Big Data no contexto de dados acadêmicos: o uso de *machine learning* na construção de sistema de organização do conhecimento

Fabio Piola Navarro
Universidade Estadual Paulista Júlio Mesquita Filho – UNESP, E-mail: navarro@univem.edu.br

Caio Saraiva Coneglian
Universidade Estadual Paulista Júlio Mesquita Filho – UNESP, E-mail: caio.coneglian@gmail.com

José Eduardo Santarem Segundo
Universidade de São Paulo – USP, E-mail: santarem@usp.br

RESUMO

A quantidade de dados no âmbito acadêmico tem aumentado significativamente, sendo causado principalmente pelo aumento da quantidade de documentos científicos. Neste sentido, há uma poderosa fonte de informação, enquadrada no contexto do *Big Scholarly Data*, que pode ser utilizada para análise e o aprimoramento das tradicionais técnicas da Ciência da Informação, ao fornecer um conhecimento relevante para a construção de Sistemas de Organização do Conhecimento que reflitam com mais veracidade dados contidos nestes ambientes. Adicionalmente, uma forma de realizar o tratamento e a análise desse conglomerado de dados é relativo ao *machine learning*. No entanto, é necessário refletir como as técnicas de aprendizado de máquina quando aplicado a grandes volumes de dados podem favorecer a construção de sistemas de organização de conhecimento. Assim, o objetivo deste trabalho é discutir a aprendizagem de máquinas, em especial o método *topic modeling*, no processo de construção de sistemas de organização do conhecimento, visando discutir como dados pertencentes ao domínio do BSD podem fornecer informações necessárias para aprimorar os instrumentos de organização do conhecimento. Para isso, utilizou-se de uma metodologia com característica tanto bibliográfica, quanto aplicada. Enquanto resultados, primeiramente, foi proposto um modelo teórico, que vincula os dados de repositórios digitais, com as técnicas de *machine learning*, para a construção de sistemas de organização do conhecimento. Para validar esse modelo, realizou-se uma prova de conceito que utilizou o método *topic modeling* e o LDA para identificar tópicos de interesse dentro de um corpus científico de um repositório digital no intuito de organizar, sumarizar e entender seu conteúdo, visando fornecer uma base para a construção de sistemas de organização do conhecimento. Por fim, concluiu-se que a aplicação das técnicas de *machine learning* contribui para fornecer um panorama amplo sobre os dados contidos em repositórios digitais, sendo uma base bastante rica para apoiar e aprimorar a construção de sistemas de organização do conhecimento.

Palavras-chave: Machine learning. Repositórios digitais. Topic modeling. Sistema de organização do conhecimento.

---

1 INTRODUÇÃO

Com a temática "Sujeito informacional e as perspectivas atuais em Ciência da Informação", o XIX ENANCIB propõe uma análise de como a Ciência da Informação deve ter papel central no cenário atual científico e tecnológico, em que os sujeitos informacionais estão cada vez mais se tornando miméticos, ou seja, nos ambientes informacionais digitais, os produtores e os consumidores de informação se fundem e se confundem, trazendo complexidade à área e uma consequente necessidade de estudo.

Segundo Araújo (2013), o estudo dos sujeitos informacionais tem crescido significativamente nos últimos vinte anos, de forma mais específica, as perspectivas de pesquisa que mais cresceram foram aquelas que buscaram integrar o caráter individual e coletivo do comportamento dos usuários nos mais variados contextos.

Neste sentido, a busca por informação é recorrente no ambiente acadêmico, e normalmente esta informação, sob a forma de documentos em formatos digitais, armazenados em repositórios digitais, muitas vezes não possuem formas holísticas de extração de informação. Entende-se por holística, neste contexto, como uma forma que se tenha um entendimento do todo em relação aos repositórios digitais, não somente na recuperação de cada um dos arquivos e análises singulares.

Para atingir essa forma holística de se extrair, e de compreender cada documento, a Ciência da Informação possui uma série de ferramentais, em destaque os Sistemas de Organização do Conhecimento. Carlan, Carlan e Medeiros (2012) destacam que os sistemas de organização do conhecimento são sistemas conceituais semanticamente estruturados que possuem termos, definições, relacionamentos e propriedades relacionados aos conceitos e que possuem em sua forma estrutural desde um esquema simples até estruturas multidimensionais que auxiliam na recuperação da informação.

Uma forma de se ter este acesso mais abrangente sobre o conteúdo de um repositório digital, muitas vezes dotados de um grande volume de dados, é por meio do empenho humano em classificar e organizar arquivos por assuntos, o que pressupõe especialistas sobre vários assuntos e, além disso, algum recurso computacional para análise e recuperação de informação de maneira mais assertiva.

Este cenário ao ser auxiliado por recursos computacionais específicos do contexto da inteligência artificial ganha uma capacidade de propiciar análises holísticas de qualidade sobre repositórios digitais. Um modelo que pode auxiliar significativamente nessas análises é o *topic modeling*, um método de análise automática de textos conhecido. Este método explora e tem a capacidade de sumarizar uma coleção de documentos sem necessariamente ter conhecimento prévio do contexto sobre o corpus analisado. (BOYD-GRABER; HU; MIMNO, 2017).

A necessidade dessas técnicas de aprendizado de máquina e inteligência artificial se fazem mais necessárias quando se reflete sobre a perspectiva da grande quantidade de dados disponíveis nos ambientes digitais. Esse cenário conhecido como *Big Data* apresenta como característica além do volume dos dados, a variedade, em que dados estruturados e não estruturados em variados formatos estão disponíveis, e a velocidade com que eles são criados e modificados (3 V's – Volume, Variedade e Velocidade).

