MINERAÇÃO DE TEXTOS: contribuições à indexação de obras e aos usuários de bibliotecas

Text mining: contributions to the indexing of works and to library users

Keli Rodrigues do Amaral Benin (1), Luiz Fernando Puttow Southier (2), Lovania Roehrig Teixeira (3), Jefferson Tales Oliva (4), Marcelo Teixeira (5)

(1) Universidade Tecnológica Federal do Paraná, Brasil, keliamaral@utfpr.edu.br

(2) luizsouthier@gmail.com

(3) lovaniateixeira@gmail.com

(4) jeffersontalesoliva@gmail.com

(5) mtex@utfpr.edu.br

Resumo

Em sistemas de busca por obras bibliográficas é comum o uso de vocabulários controlados para indexação, os quais possuem uma terminologia padronizada e estável ao longo do tempo. No entanto, essa formalização nem sempre reflete a linguagem cotidiana dos leitores, que pode ser involuntariamente transferida para o sistema de busca pelos usuários, especialmente em ambientes digitais, o que pode limitar a recuperação da informação. Este artigo propõe uma abordagem baseada em mineração de textos que visa complementar a indexação formal de obras com termos extraídos de comentários espontâneos de leitores, coletados em plataformas online. A proposta foi aplicada ao catálogo de uma biblioteca universitária, utilizando como fonte auxiliar os comentários disponíveis no comércio eletrônico de livros da Amazon. Foram analisados comentários sobre 10 obras com volume representativo de interações. Os resultados indicam que aproximadamente 75% dos termos recomendados pela abordagem não estavam presentes na indexação original, embora estivessem semanticamente relacionados ao conteúdo das obras. A análise sugere que os termos adicionais têm potencial para enriquecer o sistema de indexação, ampliar a sua cobertura temática e fortalecer a mediação semântica entre o vocabulário informal e os descritores formais de indexação.

Palavras-chave: Mineração de textos; Indexação; Significado. Representação de conhecimento.

BENIN, Keli Rodrigues do Amaral; SOUTHIER, Luiz Fernando Puttow; TEIXEIRA, Lovania Roehrig; OLIVA, Jefferson Tales; TEIXEIRA, Marcelo. Mineração de Textos: contribuições à indexação de obras e aos usuários de bibliotecas. *Brazilian Journal of Information Science: research trends*, vol.19, publicação contínua, 2025, e025022. DOI: https://doi.org/10.36311/1981-1640.2025.v19.e025022.

Abstract

In bibliographic search systems, the use of controlled vocabularies for indexing is common. These vocabularies feature standardized and stable terminology over time. However, such formalization does not always reflect the everyday language of readers, which may be unintentionally transferred to the search system by users, especially in digital environments, potentially limiting information retrieval. This article proposes a text mining-based approach aimed at complementing the formal indexing of works with terms extracted from spontaneous reader comments collected on online platforms. The proposed method was applied to a university library catalog, using reader reviews from Amazon's book marketplace as an auxiliary source. Comments on 10 works with a significant volume of interactions were analyzed. The results show that approximately 75% of the terms recommended by the approach were not present in the original indexing, although they were semantically related to the content of the works. The analysis suggests that these additional terms have the potential to enrich the indexing system, broaden its thematic coverage, and strengthen the semantic mediation between informal vocabulary and formal indexing descriptors.

Keywords: Text mining; Indexing; Meaning; Knowledge representation.

1 Introdução

Na atual era do *big data*, um crescente volume de dados tem sido gerado pela humanidade a cada segundo. Estima-se que, a cada ano, mais de 150 zettabytes **(1)** de dados sejam gerados, resultado de interações cotidianas em redes sociais, plataformas digitais e ambientes *online* diversos (Santos *et al*. 2021; Lv et al. 2017). Esse crescimento no volume de dados não apenas desafia as formas tradicionais de armazenamento e processamento computacional, mas também altera a maneira como as pessoas se comunicam, leem, escrevem e interagem com as informações. A linguagem cotidiana tornou-se mais dinâmica, fragmentada e adaptada ao meio digital, influenciada por fatores socioculturais e suas constantes mudanças. Como consequência, a indústria e a academia têm despendido esforços na tentativa de propor alternativas para o tratamento de conjuntos volumosos de dados, o que passa, sobretudo, por sua leitura, pré-processamento e caracterização semântica.

Nesse cenário, muitos hábitos e formas de interação com a informação e o conhecimento foram transformados. Entre as práticas impactadas, destaca-se a indexação de conteúdos, especialmente em bibliotecas. O vocabulário informal, típico dos ambientes digitais, distancia-se do vocabulário controlado empregado nesses sistemas, gerando um descompasso entre a linguagem do usuário e a dos mecanismos de busca. Como resultado, os sistemas tradicionais de indexação revelam-se limitados na representação temática das obras, o que compromete a eficácia

da recuperação da informação, sobretudo entre os usuários que adotam padrões linguísticos mais atuais e menos formais.

A literatura tem investigado a representação do conhecimento em sistemas de recuperação da informação, com destaque para a indexação em ambientes bibliográficos. Estudos como os de Lancaster (2004) e Boccato (2005) apontam que os vocabulários controlados garantem padronização, mas apresentam baixa flexibilidade diante das transformações linguísticas e culturais. Em resposta, algumas pesquisas têm proposto abordagens complementares, como a folksonomia ou indexação social (Catarino e Baptista 2007; Santos e Corrêa 2018; Guedes *et al* 2011), nas quais os próprios usuários contribuem para a atribuição de rótulos temáticos. Coneglian (2020) e Jo (2024) defendem o uso de técnicas de inteligência artificial e mineração de textos como estratégias para ampliar a capacidade semântica desses sistemas. No entanto, não se dispõe, ainda, de iniciativas que explorem, de forma automatizada, a geração de vocabulário auxiliar a partir de comentários espontâneos de leitores em ambientes digitais, como os encontrados em plataformas de comércio eletrônico. Em tese, caso se dispusesse de um mecanismo automático, capaz de monitorar continuamente esses comentários, interpretá-los, e integrá-los dinamicamente aos mecanismos de indexação, os sistemas de busca poderiam se tornar automaticamente reconfiguráveis e sensíveis às variações lexicais, ampliando sua eficácia e adaptabilidade na recuperação da informação.

