIDIS, G. **Appreciation of structured and unstructured content to aid decision making - from web scraping to ontologies and data dictionaries in healthcare.** Tese de Doutorado. University of Surrey. 2016.

NAVARRO, F. P.; CONEGLIAN, C. S.; SEGUNDO, J. E. S. Big Data no contexto de dados acadêmicos: o uso de Machine learning na construção de Sistema de Organização do Conhecimento. In: XIX Encontro Nacional De Pesquisa Em Ciência Da Informação, 2018, Londrina. **Anais..** Londrina, 2018.

PIMENTA, R. M. Big data e controle da informação na era digital: tecnogênese de uma memória a serviço do mercado e do estado. **Tendências da Pesquisa Brasileira em Ciência da Informação**, v. 6, n. 2, 2013.

PIMENTA, R. M. Ciberespaço, internet e habitus: uma reflexão bourdieusiana sobre a era digital. In MARTELETO, Regina M., PIMENTA, Ricardo M. (orgs.) **Pierre Bourdieu e a produção social da cultura, do conhecimento e da informação.** Rio de Janeiro: Garamond, 2017.

PINTO, André Moreira. **Mineração de textos e gestão do conhecimento: aplicação na experiência operacional em geração de energia nuclear nas Usinas de Angra I e II.** 369 f. Tese (Doutorado) –Escola de Ciências da Informação, UFMG, 2007.

POMERANTZ, Jeffrey. **Metadata.** Cambridge, Massachusetts ; London, England: The MIT Press, 2015.

ROUVROY, A. BERNIS, T. Gouvernamentalité algorithmique et perspectives d'émancipation: Le disparate comme condition d'individuation par la relation? **Réseaux**, v. 177, n. 1, p. 163, 2013.

SARACEVIC, T. Interdisciplinary nature of information science. **Ciência da Informação**, v. 24, n. 1, p. 36-41, 1995.

_____. The concept of "Relevance" in information science: a historical review. In: **Introduction to information science.** New York, Bowker, 1970.

_____. Relevance: a review of and a framework for the thinking on the notion in information science. **JASIST**, v.26 n.6 p.321, Nov. 1975.

SEJNOWSKI, T. J. **The deep learning revolution.** Cambridge, Massachusetts: The MIT Press, 2018.

SEMELER, A. R.; PINTO, A. L. Os diferentes conceitos de dados de pesquisa na abordagem da biblioteconomia de dados. **Ciência da Informação**, v. 48, n. 1, 3 maio 2019.

SHMUELI, G. To Explain or to Predict? **Statistical Science**, v. 25, n. 3, p. 289–310, ago. 2010.

SOUZA, Renato Rocha; ALMEIDA, Maurício Barcellos; BARACHO, Renata Maria Abrantes. Ciência da Informação em transformação: Big Data, Nuvens, Redes Sociais e Web Semântica. **Ciência da Informação**, v. 42, n. 2, 2013.

SOUZA, Renato Rocha; CAFÉ, L. M. A. Análise de sentimento aplicada ao estudo de letras de música. **Informação & Sociedade: Estudos**, v. 28, n. 3, 28 dez. 2018.

STEINER, Tobias. Metadaten und OER: Geschichte einer Beziehung. **Synergie: Fachmagazin für Digitalisierung in der Lehre**. 4, S. 51-55. 2017.

XX ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO – ENANCIB 2019
21 a 25 de outubro de 2019 – Florianópolis – SC

VIANA, B. Journalism in the context of the Semantic Web. **Media & Jornalismo**, v. 18, n. 32, p. 187–199, 2018.

VOSTAL, F. **Accelerating Academia: The Changing Structure of Academic Time**. Springer, 2016.