Quando se trabalha com dados no contexto do *Big Data* em cenário acadêmico, tem-se o *Big Scholarly Data* (BSD) como campo de estudo. O BSD é o termo criado para abarcar e analisar este novo fenômeno na área da ciência em relação à pesquisa científica, ou seja, a grande quantidade de dados das mais variadas formas dentro de um ambiente acadêmico de pesquisa (XIA et al., 2017). O BSD trata especialmente de informações sobre: temas de pesquisa, autores, citações, figuras, tabelas além de todo arcabouço colaborativo entre as partes.

Esse cenário de uma grande quantidade de dados acadêmicos, em que os repositórios digitais são uma das principais fontes, traz grandes desafios para a Ciência da Informação, quando se reflete nos sistemas de organização do conhecimento, e como tais sistemas podem refletir efetivamente os cenários encontrados. Assim, o presente trabalho apresenta como questão de pesquisa: Como técnicas de aprendizado de máquina quando aplicado a grandes volumes de dados podem favorecer a construção de sistemas de organização de conhecimento?

Desta forma, o objetivo deste trabalho é discutir a aprendizagem de máquinas, em especial o método *topic modeling*, no processo de construção de sistemas de organização do conhecimento. Busca-se ainda, discutir como os dados pertencentes ao domínio do BSD podem fornecer informações necessárias para aprimorar os instrumentos de organização do conhecimento.

Salienta-se que este estudo é parte de um conjunto de estudos, em que a extração de tópicos em corpus textuais é parte do processo, visando um objetivo final de se desenvolver modelos sociais complexos que, auxiliado pelo poder computacional e intensiva coleta de dados possibilite a análise de problemas complexos por meio de sistemas de organização do conhecimento mais assertivos com o uso de tecnologias aplicadas.

Para o desenvolvimento deste trabalho, aplicou-se uma metodologia exploratória e aplicada. Primeiramente, realizou-se uma pesquisa bibliográfica e documental sobre as temáticas de *Big Scholarly Data*, Sistemas de Organização do Conhecimento, Recuperação da Informação, Aprendizado de Máquinas e *topic modeling*. Posteriormente, realizou-se a discussão em que se vincula e reflete-se acerca das relações dessas temáticas, apresentando uma pesquisa aplicada, com uma análise documental para extração dos documentos (421 artigos no total) com a técnica de *web scraping* e a construção de um ambiente computacional que possibilite carregar estes artigos e proporcione capacidade de análise, usando o método *topic modeling* a fim de comprovar a capacidade latente dos repositórios digitais em proporcionar informações para análise e a construção de Sistemas de Organização do Conhecimento.

2 REFERENCIAL TEÓRICO

O referencial teórico desse trabalho está dividido em quatro partes: *Big Scholarly Data*, Sistemas de Organização do Conhecimento, Recuperação da Informação e Aprendizado de Máquina e *Topic Modeling*. Esse referencial teórico é essencial para realizar as discussões e a pesquisa aplicada que será demonstrada na seção de resultados e discussões.

2.1 *BIG SCHOLARLY DATA*

O conceito de *big data* já está bem difundido nos meios científicos, o qual contempla a rápida aquisição de dados, o alto de volume destes dados e uma grande variedade destes dados. No entanto, tem havido uma maior ramificação da granularidade dos domínios destes dados, ou seja, cada domínio tem seu próprio conjunto de dados, armazenados em diferentes tipos de arcabouços computacionais, e um destes cenários são os repositórios de documentos, foco deste trabalho.

Segundo Khan et al. (2017) um destes cenários é o de dados científicos, ao qual foi dado o nome de *Big Scholarly Data* e que segundo o autor recebe menos atenção no que tange a aplicações e resultados para este domínio no qual os dados estão armazenados e que carece de estudos que explorem cada vez mais este tipo de cenário específico, mas que ainda sim herda todos os conceitos e características de *big data*.

Neste sentido, os autores relatam que:

> A necessidade de pesquisa em "*Big Scholarly Data*" e suas análises pode ser resumida como a falta de plataformas acadêmicas e ferramentas que podem usar essa imensa base de dados para criar aplicativos que podem beneficiar a comunidade acadêmica em geral. O gerenciamento efetivo e eficiente de grandes dados acadêmicos usando a infraestrutura de nuvem pode facilitar os processos envolvidos na análise de *Big Data*, como aquisição, armazenamento, processamento, análise e visualização de dados para suportar o gerenciamento de dados de pesquisa e seus usos analíticos. Documentos acadêmicos são gerados diariamente na forma de trabalhos de pesquisa, propostas de projetos, relatórios técnicos e documentos acadêmicos, por pesquisadores e estudantes de todo o mundo. [...] No entanto, é importante notar que esta é uma descrição generalizada e a definição pode variar de uma comunidade acadêmica para outra. [...] Devido ao grande volume desses recursos digitais, os dados precisam ser considerados a partir da perspectiva do *Big Data*.(KHAN et al., 2017, tradução nossa)

Baseado no excerto, existe uma necessidade latente de pesquisa e de se explorar o chamado *Big Scholarly Data*, pois há uma falta de plataformas e ferramentas acadêmicas que possam usar este enorme reservatório de dados para criar aplicações e análises que possam beneficiar a comunidade de pesquisa e por consequência toda a comunidade.

De forma genérica, o ciclo de vida do fenômeno *big data* pode ser dividido em quatro categorias: geração de dados, aquisição, armazenamento e processamento dos dados. No entanto, segundo Assunção et al. (2015) de forma mais direta e aplicada ao ciclo de vida do *big scholarly data*, o mesmo divido o ciclo em três categorias: gerenciamento de dados, um segundo ciclo chamado de análise (*analytics)* e o terceiro ciclo, como resultado para os dois primeiros, chamado de aplicações e visualizações (*applications and visualization)*.