Para esse propósito, a Aprendizagem de Máquina (AM) surge como uma alternativa promissora. Um algoritmo de AM consiste na aplicação sistemática de processamento computacional para o reconhecimento de padrões e extração de conhecimento em dados (Mitchell 1997). A AM tem sido aplicada com sucesso em áreas como a medicina, a segurança e a indústria (Srinivas et al. 2021). Porém, o desempenho dessas abordagens na representação de conhecimento depende de como os dados são organizados **(2).**

A mineração de texto surge como uma técnica de AM que visa extrair informações úteis desses textos não estruturados, a fim de munir gestores a tomarem decisões mais assertivas e amparadas de percepção artificial (Kao 2007). Embora o conceito de mineração de textos não seja novo, ainda se percebe espaço para inovações, em particular no que tange às suas aplicações (Jo

2024). Um exemplo de aplicação, que é de interesse particular deste trabalho, advém dos sistemas de indexação de obras, como aquelas das bibliotecas que compõem importante esfera das instituições de ensino (Nunes 2004; Madureira; Santana; Santos 2022; Santos 2011). Nesse nicho de aplicação, o acervo é catalogado por bibliotecários, que usam palavras-chave para caracterizar a obra, o que repercute na forma como os usuários encontrarão as obras por meio de sistema de busca e recuperação de informação. Os conceitos de representação descritiva e representação temática são cruciais para a organização e a recuperação da informação. A representação descritiva (catalogação) reúne os aspectos físicos e formais de um item bibliográfico com intuito de torná-lo único entre os demais em um determinado acervo, permitindo identificá-lo, localizá-lo e representá-lo nos catálogos (Lima 2020). A representação temática (ou indexação) é aplicada na análise do conteúdo temático do documento, isto é, colabora na identificação dos assuntos tratados no texto usando vocabulário controlado ou não (Lancaster 2004) e é sobre esse ponto que este artigo reflete.

Este artigo propõe, portanto, uma abordagem baseada em mineração de textos para identificar, extrair e integrar novos indexadores, gerados a partir do refinamento terminológico de comentários de leitores em bibliotecas digitais na internet, aos indexadores formais das obras bibliográficas. Ao comentar uma obra, espera-se que o leitor se expresse por meio de um vocabulário informal e espontâneo, que reflete sua percepção e a linguagem de seu cotidiano. Este artigo confronta esse vocabulário emergente com o vocabulário controlado **(3)** e, a partir dessa comparação, propõe a integração dos termos informais aos indexadores formais já existentes. Observando esse processo, pode-se afirmar que as palavras-chave disponíveis nos catálogos, em muitos casos, podem estar desatualizadas em relação ao conjunto lexical usado atualmente pelos falantes nas interações, algo em constante evolução (Labov 2008, 2010). Especialmente, por se tratar de língua natural e essa estar sujeita às múltiplas influências de aspectos extralinguísticos, tais como o crescente uso das mídias sociais, o surgimento de palavras e o desuso de outras.

Essa integração amplia e potencializa os resultados de busca, ao incorporar um léxico autêntico e compartilhado por leitores que já conhecem as obras, aproximando o sistema da linguagem dos usuários. A eficácia da abordagem é avaliada comparando-se o desempenho do

sistema antes e depois da integração, evidenciando uma melhoria significativa na recuperação da informação bibliográfica.

A eficácia da abordagem proposta na recuperação da informação é comparada com o desempenho do sistema antes da integração, e os resultados sugerem uma melhoria substancial na recuperação correta das obras. Estruturalmente, assim, na Seção 2, apresenta-se a fundamentação teórica; na Seção 3, introduz-se a metodologia e os resultados; e na Seção 4 discutem-se as conclusões e as perspectivas futuras.

2 Fundamentação teórica

Nas áreas de Biblioteconomia e Ciência da Informação, especificamente em seus desmembramentos voltados à recuperação, à catalogação, aos aspectos linguísticos e à gestão de informação, a representação temática eficiente de obras é elementar. Além de auxiliar na recuperação correta de documentos, essa representação também expõe conteúdos internos dos documentos, potencializando assim uma relação harmônica e proveitosa entre leitor e obra (Maimone e Gracioso 2007).

Nesse contexto, representar a informação implica em conhecer as necessidades informacionais de quem a usa (Santos, Oliveira e Lima 2017). Na prática, espera-se que a informação simplifique a interface entre usuário e sistema (Boccato 2005). Porém, esse nem sempre é o caso, haja visto que os sistemas de informação modernos, em geral, geram e estão sujeitos a grandes volumes de dados e às interferências causadas por linguagem natural, que tende a sofrer variações ao longo do tempo, reflexo de relações entre variáveis linguísticas (léxico usado em determinada região, gírias de determinado grupo etário, linguagem técnica de determinada área, etc) e não-linguísticas (idade, condição social, nível de escolaridade, etc). Isso tudo dificulta encontrar e recuperar a informação. É nesse sentido que os sistemas de indexação se aplicam e em vista desses elementos que seu potencial pode ser desenvolvido. Nesse contexto, este artigo discute a Organização da Informação, constituída de processos de representações de objetos informacionais destinados a facilitar a recuperação eficaz por parte dos usuários (Lima 2020 p. 63).

Alguns estudos recentes têm explorado formas de incorporar aspectos de linguagem dos usuários em processos de indexação automatizada. Por exemplo, Leite e Martins (2024) propõem um modelo baseado na extração de termos a partir de trechos destacados por leitores em documentos acadêmicos, aproximando a organização da informação da percepção do usuário final. Leavy *et al* (2019) e Leavy, O'Sullivan e O'Connor (2023) propõem uma abordagem voltada à curadoria temática de grandes coleções literárias por meio da combinação de *embeddings* neurais e conhecimento especializado, permitindo a construção de léxicos adaptados a diferentes contextos de análise textual. Em termos de revisões sistemáticas da literatura, Li *et al* (2022) exploram o potencial de técnicas de mineração de conteúdo gerado por usuários para enriquecer sistemas de recomendação com base em comentários e avaliações informais; enquanto; Golub (2021) expõe os princípios básicos que norteiam a indexação automatizada de assuntos, principais abordagens e possíveis aplicações em sistemas de bibliotecas reais, bem como lacunas ainda abertas na literatura.

Além das abordagens relacionadas ao vocabulário, estudos têm investigado aspectos comportamentais e de experiência de uso em sistemas de informação. Kreutz *et al* (2023) propõem um modelo formal para avaliar estratégias de busca em bibliotecas digitais e identificar lacunas entre as necessidades dos usuários e as funcionalidades oferecidas pelos sistemas. Kuhar e Mercun (2022) combinam técnicas de *eye-tracking* e questionários para analisar a experiência do usuário em bibliotecas digitais, evidenciando que elementos como a posição da caixa de busca impactam significativamente na percepção de usabilidade. Greenberg e Bar-Ilan (2017) analisam registros de uso em uma biblioteca universitária e mostram que muitos usuários preferem buscadores como o *Google Scholar* em detrimento ao sistema institucional; já Zhang, Niu e Promann (2017) destacam que a experiência de descoberta e uso de *e-books* está diretamente ligada à navegabilidade, aos recursos de destaque e anotação, e à consistência entre plataformas.

Embora esses estudos avancem na direção de tornar os sistemas de busca mais sensíveis à linguagem do usuário, eles, em geral, não tratam diretamente da integração dinâmica entre vocabulário informal, extraído de comentários espontâneos sobre obras, e a terminologia formalmente adotada em sistemas de indexação de bibliotecas. Além disso, apesar de identificarem

comportamentos e preferências dos usuários, tais investigações raramente exploram como incorporar esse comportamento à indexação formal.

Este estudo busca preencher essas lacunas ao integrar não apenas o percurso e as interações do usuário, mas também o vocabulário informal gerado nos comentários em plataformas públicas, conectando-o aos descritores formais dos catálogos bibliográficos das bibliotecas. Essa abordagem visa tornar os sistemas de busca mais responsivos, interpretativos e adaptáveis às variações linguísticas dos leitores contemporâneos.

2.1 Indexação e folksonomia

A indexação faz parte de um conjunto de atividades que visam caracterizar o conteúdo de documentos e, possivelmente, favorecer a escolha de termos que permitam representar um determinado assunto, facilitando a sua recuperação (Madureira, Santana e Santos 2022; Normas Técnicas 1992; Nunes 2004; Santos 2011).

Na prática, o processo de indexar documentos depende fundamentalmente da percepção e interpretação humana sobre a temática da obra e sobre como bem representá-la. Como tal, esse processo pode ser complexo, já que é preciso antecipar os termos que, em tese, podem satisfazer a um conjunto numeroso de possíveis consultas (Lancaster 2004). Desse modo, é indispensável que os indexadores usados sejam amplos, condizentes e representativos em relação ao tema da obra, sem comprometer a integridade da informação no momento da busca. Somado a isso, estão os fatores linguísticos que podem dificultar a seleção de uma palavra ou expressão na indexação de obras bibliográficas. Por exemplo, se tomarmos o léxico do Português Brasileiro nos deparamos com palavras ambíguas (Abbott 1997; Chierchia 2003;) (com mais de um sentido dependendo do contexto de uso) como "banco" que pode ser um móvel (banco da praça), uma instituição financeira (Banco do Brasil) ou coleção/conjunto (banco de sangue/banco de dados) que se usadas inadequadamente podem levar a uma obra pela qual o usuário não busca **(4)**. A partir disso, a escolha dos termos indexadores deve envolver não apenas informações sobre o conteúdo da obra, mas também a pluralidade linguística, já que os usuários dos sistemas de busca de bibliotecas são, também, usuários da língua.

Assim, a concepção da indexação, conforme tomada neste artigo, é aquela oriunda da *web* 2.0, para a qual a representação temática é concebida como uma indexação social, ou seja, ela resulta da ação colaborativa de usuários, denominada "folksonomia". Esse conceito surgiu da criação da *web* e da necessidade de lidar com a multiplicação de conteúdos que, em sua grande parte, decorrem da interação entre múltiplos usuários (Catarino e Baptista 2007). Como tal, é importante que haja mecanismos para a etiquetagem de conteúdos conforme sua classificação. É nesse contexto que emerge o termo "folksonomia" do inglês em que se tem "folksonomy" - "folk" (povo) + "sonomy" (taxonomia) (Santos e Corrêa 2018; Lopes e Albuquerque 2022).

Wal (2007) indica que "folksonomia" pode ser o resultado de uma etiquetagem de informações e objetos realizada por uma pessoa para sua própria recuperação, mas também pode ser feita em ambientes coletivos. Assim, qualquer informação com um *Uniform Resource Locator* (URL) pode constituir uma folksonomia, ou seja, pode ser etiquetada colaborativamente na *web* (Brandt e Medeiros 2010 p. 112; Vignoli, Almeida e Catarino 2014 p. 125). Assim, os usuários da *web* ao utilizar *tags* ou ao etiquetar conteúdos por meio de palavras-chave/termos estão realizando a "etiquetagem social", isto é, indexando conteúdos de forma colaborativa (Wal 2007; Vignoli, Almeida e Catarino 2014;). A etiquetagem social, nesse contexto, representa o ato de etiquetar realizado pelo próprio usuário da informação e não é uma atividade realizada nem pelo autor, nem pelo profissional da informação.

A folksonomia, assim, define uma forma de rotulagem de conteúdos, seja por *tags*, etiquetas ou palavras-chave, que depende basicamente da forma como usuários, sejam eles humanos ou não, interagem e colaboram em ambientes digitais acerca de certa temática. O resultado dessa interação é um vocabulário representativo da obra, descrito em linguagem livre pelos próprios usuários da informação, conferindo um teor natural à escolha dos rótulos (Guedes *et al* 2011). Conforme apontam Vignoli, Almeida e Catarino (2014 p. 126), em folksonomias, a indexação ocorre em contexto adverso ao das linguagens controladas utilizadas por Profissionais da Informação, já que o vocabulário é construído pelos usuários utilizando a linguagem livre para etiquetar ou indexar seus conteúdos. Folksonomias, assim, se diferenciam de taxonomias, pois as primeiras são um recurso de vocabulário não controlado e carecem de precisão na recuperação de

informações, já as segundas são vocabulários controlados constituídos de regras rígidas que facilitam a recuperação da informação (Vaidya e Harinarayana 2016).

Uma etapa importante da folksonomia diz respeito à identificação de padrões que culminam nas preferências de cada usuário sobre uma obra. A tarefa de identificar tais padrões pode ser trabalhosa se executada manualmente, o que justifica esforços para a sua automação. Dentre as ferramentas que podem ser usadas para observar padrões na interface entre usuários e obras está a mineração por meio computacional, abordada a seguir.

2.2 Mineração de textos

A mineração de textos é uma subárea da mineração de dados (Feldman e Sanger 2007). Enquanto a mineração de dados se utiliza, em geral, de dados estruturados, que possuem uma organização bem definida, a mineração de texto consiste na extração de informações a partir de texto não estruturado. Usualmente, textos são compostos por uma sequência de palavras e frases que carregam um contexto semântico complexo, cuja determinação computacional pode envolver tarefas não triviais de processamento (Jo 2024).

Em essência, a mineração de textos tem como objetivo principal identificar termos em documentos e compreender os padrões associados à ocorrência desses termos, o que permite montar grupos temáticos baseados nesses padrões (Tuffery 2023; Aranha e Passos 2006). Para transformar dados textuais brutos em informações úteis, é indispensável representá-los adequadamente para, então, estabelecer os padrões de ocorrência, a regularidade como os termos ocorrem, e as tendências estatísticas para as próximas ocorrências. A arquitetura computacional responsável por essas tarefas compõe o núcleo da mineração de textos.

Tecnicamente, a mineração de textos se apoia em diversos fundamentos do Processamento de Linguagem Natural, como a modelagem de tópicos, a extração de entidades nomeadas, a análise de sentimentos e a classificação de documentos (Vajjala *et al.* 2020). Tais tarefas ampliam a capacidade de interpretar contextos, desambiguar sentidos e agrupar textos com base em similaridade semântica. Por exemplo, técnicas como *Latent Dirichlet Allocation* e *Non-Negative*

Matrix Factorization facilitam descobrir tópicos latentes em grandes coleções de texto, auxiliando na organização temática de acervos e no refinamento da busca por conteúdos específicos.