As bases de dados pertencentes ao contexto do *Big Scholarly Data* podem contribuir na construção de Sistemas de Organização do Conhecimento que contemplem os conhecimentos

obtidos por meio de análises de dados. Na sequência, apresentam-se os conceitos chave dos sistemas de organização do conhecimento.

2.2 SISTEMAS DE ORGANIZAÇÃO DO CONHECIMENTO

A oportunidade para exploração e proposição de soluções para repositórios de documentos científicos e com isso proporcionar benefícios para a comunidade científica necessita de auxílios dos recursos computacionais como estrutura dos sistemas de organização do conhecimento que segundo (INTERNATIONAL SOCIETY FOR KNOWLEDGE ORGANIZATION.; CHAPTER, 1993) esta define organização do conhecimento como:

"A ciência que estrutura sistematicamente e organiza as unidades de conhecimento ou conceitos de acordo com seus elementos (características) e a aplica estes conceitos e classes de conceitos em uma organização conhecida como sujeito e objeto" (tradução nossa)

Segundo Bräscher (2014), linguística, filosofia, psicologia, ciência da informação e inteligência artificial são algumas das áreas que lidam com diferentes aspectos da representação do conhecimento. Neste sentido é inevitável que a interdisciplinaridade entre ciência da informação e ciência da computação aconteça de forma natural para resolução e demanda dos novos cenários ou problemas relacionados à informação.

É plausível que a partir destas definições e colocações sobre a interdisciplinaridade entre as mais variadas ciências acrescentar que sistemas de organização do conhecimento são muito importantes em muitos aspectos quando, de forma geral, pensa-se sobre a interação humano e sistemas. No entanto, atualmente esta área vem crescendo em relação a sua utilização como estrutura base para sistemas inteligentes que nem sempre tenham ou terão um ser humano como possível ator de interação neste cenário (CARLAN; CARLAN; MEDEIROS, 2012).

2.3 RECUPERAÇÃO DA INFORMAÇÃO

O termo "*information retrieval*" remonta à década de cinquenta com Calvin Mooers, quando o mesmo delimitou os problemas a serem resolvidos por esta nova disciplina.

Saracevic em 1995 explora muito bem a interdisciplinaridade entre ciência da informação e ciência da computação ao dizer que a área de recuperação da informação seria o braço tecnológico da ciência da informação.

Segundo Ferneda (2003), a recuperação da informação está baseada em dois eixos: o primeiro relacionado à preparação do corpus a ser utilizado como base para a recuperação da informação, no qual ainda se contempla a representação da expressão de busca e a função de busca.

No segundo eixo, está definida a usabilidade, na qual estão envolvidos, a expressão de busca, os resultados de busca e o próprio usuário, que a partir de sua busca, ou seja, sua necessidade de informação deve interagir com algum sistema que propicie esta interação humano computador e que resulte em documentos que atendam sua necessidade.

É neste momento que nosso trabalho difere e de certa maneira propicia uma nova abordagem na recuperação da informação, pois, normalmente o que se possui é de fato o cenário proposto por Ferneda, no entanto, atualmente temos repositórios de documentos no qual o corpus possui muitos terabytes de dados, o que de certa forma pode limitar o resultado de uma busca caso o usuário não tenha conhecimento sobre o corpus em si.

Assim, este trabalho propõe a utilização de técnicas e métodos ligados à ciência da computação e mais especificamente ligados à área da inteligência artificial para auxiliar na recuperação da informação, de forma específica, a repositórios de documentos científicos utilizando-se de tecnologias que propiciem um maior conhecimento sobre o corpus, ou seja, sobre os assuntos, sobre os domínios do corpus no qual a informação está disposta, mas muitas vezes, de forma desordenada, o que levaria a resultados de busca menos assertivos.

2.4 APRENDIZADO DE MÁQUINA, *TOPIC MODELING e LDA*

O campo do aprendizado de máquina ou ainda pelo termo corrente em inglês, *machine learning*, advém dos anos de 1950 logo após a criação do teste de Turing com os primeiros softwares que conseguiam aprender a partir de dados de entrada.

Aprendizado de máquina é um subconjunto da inteligência artificial no qual programas são usados para aprender através deles mesmos, ou seja, de forma autônoma a partir de dados e informação.

Segundo Mitchell (1997) o aprendizado de máquina trata da questão da construção de programas que automaticamente melhorem sua atuação com base na sua própria experiência, ou seja, os programas de computadores conseguem melhorar sua assertividade quanto mais são utilizados dentro de um determinado domínio de aplicação.

Este trabalho propõe a utilização de algoritmos de aprendizado de máquina para fazer buscas holísticas em bases de documentos científicos, que mesmo sem prévio conhecimento do conteúdo dos repositórios de documentos, vão permitir inferências para melhorar a assertividade sobre o conhecimento de um determinado corpus de documentos.

Uma técnica de aprendizado de máquinas é o *topic modeling*. *Topic modeling* é uma subárea da grande área de pesquisa *text mining*, método cunhado nos anos 60 com o aumento da pesquisa na área de recuperação da informação. Mas foi nos anos 90 que a área ganhou maior capacidade de expansão e de pesquisa devido ao crescimento da quantidade de documentos digitais e também pela capacidade de processamento dos computadores.

O método de *topic modeling* ou sua tradução modelo de tópicos teve seu desenvolvido em 2003 com estudos de Blei que nomeou seu primeiro estudo na área como LDA ou *Latent Dirichlet Allocation*. LDA permite análise de estruturas temáticas não estritamente expostas o que possibilita extração de informação semântica para grandes volumes de texto (BLEI, 2012).