Metodologicamente, o processo de mineração de textos pode ser sumarizado pelas seguintes etapas: 1) coleta, 2) pré-processamento, 3) extração de características, 4) mineração, 5) análise dos resultados (Aranha 2007; Feldman e Sanger 2007; Jo 2024), as quais são descritas a seguir por fazerem parte da metodologia aplicada neste estudo.

- **Coleta**: envolve a obtenção de textos que servem como base para análise. Esse processo envolve a identificação das fontes de onde os textos são extraídos, como websites, redes sociais, bancos de dados, documentos digitais, e-commerce e arquivos de bibliotecas. A extração, em si, em geral se utiliza de meio automatizado, como web *scraping* (raspagem de dados) e *crawling* (navegação automatizada).
- **Pré-processamento**: uma vez coletados, os dados podem precisar passar por procedimentos de "limpeza", fusão, ou decomposição, a fim de que sejam padronizados para um formato único que possa ser processado. A etapa de pré-processamento constitui essas ações e é crucial para preparar os textos para análise. Trata-se de uma etapa construtiva de integração e remoção de ruídos, como *stopwords* (palavras irrelevantes para a análise, como "e", "de", "o"), tokenização (divisão do texto em unidades básicas, geralmente palavras ou frases) e normalização de termos (por exemplo, lematização e *stemming*, que transformam palavras em sua forma base).
- **Extração de características**: após a limpeza dos dados, a extração de características é a etapa responsável por identificar e extrair padrões e características relevantes do texto. Uma técnica comum para tal é a construção de vetores de características, como o *Term Frequency-Inverse Document Frequency* (TF-IDF), que tem a finalidade de quantificar a importância de uma palavra em relação ao conjunto de documentos.
- **Mineração**: os textos são analisados usando algoritmos de AM e técnicas estatísticas. O objetivo é identificar padrões, relações e tendências que possam ser úteis para a classificação, agrupamento (*clustering*) ou associação de dados. Modelos, como as redes neurais e os algoritmos de aprendizado são usualmente aplicados nesta etapa.

- **Análise de Resultados**: a partir do modelo de dados processado, os resultados precisam ser interpretados e expostos de modo compreensível e que visualmente facilitem a compreensão dos padrões e insights descobertos. Técnicas de visualização, como nuvens de palavras e gráficos de co-ocorrência, ajudam a tornar as descobertas mais acessíveis e compreensíveis para os tomadores de decisão.

A aplicação de mineração de texto em sistemas de bibliotecas universitárias tem se mostrado uma abordagem promissora para a melhoria da recuperação de informações. Ao integrar terminologias não padronizadas, como gírias e jargões coletados de fontes externas (por exemplo, resenhas de livros em sites de comércio eletrônico), com a terminologia técnica adotada pelos sistemas próprios de indexação, é possível ampliar a cobertura semântica (os sentidos e conteúdos veiculados) e melhorar a experiência de busca (Aggarwal e Zhai 2012).

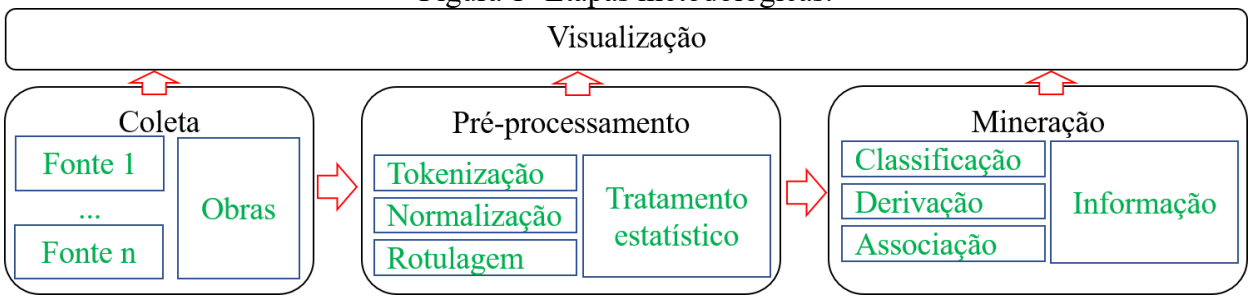
Nos dois ambientes envolvidos na análise, a biblioteca universitária e o catálogo da *Amazon*, ocorrem diferentes tipos de indexação que buscamos integrar. No primeiro, a indexação tradicional ou taxonomia é conduzida por especialistas que utilizam vocabulários controlados, ou seja, um conjunto padronizado de termos para descrever o conteúdo de obras bibliográficas, por exemplo. O vocabulário controlado garante a consistência e a uniformidade na indexação, facilitando a recuperação da informação. No segundo ambiente ocorre uma indexação social ou folksonomia, constituída de vocabulário não-controlado, que carece de precisão na recuperação da informação (Vaidya e Harinarayana 2016).

Essa integração de terminologias tende a permitir uma recuperação de informações mais eficiente e também uma compreensão mais profunda do comportamento do usuário e de suas preferências, gerando *insights* que podem ser utilizados para aprimorar os serviços oferecidos pelas bibliotecas. A seguir, apresenta-se a proposta deste artigo alinhada a essa temática.

3 Proposta para a indexação de obras

A proposta metodológica deste trabalho considera as etapas descritas e encadeadas conforme a Figura 1.

Figura 1- Etapas metodológicas.

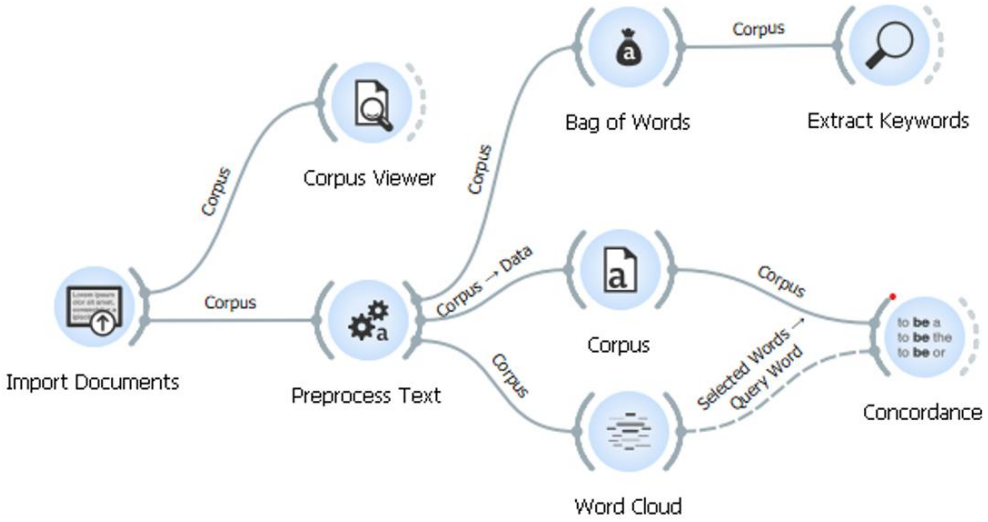


Fonte: os autores.