Modelos de tópicos fazem uma busca por padrões nas relações entre documentos e termos, tais padrões, podem estar latentes e podem ser significativos para o entendimento dessas relações. Este modelo pode trazer como resultado um conjunto de termos que são importantes para um ou mais temas, ou ainda, ranquear documentos com mais relevância para um ou mais temas.

> Os tópicos são estruturas com valor semântico e que, no contexto de mineração de texto, formam grupos de palavras que frequentemente ocorrem juntas. Esses grupos de palavras quando analisados, dão indícios a um tema ou assunto que ocorre em um subconjunto de documentos. A expressão “tópico” é usada levando-se em conta que o assunto tratado em uma coleção de documentos é extraído automaticamente, ou seja, tópico é definido como um conjunto de

> palavras que frequentemente ocorrem em documentos semanticamente relacionados. (FALEIROS; LOPES, 2016, p.9).

Como pode ser analisado no excerto, o modelo de tópicos não é simplesmente um ranqueamento das palavras que mais ocorrem em um documento, mas sim uma análise de tópicos que são na verdade um conjunto de palavras que possuem maior frequência semântica e de forma relacionada dentro de cada documento.

Portanto, por hipótese a ser comprovada pelo caso exposto, a utilização de modelos de tópicos (*machine learning*) como recurso computacional de apoio para o desenvolvimento de sistemas de organização de conhecimento possibilita a recuperação da informação de maneira mais inteligente, já que em meio ao contexto de *big data* inerente de grandes repositórios de documentos científicos, este método proporciona o descobrimento de tópicos que a priori não eram conhecidos e desta maneira pode apoiar de forma assertiva a recuperação e conhecimento de dados e de informação.

Uma técnica relacionada e complementar ao *topic modeling* é o *Latent Dirichlet Allocation* (LDA). Segundo Blei (2003), o LDA é um modelo que possibilita a extração de palavras de cada documento e alocação destas palavras em tópicos sobre um corpus de documentos em repositórios digitais. Cada documento pode ser visto como uma coleção de tópicos e dado o caráter do grande volume de dados nestes repositórios digitais, este método possibilita a identificação destes tópicos, que refletem o verdadeiro conteúdo de cada um documento, assim este modelo possibilita a construção de instrumentos para sistemas de organização do conhecimento, mesmo que não haja um prévio conhecimento dos conteúdos de qualquer repositório digital.

Partindo dos conceitos discutidos no referencial teórico, a seguir apresenta-se os resultados e discussões desse trabalho, em que buscasse atingir os objetivos dessa pesquisa.

3 RESULTADOS E DISCUSSÕES

A quantidade de dados acadêmicos vem aumentando significativamente nos últimos anos, reflexo de uma expansão da ciência, especialmente em países em desenvolvimento. Na esteira dessa expansão, a publicações de acesso aberto cresceu de forma exponencial, permitindo a criação de diversas ferramentas para a disponibilização dos documentos em acesso aberto.

Atualmente, a principal tecnologia para a disponibilização de publicação intelectual de acessos aberto é o repositório digital. Um repositório é capaz de armazenar grandes quantidades de documentos (entre eles, artigos, livros, teses e dissertações), juntamente com os metadados que descrevem esses objetos.

Esse cenário demonstra a diversidade e a quantidade de informações que estão inseridas nos repositórios digitais. Neste contexto, informações tanto estruturadas quanto não estruturadas fazem parte dessa massa de dados que os repositórios digitais contém. Refletir e aplicar técnicas vinculadas ao *Big Data* se faz necessário para compreender e extrair informações valiosas desses dados.

Evidenciando a questão da quantidade de informações que os Repositórios Digitais contém, o serviço *Registry of Open Access Repositories* (ROAR)[1], que reúne informações diversas atualizadas sobre os repositórios digitais, aponta que atualmente há 4666 repositórios digitais ativos e registrados, espalhados por dezenas de países.

Esse número demonstra a relevância dos repositórios, tanto pelo alto número de implantações, quanto pela variedade de países e tipos. Assim, ao pensar na quantidade de registros que podem e estão armazenados nesses repositórios, identifica-se o volume de dados desses ambientes, bem como o potencial de obtenção de conhecimento, quando se reflete sobre análise e mineração de dados nesse contexto.

O cenário apresentado permite enquadrar os dados contidos nos repositórios como inseridos no contexto do *Big Data*. Quando reflete-se nos três "V's" que compõem o *Big Data*, Volume, Variedade e Velocidade, identifica-se que os dados dos repositórios apresentam um alto volume, muitos registros em milhares de repositórios, grande variedade, os dados dos repositórios são informações estruturadas (metadados), mas também variados documentos como textos, imagens, vídeos, *datasets* e áudio, e velocidade, pelo aumento exponencial na quantidade de registros armazenados e pela necessidade constante de permitir que os dados estejam sempre atualizados.

[1] Registry of Open Access Repositories. Disponível em: <http://roar.eprints.org/>. Disponível em: <http://roar.eprints.org/>. Acesso em: 30 jul. 2018.

Complementarmente, quando refletimos sobre *Big Data* no contexto de dados acadêmicos, surge o conceito de *Big Scholarly Data*, em que estão inseridos os dados acadêmicos de artigos, autores, citações e demais relacionados, que estão no âmbito do *Big Data*. Como relatado anteriormente, o *Big Scholarly Data* vem se tornando mais relevante pelo aumento das produções científicas, e pela necessidade de se gerar valor e extrair conhecimento dos dados acadêmicos.

Aprofundar os estudos nessa área é necessário pelos novos desafios que são dados a ciência atual, em que o tratamento e o uso de se tornou elementar e fundamental. Análises de dados no contexto do *Big Scholarly Data* podem auxiliar a compreender os cenários existentes, além de gerar conhecimentos para a realização de novas pesquisas e para melhorar como os pesquisadores compreendem o seu processo produtivo.

Esses desafios e oportunidades estão inseridos na seara da Ciência da Informação, uma vez que esta disciplina pode aprimorar a sua compreensão dos cenários de dados científicos, melhorando técnicas e ferramentas que são criadas no contexto da Ciência da Informação e utilizadas pelas diversas áreas do conhecimento. Por outro lado, a Ciência da Informação pode levar os seus processos tradicionais para essas análises de dados, inserindo questões fundamentais para que o *Big Scholarly Data* considere elementos fundamentais para o cenário dos dados científicos e acadêmicos.

Nessa mútua cooperação, os sistemas de organização de conhecimento (SOC) podem ser significativamente aprimorados. Esses sistemas que estão inseridos dentro da Ciência da Informação vêm constantemente evoluindo, com as tecnologias da Web Semântica, com a popularização das Tecnologias da Informação e Comunicação, e mais recentemente, com o uso do *Big Data*.

As análises de dados no contexto do *Big Data* podem permitir que os SOC sejam aprimorados, pois tais análises obtêm informações importantes sobre termos e relações que compõem esses sistemas. Desta forma, o contexto atual das tecnologias permite uma evolução e uma consequente melhora no modo como as informações podem ser representadas a partir dos SOC.

No entanto, há uma série de técnicas existentes para se extrair algo de valor desse volume extremo de dados. Atualmente, as técnicas de *Machine Learning*, advindas da Inteligência Artificial estão trazendo grandes contribuições para os estudos de Big Data. Isso ocorre pela necessidade de haver grandes massas de dados, para se extrair padrões e comportamentos, que podem ser obtidos e explorados a partir de implementações de técnicas de *Machine Learning*.

Como descrito, o modelo de tópicos é uma técnica de *Machine Learning*, e no âmbito deste trabalho foi o principal elemento utilizado para demonstrar como Big Data e *Machine Learning* podem contribuir para o aperfeiçoamento dos SOC.

O principal foco do uso do modelo de tópicos para auxiliar na construção dos SOC está na obtenção de termos relevantes para os cenários. Isso ocorre porque quando se aplica técnicas de *Topic Modeling* em grandes massas de dados, obtém-se um conjunto de termos relevantes e contextualizados sobre a massa de dados analisada.

Neste sentido, a aplicação de *Topic Modeling* em um repositório digital é capaz de gerar tal conjunto, no contexto de dados acadêmicos. Essa técnica é capaz de encontrar termos que refletem os conjuntos de dados, considerando aspectos de como os dados estão relacionados e considerando o repositório digital na totalidade.

Adicionalmente, um complemento ao *topic modeling* está na aplicação do LDA. A aplicação das técnicas de LDA irá permitir que os tópicos e os termos obtidos, sejam contextualizados e relacionados, permitindo assim a obtenção de um conhecimento mais abrangente dos repositórios digitais.

Sintetizando, ao aplicar as técnicas de *topic modeling* e LDA, gera-se um conjunto de termos vinculados a um determinado tópico. Tais termos são classificados de acordo com as relações semânticas existente nas bases de dados dos repositórios digitais, em que tanto o contexto dos documentos frente ao repositório quanto a relação do termo frente ao documento são considerados.

A partir desse conhecimento obtido com a aplicação de tais técnicas, é possível construir sistemas de organização do conhecimento que reflitam o conteúdo dos repositórios. Esse processo se baseia no fornecimento dos termos e do quão significativos tais termos são no âmbito do repositório, além das relações que são obtidas, entre os tópicos e os termos.

Posteriormente, termos e relações podem ser usados como fonte na construção dos Sistemas de Organização do Conhecimento, como taxonomias, tesauros e ontologias. O profissional da informação utilizaria o conhecimento obtido com as técnicas de *machine learning* para apoiar e aperfeiçoar a construção desses instrumentos.

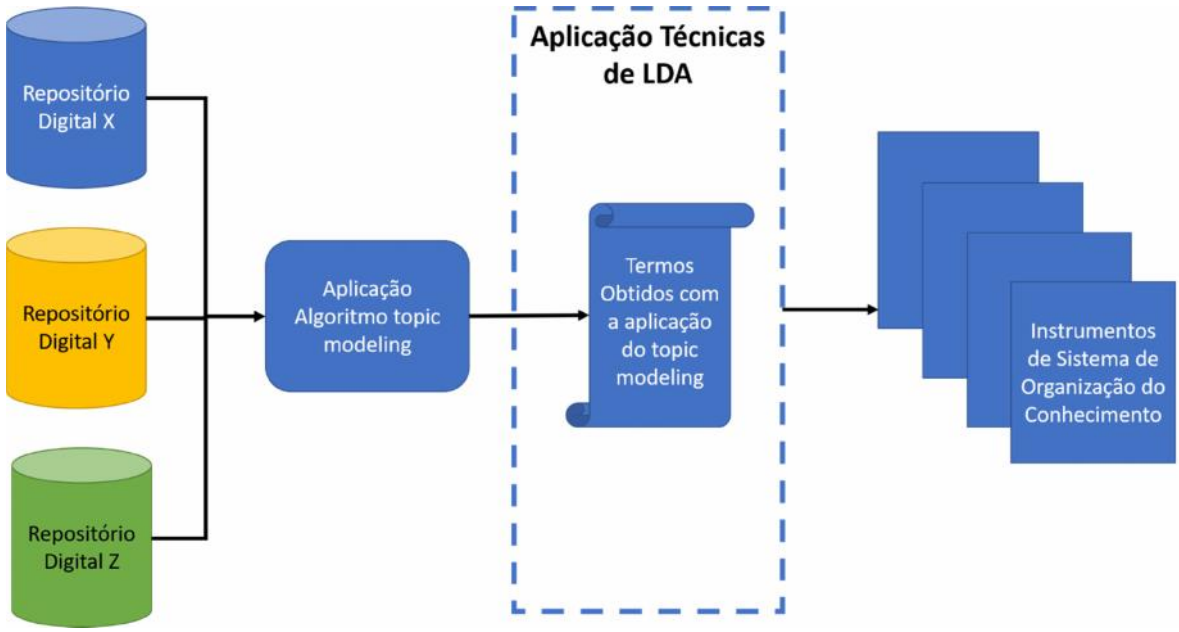
A partir da união das técnicas de *topic modeling* e LDA aplicados em conjuntos de dados obtidos a partir de repositórios digitais, é possível extrair um conhecimento acerca do conteúdo dos documentos contidos nestes repositórios, que aglomera os principais termos que representam cada documento e o ambiente como um todo. Esse conhecimento obtido pode ser apresentado por meio dos próprios termos, das relações, das porcentagens de cada tópico dentro de um documento, sendo a base para uma futura construção dos sistemas de organização do conhecimento.