Inicialmente, proceder-se-á com a coleta dos dados a partir dos repositórios alvo, os quais serão então pré-processados, analisados estatisticamente e classificados conforme a terminologia correlata e, por fim, proceder-se-á com a interpretação de aderência que culminará na proposta de indexação complementar.

Para que esta metodologia possa ser aplicada na prática, este trabalho propõe a sua implementação no contexto de uma ferramenta de análise de texto. Foi escolhida para este trabalho a ferramenta *Orange Data Mining*, uma plataforma computacional que se destaca pelo modo intuitivo como disponibiliza e permite o uso de rotinas de inteligência artificial. Espera-se que essa simplicidade favoreça o uso da abordagem para comunidades além da computacional, socializando o estudo e a aplicação de inteligência artificial em múltiplos domínios. A versão da metodologia proposta, quando implementada no Orange, é apresentada na Figura 2.

Figura 2- Fluxo metodológico implementado no Orange Data Mining.



Fonte: os autores.

A seguir, cada uma dessas etapas é individualmente detalhada no contexto de cada passo metodológico correspondente.

3.1 Coleta de dados

O primeiro passo para a condução da abordagem de mineração de textos proposta foi extrair os dados de fontes originais. Na Figura 2, isso corresponde ao *widget* "Import Documents", que recebe uma lista, em formato txt, com o título da obra e as anotações correspondentes, como os comentários dos leitores ou os indexadores. A coleta dos dados visa, portanto, prover essa entrada para a ferramenta para que, a partir disso, se possa proceder às etapas subsequentes. O *widget* auxiliar "Corpus Viewer" pode ser utilizado como ferramenta de conferência e consulta aos dados coletados.

3.1.1 Descrição das fontes de dados

Neste trabalho, foram consideradas 2 fontes distintas de coleta de dados: (i) uma biblioteca digital da internet - o *site* de comércio eletrônico da Amazon; e (ii) o catálogo de uma biblioteca universitária de uma instituição de ensino superior pública do Paraná.

O *website* da Amazon foi escolhido por apresentar interações dos usuários ao avaliar obras, compartilhar percepções e comentar a experiência com o produto. Ao processar essas percepções, se torna possível entender o usuário e sua aderência aos produtos, conforme recursos característicos da web 2.0 como o tagueamento resultante da folksonomia (Campos 2007).

As etiquetas e as classificações de livros na *Amazon*, especialmente quando feitas por usuários, são exemplos de etiquetagem social, pois trata-se de uma modalidade de indexação em língua natural realizada pelos próprios consumidores do *website*. Assim, essa atividade é resultante da necessidade desses sujeitos e orientada pela natureza informal desse contexto (Vignoli, Almeida e Catarino 2014). Em um ambiente como esse, por ser virtual, de cunho informal e que leva os usuários a tecerem comentários subjetivos sobre determinada obra, estimula-os a usarem uma linguagem menos formal e, assim, mais próxima da linguagem usada no dia a dia.

Nesse sentido, a *Amazon*, como outras plataformas *online*, utiliza a etiquetagem colaborativa, onde os usuários atribuem etiquetas aos livros, adicionando informações não presentes nos metadados oficiais. Essa etiquetagem é utilizada para que os dados coletados dos usuários influenciem a classificação e a visibilidade dos produtos, incluindo livros.

Especificamente, a *Amazon* utiliza folksonomia larga, já que todos os usuários do *website* podem marcar qualquer produto e, assim, indexar produtos que estão sendo comercializados (Meira *et al* 2025). É possível selecionar uma *tag*, e assim encontrar discussões e pessoas relacionadas àquela marcação, formando-se comunidades direcionadas para cada categoria ou produto específico. A folksonomia no *website* da *Amazon* direciona a decisão de compra de um produto, uma vez que os aspectos positivos e negativos são evidenciados diretamente pelo público consumidor (Meira *et al* 2025).

Tecnicamente, assim, o ambiente escolhido permite a etiquetagem dos produtos pelos usuários. É por meio dessas etiquetas que a empresa classifica os produtos, materializando a ideia de como a folksonomia influencia na estruturação da taxonomia. Como resultado, por exemplo, a visualização de um produto implica na sugestão de produtos similares, que já foram comprados por outras pessoas e que podem interessar a determinado perfil de usuário, oferecendo assim uma experiência de navegação personalizada.

Em relação ao catálogo de uma biblioteca universitária, esse tem como público-alvo docentes, discentes, pesquisadores e funcionários da instituição e usuários externos. A maior demanda da biblioteca é atendida por meio do catálogo *online* da rede "Pergamum", que é um sistema de automação de bibliotecas que operacionaliza cadastro, empréstimo, localização, relatórios e outras funções. Esse contexto, diferentemente do anterior, é mais formal, pois se trata de um ambiente de uma instituição de ensino superior, e mais controlado, pois envolve usuários que, teoricamente, tem um nível de instrução mais homogêneo do que os usuários de um *website* de compras como a Amazon. Mas, mesmo assim, deve-se pensar que esses mesmos sujeitos fazem uso da língua em ambientes formais e informais e que podem se favorecer por um sistema de indexação mais próximo do seu cotidiano linguístico informal.

Aqui ressalta-se também que os domínios considerados (biblioteca e *Amazon*), possuem diferenças que impactam a recuperação da informação. Catálogos de bibliotecas universitárias, fazem uso de vocabulários controlados, garantindo precisão; a web depende primariamente da linguagem natural, exigindo sistemas que lidem com a inerente ambiguidade para conectar usuários e informações. Assim, a recuperação de informação nas bibliotecas é marcada pela especificidade e pela exaustividade: "exaustividade diz respeito ao âmbito de cobertura, a especificidade refere-se à profundidade de tratamento do conteúdo" (Lancaster 2004 p. 203); enquanto em um domínio como a *Amazon* é possível que se encontro exaustividade, mas não especificidade.

3.1.2 Seleção das obras

Para a seleção efetiva das obras, foram consideradas as 20 mais vendidas do catálogo da *Amazon* no período da coleta (denotadas OE - Obra Escolhida), das quais foram selecionadas 10 que possuíam pelo menos 10 comentários de leitores em cada uma. Esse critério busca garantir uma base mínima de textualidade por obra, suficiente para aplicar técnicas como TF-IDF com relevância léxica. Estudos similares em mineração de opinião sugerem que essa quantidade de comentários pode já tender ao propósito de revelar padrões terminológicos significativos (Liu 2012). Embora não estabeleça um número fixo de textos necessários, Liu (2012) argumenta que, em tarefas de extração de opinião e análise léxica, mesmo conjuntos pequenos de comentários já

podem revelar padrões relevantes e opiniões recorrentes, especialmente em abordagens exploratórias, como a aqui apresentada.

As obras selecionadas foram:

OE1 – "A coragem de ser imperfeito", Brené Brown 2013;

OE2 – "Assassinato no Expresso do Oriente", Agatha Christie 2020;

OE3 – "Cristianismo: puro e simples", C. S. Lewis 2017;

OE4 – "Essencialismo: a disciplinada busca por menos", Greg Mckeown 2015;

OE5 – "Mulheres que correm com os lobos: mitos e histórias do arquétipo da mulher selvagem", Clarissa Pinkola Estés 2018.

OE6 – "O homem mais rico da babilônia", George S Clason 2017;

OE7 – "Orgulho e preconceito", Jane Austen 2018;

OE8 – "Os segredos da mente milionária: aprenda a enriquecer mudando seus conceitos sobre o dinheiro e adotando os hábitos das pessoas bem-sucedidas", T. Harv Eker 1992;

OE9 – "Pai rico, pai pobre: o que os ricos ensinam a seus filhos sobre dinheiro", Robert T. Kiyosaki 2017.

OE10 – "Rápido e devagar: duas formas de pensar", Daniel Kahneman 2012.

3.2 Pré-processamento dos dados

A etapa de pré-processamento corresponde ao *widget* "Preprocess Text" da Figura 2. O objetivo é filtrar o texto para dividi-lo em unidades menores (*tokens*), executando as respectivas etapas de normalização (derivação, lematização), criação de n-gramas e rotulagem de *tokens* conforme classes gramaticais.

Para o pré-processamento dos textos, foram aplicadas a tokenização e a remoção das *stopwords*. A primeira visou remover caracteres desnecessários para a análise, o que é determinante para a extração de trechos mínimos de textos que sustentam a análise sem ruídos. Já para a remoção das *stopwords*, foi criada uma lista, também conhecida como *stoplist*, contendo os

termos que adicionam pouco valor semântico ao trecho, por exemplo, artigos, preposições, conjunções e pronomes.

O pré-processamento foi iniciado com a conversão dos textos para letras minúsculas e a eliminação de URLs e elementos HTML dos tokens. Posteriormente, realizou-se a tokenização, processo que divide o texto em unidades menores, como palavras, sentenças ou bigramas. Além disso, utilizou-se a função *Regex*, que permite identificar padrões de cadeias de caracteres de forma concisa e eficiente, como caracteres específicos, palavras ou sequências definidas. Esse recurso foi empregado para remover palavras que correspondiam às expressões regulares previamente estabelecidas. As expressões regulares configuradas para exclusão incluíam os seguintes caracteres especiais (onde a notação "|" indica a separação entre diferentes caracteres especiais):

```
|.|,|:|;|!|\?|\(|\)|\||\+|\'|\"|‘|’|"|"|\'|\’|…|\-|–|—|\$|&|\*|>|<|\/|\[|\]
```

Por fim, realizou-se a remoção das *stopwords*. Identificaram-se as palavras que não influenciam na análise e interpretação dos dados e, em seguida, processou-se a *stoplist*, o que resultou na exclusão dessas palavras dos textos. Além disso, foram analisados possíveis erros semânticos no documento, buscando avaliar a relevância real de cada termo no contexto de suas frases. Esse procedimento permitiu identificar e quantificar as palavras-chave mais significativas presentes no documento.

3.3 Extração de características

Após a verificação de erros semânticos, utilizou-se a saída do *widget* "Preprocess Text" associada ao *widget* "Bag of Words" para construir uma matriz de documento-termo com a configuração Count + IDF. O *widget* "Bag of words" implementa um algoritmo computacional que visa associar o texto de entrada a atributos numéricos que quantificam a presença, a contagem ou a frequência total dos termos de um determinado documento.

Para isso, os parâmetros escolhidos foram o número de vezes que uma palavra aparece no documento; e a frequência do documento que contém o termo. Assim, foi possível selecionar os termos com mais relevância do documento.

Em seguida utilizou o *widget* "Extract Keywords", para encontrar as palavras-chave mais significativas nos textos. O objetivo da extração de palavras-chave é encontrar frases que melhor descrevam o conteúdo de um documento automaticamente. Frases-chave, termos-chave, segmentos-chave ou simplesmente palavras-chave são as terminologias usadas para definir os termos que indicam as informações mais relevantes contidas no documento. O *widget* "word cloud" é um complemento importante nesse passo e pode ser usado para a visualização da distribuição das palavras no texto. A análise de texto utilizando esse recurso permite uma visão intuitiva e concisa das palavras-chave, temas e padrões mais relevantes em um texto, o que pode ser relevante para uma visão geral e pare determinar escolhas em termos de processamento.

Na configuração do *widget* foi adotada a técnica TF-IDF (*Text Frequency-Inverse Document Frequency*) simples, para evidenciar as palavras com a pontuação TF-IDF mais alta. O TF − IDF avalia o quanto um determinado termo importa em um documento (Tabosa e Al 2020). O Quadro 1 apresenta o resultado da aplicação do widget "Bag of Words" sobre o widget "Extract Keywords".

Quadro 1- Cálculo de TF- IDF das palavras.

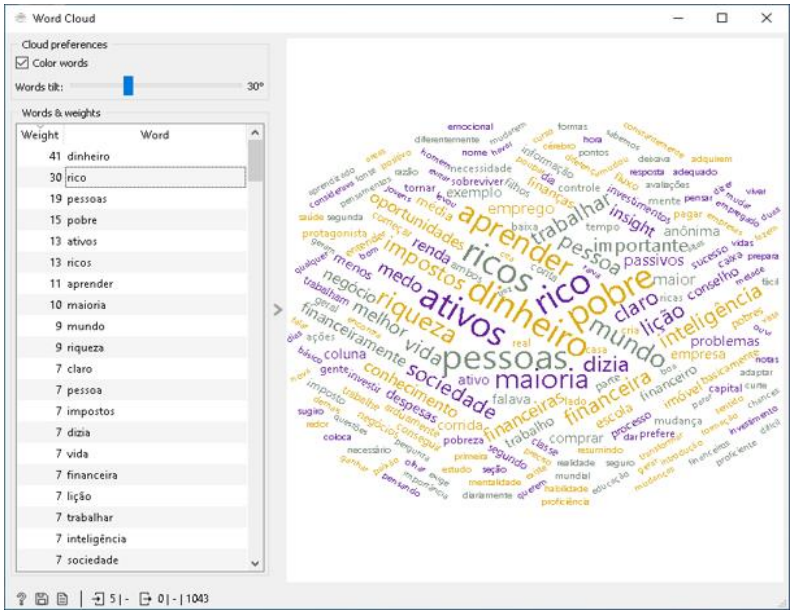
Palavra-chave	**TF-IDF**
dinheiro	1,131
rico	0,096
ativos	0,086
evitar	0,074
passivos	0,067
avaliações	0,059
bom	0,049
metade	0,049
preciso	0,049
absolutamente	0,048
entendem	0,048
entendo	0,048

Fonte: Dados da pesquisa

Em seguida, para identificar os termos encontrados nos textos inseridos na etapa anterior, conectou-se o "Preprocess Text" com o "Word Cloud", a fim de classificar as palavras em relação à frequência e ao peso semântico no texto. A *widget* em discussão, após coletar e analisar os textos,

identificou a presença de 200 palavras distintas, excluindo as *stopwords*. A Figura 3 ilustra as proporções estatísticas ocupadas pelas palavras-chave em relação aos textos.