Além disso, os instrumentos construídos serão utilizados para a Recuperação da Informação nos repositórios digitais que foram base para a sua construção. Tais instrumentos serão capazes de fornecer um panorama geral dos repositórios, auxiliando com que os usuários encontrem informações mais aprimoradas a suas necessidades.

Neste sentido, a aplicação de técnicas de *machine learning*, contribui efetivamente para tornar a Recuperação da Informação mais efetiva, pois os algoritmos irão contribuir para os sistemas de organização do conhecimento representar efetivamente os conteúdos, e assim, melhorar a recuperação da informação nestes ambientes.

A figura 1 apresenta um modelo conceitual de como as técnicas de *machine learning*, *topic modeling* e LDA, são aplicadas as bases dos repositórios digitais, e como isso é utilizado para a construção de instrumentos de sistema de organização do conhecimento.

Figura 1 – Modelo baseado em *Machine Learning* para Construção de SOC



Fonte: Elaborado pelos autores (2018).

O modelo apresentado na figura 1 demonstra como os dados dos repositórios digitais são utilizados para a construção dos sistemas de organização do conhecimento. Em suma, aplica-se os algoritmos de *topic modeling* sobre os repositórios digitais. O que é obtido por meio desse algoritmo, juntamente com as técnicas de *machine learning*, gera uma lista de termos contextualizados. A partir dessa lista, aplica-se as técnicas LDA, obtendo as relações e demais informações sobre os termos obtidos, que é utilizado para a construção dos SOC. Essas informações podem ser centrais na reflexão dos SOC, uma vez que permitirão escolher os termos e relações que a princípio não são identificadas pelos profissionais.

A partir do modelo construído, apresenta-se a seguir um caso que visa ser uma prova de conceito desta proposta. Busca-se nesse caso demonstrar a viabilidade da aplicação das técnicas de *machine learning* para a construção de sistemas de organização do conhecimento.

3.1 CASO – *TOPIC MODELING* e LDA APLICADO NA CONSTRUÇÃO DE SISTEMAS DE ORGANIZAÇÃO DO CONHECIMENTO

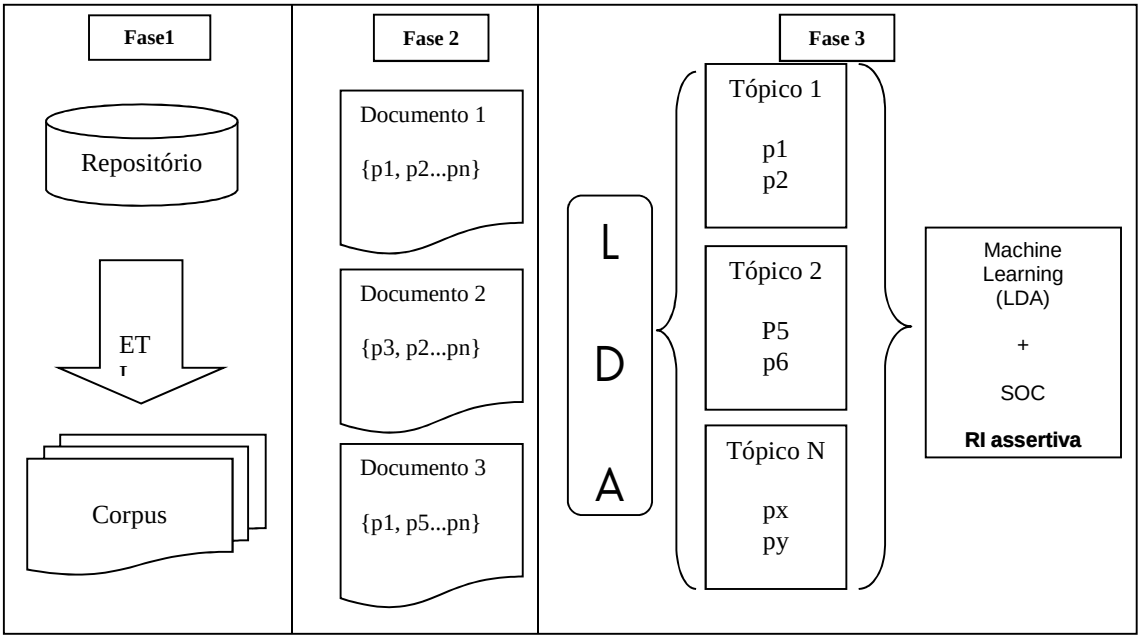
Foi desenvolvida uma prova de conceito visando comprovar o modelo proposto. Para isso, utilizou-se um repositório digital contando com aproximadamente 255 documentos, como fonte de informação para a aplicação das técnicas relatadas. Na sua maioria, os documentos desse repositório são trabalhos de conclusão de curso.