Figura 3- Nuvem de palavras referente à frequência dessas palavras encontradas nos textos.



Fonte: os autores.

3.4 Mineração

Por fim, a proposta alcança a etapa de mineração, cujo objetivo é derivar informações pós-estatísticas sobre as obras. A base para essa etapa é o *widget* "concordance" que, dentre outras coisas, exibe a palavra escolhida no item anterior ("Word Cloud"), localizada no texto com 10 caracteres adjacentes à direita e à esquerda, como exemplo. Essa ferramenta também permite inferir informações relevantes sobre o contexto da palavra selecionada, identificando, por exemplo, se está acompanhada de um substantivo ou adjetivo, elementos que podem alterar seu significado na oração. Além disso, essa funcionalidade pode ser útil na reorganização textual, ao possibilitar ajustes nas posições das palavras. É possível, ainda, configurar a quantidade de palavras mostradas à esquerda ou à direita da palavra selecionada, com o objetivo de esclarecer ainda mais seu sentido.

Utilizando a classificação de "peso" e a medida estatística TF-IDF, quantificou-se a palavra "pessoas" com o valor (0.053) entre os termos, ou seja, aparecendo dezenove vezes no "corpus".

Porém a palavra "pessoa" é um substantivo e, como tal, não poderia ser um assunto, independentemente do peso e valor serem altos. Essa dedução foi construída a partir do *widget* "concordance" ao analisar os contextos em que a palavra aparece no texto. Também outros termos com peso e valor elevados, como "maioria" e "claro", não agregam poder semântico, ou seja, não representam um assunto, e não foram consideradas palavras-chave na análise.

Observa-se, no Quadro 2, que o *widget* "concordance" recupera a palavra consultada em um texto e exibe o contexto sob o qual ela está sendo utilizada. No exemplo a seguir, foi examinada a palavra "dinheiro".

Quadro 2 - Ocorrência de "dinheiro" em textos com separação antes e após.

Texto antes de "dinheiro"	**descritor**	**Texto depois de "dinheiro"**
Repetitivo, Foca bastante na diferença entre investimento que tiram	dinheiro	da sua conta e investimentos que colocam dinheiro na sua
que tiram dinheiro da sua conta e investimentos que colocam	dinheiro	na sua conta, corra qualquer livro escrito por um
os estudos, e nunca teve uma aula sequer sobre	dinheiro	investimento, juros, etc, Ou seja,
menos impostos é que "os ricos não trabalham por	dinheiro	Introdução Na introdução da obra ele descreve o
mais rico do Havaí. Basicamente a ideia sobre	dinheiro	fazia esse contraste entre eles! Falar sobre financeiro é
é eterna… Pai pobre dizia que não ligava para	dinheiro	e que o mesmo não era importante enquanto o pai
mesmo não era importante enquanto o pai rico dizia que	dinheiro	é poder… Conclusão deste início: pai pobre não
vida mais abundante financeiramente… Disto podemos tirar que o	dinheiro	pode ser uma fonte de poder, contudo mais poderosa
Contudo, mais poderosa ainda é a educação financeira! O	dinheiro	vem e vai, mas se tiver sido educado
Arcuri e Thiago Nigro (recomendado todos livros deles):	dinheiro	não aceita desaforo!! Lição 1: ricos não trabalham
aceita desaforo!! Lição 1: ricos não trabalham por	dinheiro	os ricos fazem com que o dinheiro trabalhe para eles
não trabalham por dinheiro os ricos fazem com que o	dinheiro	trabalhe para eles, diferentemente da classe média e baixa
diferentemente da classe média e baixa que trabalham por	dinheiro	Logo, estes últimos tornam-se escravos do
Logo, estes últimos tornam-se escravos do	dinheiro	Oportunidades vêm e vão. Ser capaz de tomar
uma combinação de raiva e amor. No caso do	dinheiro	a maioria das pessoas é conduzida pelo medo e por

Fonte: Dados da pesquisa

Por fim, realizou-se um recorte das palavras-chave com "peso e valor" de um 1 a dez 10. Então, para definir as palavras-chave de cada livro oriundo da *Amazon*, e compará-las com as mesmas palavras do catálogo da biblioteca, aplicou-se uma nova filtragem a partir da qual foram definidas até 5 palavras-chave com maior aderência à obra.

Quadro 3 - Extração de palavras-chave conforme peso e valor.

Obras	**Extração de Palavras- chave (TF–IDF de 1 a 10)**	**Concordâncias (1 a 5)**	**Palavras-chave (1 a 3)**	**Assuntos**
OE1	vulnerabilidade, autoajuda, vulnerabilidades, pesquisa, vida, vale, melhor, fechamos, realmente, coragem	vulnerabilidade, coragem, vulnerabilidades, autoajuda	vulnerabilidade, autoajuda, coragem	vulnerabilidade, autoajuda, coragem
OE2	assassinato, expresso, Poirot, história, Oriente, detetive, trama, clássico, espera, personagens	assassinato, Poirot, detetive, trama, personagens	assassinato, detetive, trama	assassinato, trama
OE3	cristão, vida, cristianismo, cristã, bastante, cristãos, perguntas, pontos, qualidade, analogias	cristão, vida, cristianismo, cristã, cristãos	cristão, vida, cristianismo, cristã	cristão, vida, cristianismo, cristã
OE4	essencialismo, melhor, essencial, vida, sobre, frases essencialista, delicados, tempo, atividade,	essencialismo, essencial, vida, essencialista	essencialismo, essencial, vida	essencialismo, vida
OE5	vida, histórias, mulher, mulheres, aprendendo, importância, mundo, análise, interessantes, situações	vida, histórias, mulher, mulheres, aprendendo	vida, histórias, mulheres	vida, histórias, mulheres
OE6	dinheiro, riqueza, homem, investir, conhecimento, lições, ricos ensinamentos, guardar, Babilônia	dinheiro, riqueza, investir, guardar, ensinamentos	dinheiro, riqueza, investir	dinheiro, riqueza, investir
OE7	história, época, filme, família, apaixonada, senti, amei, inglesa, sociedade, irônica	história, família, inglesa, sociedade	família, inglesa, sociedade	família, inglesa, sociedade
OE8	mentalidade, ricas, dinheiro, pobre, vida, dia, mudar, importa, pensamento, ajudará	mentalidade, ricas, dinheiro, pobre, vida,	mentalidade, dinheiro, vida	mentalidade, dinheiro
OE9	dinheiro, rico, ativos, evitar, passivos, avaliação, bom, metade, preciso, absoluto	dinheiro, rico, ativos, passivos,	dinheiro, rico	dinheiro, rico
OE10	vida, rápido, decisões, termos, economia, perda, manipulados, mente, ponto, tomada	vida, decisões, economia, manipulados, mente	vida, mente	mente

Fonte: Dados da pesquisa

3.5 Análise dos resultados

Na análise da obra OE1, os termos "vulnerabilidade, coragem e autoajuda" são candidatos a palavras-chave do livro, por permitirem apontar o assunto com razoabilidade, já que a obra sugere ao leitor aceitar a sua vulnerabilidade.