O primeiro procedimento para a realização dessa prova de conceito, foi a transformação dos arquivos de ".PDF" dos documentos para ".TXT". Essa mudança é necessária para que pudessem analisar os documentos como texto puro.

Em um passo posterior, realizou-se o carregamento dos arquivos para um ambiente de desenvolvimento com a linguagem R (RStudio). Tal ambiente permite a execução de scripts na linguagem R, que por sua vez possui uma série de bibliotecas (programas escritos por outras pessoas que possam ser utilizados de forma livre). No âmbito deste trabalho, utilizou-se a biblioteca denominada TM (*Topic Modeling*), que permite a aplicação do algoritmo LDA (*Latent Dirichlet Allocation*) sobre os documentos obtidos no repositório.

Os procedimentos seguintes foram divididos em três fases, partindo desde a extração dos documentos, até a obtenção dos tópicos e termos que são base para a construção dos sistemas de organização do conhecimento. Essas fases estão apresentadas na figura 2.

Figura 2 – Fases do processo de aplicação do LDA no repositório



Fonte: Elaborado pelos autores (2018).

Na primeira fase, aplica-se a técnica chamada de ETL (*Extract, Transform, Loading*), em que, primeiramente, é realizada uma extração dos arquivos do repositório, e posteriormente, após realizadas as devidas melhorias nos arquivos e transformações necessárias, faz-se o carregamento dos arquivos que gera como resultado o corpus ao qual será aplicado o algoritmo LDA.

A fase 2 representa o momento em que os documentos já estão processados, contendo em cada documento centenas ou milhares de palavras. Esses documentos servem de base para proporcionar ao ambiente de programação a utilização plena do algoritmo de LDA, que posteriormente alimenta a fase 3.

Por fim, a fase 3 é onde de fato ocorre o processo do algoritmo LDA. A partir dos documentos o algoritmo faz uma leitura de cada palavra em cada documento e começa a fazer, em um primeiro momento, uma alocação de cada grupo de palavras para um determinado tópico.

Neste primeiro momento, a alocação de palavras para determinados tópicos não resulta em uma representação assertiva do que cada documento de fato possui como conteúdo, portanto, são necessários mais passos e execuções iterativas para que esta alocação se torne relevante.

Em um próximo passo, o algoritmo faz uma nova leitura abordando os seguintes aspectos:

a) Qual a proporção de uma determinada palavra no documento analisado, palavra esta que está atualmente atribuída a um determinado tópico?
b) Qual a proporção de atribuições para o tópico analisado, sobre todos os documentos, que venham da palavra em questão?
c) Analisando os passos a e b, retribua a palavra analisada a um novo tópico baseado na probabilidade desta palavra pertencer a este novo tópico. Desta forma aloca palavras ou termos contextualizados que possuem aproximação semântica à um mesmo tópico.

Como resultado deste processo, após a execução do algoritmo na fase 3, serão obtidos um conjunto de tópicos. Cada tópico gerado será composto de palavras que semanticamente possuem uma aproximação em relação a todo o corpus analisado.

Além disso, como cada documento é composto por uma mistura de um ou mais tópicos e cada tópico um conjunto de palavras, a busca não mais ficará baseada em palavras chave, mas sim a uma distribuição de tópicos com termos contextualizados que permitem a construção de sistemas de organização do conhecimento mais assertivos.

Uma outra vantagem na utilização de LDA (*machine learning*) é a capacidade de automaticamente reconhecer um novo documento adicionado ao repositório de documentos científicos, pois uma vez que os tópicos foram criados então cada documento tem uma porcentagem de tópicos em seu conteúdo. Se um determinado documento foi analisado pelo LDA o mesmo será composto por porcentagens de tópicos, exemplo: um determinado documento: composto por 70% tópico 1 e 30% tópico 2. Em um caso específico do caso analisado obteve-se documentos com proporção de 70% relacionado ao tópico “sistemas web” e 30% relacionado ao tópico “startups”.

Desta maneira, além da recuperação mais assertiva, tem-se também a possibilidade de análise contínua sobre todos os documentos que poderão ser adicionados ao repositório, mantendo-se desta maneira o repositório atualizado para análises futuras.

5 CONSIDERAÇÕES FINAIS

Atualmente os repositórios digitais podem ser um dos principais espaços para proporcionar análises de dados e a extração de conhecimento que até pouco tempo não estavam acessíveis de forma explícita aos pesquisadores. Esse cenário ocorria, pois, estas análises necessitam de um processamento computacional intensivo, o que há pouco tempo tinha um custo muito alto tanto financeiro quanto em tempo hábil de resolução destas análises.

Além disso, os repositórios digitais possuem grande quantidade de coleções de documentos e analisar de forma mais assertiva e efetiva torna-se um problema e um desafio já que dado o grande volume, variedade de tipos de dados e ainda a crescente velocidade de geração caracterizam necessidade de auxílio computacional. A análise de textos científicos tem grande importância já que documentos científicos possuem um vocabulário diferenciado, são portadores de conhecimento que, são e serão utilizados para tomadas de decisão e são veículos de inovação tanto para a área acadêmica quanto para o mercado.

O cenário apresentado que é caracterizado como *Big Scholarly Data*, apresenta uma gama de oportunidades para aprimorar o modo como pesquisadores interagem com os ambientes informacionais digitais, quando utilizados na construção de ferramentais que aprimoram o modo como as informações são representadas e recuperadas. Neste cenário, a Ciência da Informação é capaz de auxiliar em transpor as suas teorias e técnicas para a construção de ferramentais que utilizam as análises de dados, favorecendo com que os usuários sejam capazes de localizar aquilo que necessitam com mais precisão.

Assim, este trabalho discorre sobre como técnicas de *machine learning*, quando aplicado a cenários de grandes conjuntos de dados acadêmicos (repositórios digitais), podem auxiliar a construção de sistemas de organização do conhecimento mais precisos e contextualizados com as informações armazenadas nestes ambientes.

O modelo construído demonstra um caminho possível que relaciona esses diferentes campos de estudos, apresentando um meio de utilizar o conteúdo dos repositórios, com o uso de algoritmos de *topic modeling* e LDA. A partir dessa aplicação dos algoritmos nas bases de dados, apresenta-se um modo de aprimorar a construção de sistemas de organização do conhecimento.

Finalmente, o trabalho apresenta um caso em que o modelo proposto é aplicado, em que se utiliza de algoritmos de *machine learning* em um repositório digital real. Esse caso foi capaz de demonstrar a viabilidade dessa proposta, demonstrando o relacionamento existente entre *Big Scholarly Data*, *machine learning*, sistemas de organização do conhecimento e repositórios digitais.

Este trabalho é parte de uma pesquisa que busca estabelecer um modelo para análise de sociedades atuais, auxiliado por poder computacional, para que através de análises mais complexas permeadas pelo poder computacional e reflexão pela cibercultura de Pierre Levy, seja possível proporcionar entendimentos, análises e intuições de impacto na sociedade.

Big Data in the academic data context: the use of machine learning in the construction of knowledge organization system

ABSTRACT

The amount of data in the academic context has increased significantly, being caused mainly by the increase of the amount of scientific documents. In this sense, there is a powerful information source, framed in the context of the Big Scholarly Data, which can be used to analyze and improve the traditional techniques of Information Science, by providing relevant knowledge for the construction of Knowledge Organization Systems that more accurately reflect data contained in these environments. In addition, one way of performing the treatment and analysis of this data conglomerate is related to machine learning. However, it is necessary to reflect how the techniques of machine learning when applied to large volumes of data can favor the construction of knowledge organization systems. Thus, the objective of this work is to discuss machine learning, especially the topic modeling method, in the process of building knowledge organization systems, aiming to discuss how data belonging to the BSD domain can provide information necessary to improve organizational tools of knowledge. For this, a methodology with both bibliographic and applied characteristics was used. As results, first, a theoretical model was proposed, which links the data of digital repositories, with the techniques of machine learning, for the construction of systems of knowledge organization. In order to validate this model, a proof of concept was performed using the topic modeling method and the LDA to identify topics of interest within a scientific corpus of a digital repository in order to organize, summarize and understand its content, aiming to provide a basis for the construction of knowledge organization systems. Finally, it was concluded that the application of the techniques of machine learning contributes to provide a broad overview of the data contained in digital repositories, being a very rich base to support and improve the construction of knowledge organization systems.

Keywords: Machine learning. Digital repositories. Topic modeling. System of knowledge organization.

REFERÊNCIAS

ARAÚJO, C. A. Á. O sujeito informacional no cruzamento da Ciência da Informação com as ciências humanas e sociais. In: ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO, 14., **Anais do Enancib**, 2013.

ASSUNÇÃO, M. D. et al. Big Data computing and clouds: Trends and future directions. **Journal of Parallel and Distributed Computing**, v. 79–80, p. 3–15, 1 maio 2015.

BLEI, D. M. Introduction to Probabilistic Topic Modeling. **Communications of the ACM**, 2012.

BOYD-GRABER, J.; HU, Y.; MIMNO, D. Applications of Topic Models. **Foundations and Trends® in Information Retrieval**, v. 11, n. 2–3, p. 143–296, 2017.

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

CARLAN, E.; CARLAN, E.; MEDEIROS, M. B. B. Sistemas de Organização do Conhecimento na visão da Ciência da Informação. **Revista Ibero-Americana de Ciência da Informação**, v. 4, n. 2, 15 fev. 2012.

FALEIROS, Thiago; LOPES, Alneu de Andrade. **Modelos probabilísticos de tópicos**: desvendando o latent dirichlet. Relatórios técnicos. São Paulo: ICMC, 2016.

FERNEDA, E. **Recuperação de Informação**: estudo sobre a contribuição da Ciência da Computação para a Ciência da Informação. 2003, 147 f. 2003. Tese (Doutorado em Ciências da Comunicação) - Escola de Comunicação e Artes, Universidade de São Paulo, São Paulo, 2003.

INTERNATIONAL SOCIETY FOR KNOWLEDGE ORGANIZATION., I.; CHAPTER, I. S. FOR K. O. R. **Knowledge organization : KO.** [s.l.] INDEKS Verlag, 1993.

KHAN, S. et al. A survey on scholarly data: From big data perspective. **Information Processing & Management**, v. 53, n. 4, p. 923–944, 1 jul. 2017.

MITCHELL, T. M. **Machine Learning**. [s.l: s.n.]. Disponível em: https://www.cs.ubbcluj.ro/~gabis/ml/ml-books/McGrawHill - Machine Learning -Tom Mitchell.pdf. Acesso em: 27 jul. 2018.

SARACEVIC, T. Interdisciplinary nature of information science. **Ciência da informação**, v. 24, n. 1, p. 36-41, 1995. Disponível em: http://www.brapci.ufpr.br/documento.php?dd0=0000005946&dd1=59269. Acesso em: 05 ago. 2018.

XIA, F. et al. Big Scholarly Data: A Survey. **IEEE Transactions on Big Data**, v. 3, n. 1, p. 18–35, 1 mar. 2017.