No livro OE2, os termos "assassinato" e "trama" despontam como palavras-chave, pois trata-se de uma ficção policial.

Na obra OE3, vários termos foram postulantes a apontar assunto principal. Na obra OE4, "essencialismo" foi o termo candidato a palavra-chave, já que o próprio título da obra e sua temática reportam o que é essencial para viver.

A obra OE5 aborda mitos, histórias e contos de fada, e foi possível apontar com razoabilidade termos que representassem essa temática. Para a obra OE6, os termos "dinheiro", "riqueza" e "investir" foram apontados como candidatos, pois a obra trata de finanças pessoais.

No caso da obra OE7, encontrou-se um resultado não previsto *a priori*, referente à identificação de termos sentimentais, como "amei" e "apaixonada", para representar a obra que, de fato, são aderentes à temática do livro.

Na obra OE8, que trata de educação financeira, foi possível levantar os termos "mentalidade" e "dinheiro", os quais procedem com o assunto da obra, analogamente ao resultado da obra OE9, que também trata de finanças pessoais.

Já em relação à OE10, o livro explica como funciona a mente humana em sua forma de pensar, assim, a análise aponta para palavras-chave como "mente", como representante coerente do assunto da obra.

3.6 Análise comparativa

Agora, discute-se como as palavras-chave identificadas na Amazon, fazendo o uso de folksonomia, se correlacionam com os descritores coletados do catálogo da biblioteca, usando indexação tradicional.

Para isso, foi adotado o descritor de similaridade de Jaccard (Lescovec, Rajaraman e Ullman 2020), que permite quantificar o índice de similaridade entre dois conjuntos de dados a partir da divisão do número de elementos na intersecção, pelo número de elementos da união entre esses dois conjuntos. Matematicamente, para dois conjuntos A e B de dados, a similaridade de A com B, denotada, J(A,B), é tal que $J(A,B) = |A \cap B| / |A \cup B|$.

O Quadro 4 apresenta esses resultados comparativos. Foram considerados homônimos os termos cujas variações são apenas de número e gênero gramaticais.

Quadro 4 - Estudo comparativo.

Obra	***Amazon (A)***	***Biblioteca (B)***	**J(A,B)**
OE01	vulnerabilidade; autoajuda; coragem	comportamento (modificação); vontade; assertividade (psicologia); coragem	0.17
OE02	assassinato; trama	literatura inglesa; ficção inglesa; ficção policial inglesa	0.0
OE03	cristão; vida; cristianismo; cristã	cristianismo; cristãos; teologia dogmática	0.5
OE04	essencialismo; vida	essência (filosofia); processo decisório	0.0
OE05	vida; histórias; mulheres	mulheres (psicologia)	0.3
OE06	dinheiro; riqueza; investir	riqueza; finanças pessoais; ética empresarial	0.2
OE07	história, família, inglesa, sociedade	literatura inglesa; ficção inglesa	0.0
OE08	mentalidade; dinheiro	capitalistas e financistas (psicologia); sucesso nos negócios; riqueza (psicológia); moeda (psicológia); milionários (psicologia); criatividade nos negócios; finanças pessoais	0.0
OE09	dinheiro; rico	investimentos; finanças pessoais	0.0
OE10	mente	processo decisório; raciocínio (psicologia); intuição (psicologia)	0.0

Fonte: Dados da pesquisa

O diagrama de Venn da Figura 4 complementa visualmente a interpretação desses dados.

Figura 4 - Diagrama de Venn de interpretação dos dados

Fonte: Dados da pesquisa

Pode-se perceber que a maioria, em torno de 75%, dos termos identificados como representantes da obra conforme os comentários de usuários, não são adotadas como palavras-chave no catálogo da biblioteca. Ainda assim, tais termos possuem características semânticas que de fato permitem, indiscutivelmente, identificar o conteúdo da obra.

Ou seja, apesar de alguns termos não serem na prática usados como palavras-chave, eles compõem rótulos que permitem identificar a obra e poderiam ser mais bem aproveitados para fins de indexação. Esse fato fica evidenciado, por exemplo, na obra OE5, para a qual a indexação tradicional contém um único descritor, "Mulheres (psicologia)". Por outro lado, a abordagem proposta sugere 3 palavras representativas, induzindo a uma temática sobre mulheres, história e vida, o que aparentemente enriquece e deixa menos imprecisa a indexação tradicional usando "Mulheres (psicologia)".

Alguns termos identificados pela abordagem de mineração de texto se aproximam, em semântica, dos termos usados no catálogo da biblioteca. É o caso, por exemplo, da obra OE4, para a qual a abordagem sugere uma temática voltada a "essencialismo", enquanto o indexador da biblioteca usa o descritor "Essência (filosofia)".

Vale ressaltar que, em tese, é esperado que as indexações usadas em bibliotecas sejam mais restritas em relação a índices descobertos por meio do processamento inteligente de textos de comentários *online*. De fato, as obras do catálogo da biblioteca utilizam, em geral, um vocabulário controlado, que exerce função de padronização da terminologia. É esperado, por conseguinte, que tal padronização implique em restrição de vocabulário e perdas semânticas ao expressar as obras (ver obra OE5, por exemplo). É nesse sentido que se argumenta em favor da complementação desses mecanismos por meio de percepção artificial.

4 Considerações finais

Este artigo apresentou uma abordagem de mineração de textos, que integra a terminologia coloquial extraída de comentários *online* sobre obras com a terminologia padrão utilizada nos sistemas de indexação de obras. O objetivo foi avaliar a aplicabilidade dessa técnica como ferramenta para auxiliar os profissionais da informação na representação semântica em Sistemas de Bibliotecas, em particular as de instituições de ensino. Os resultados indicam que a mineração de texto pode desempenhar um papel significativo para automaticamente enriquecer o poder semântico da indexação e, assim, melhorar a experiência do usuário quando essa busca por obras no acervo.

A análise dos dados demonstrou que a mineração de texto complementa a representação temática da informação de modo que seria inviável de ser reproduzida da mesma forma por meio manual. A comparação entre a indexação social e tradicional evidencia o quão mais completa e moderna se torna a informação, se comparada à que usa somente o vocabulário controlado.

Além disso, a pesquisa aponta para um cenário em que a indexação tradicional e a social podem coexistir de maneira complementar. Essa integração possibilita uma avaliação mais robusta dos termos representativos das obras, enriquecendo a representação da informação. A mineração de texto mostrou-se eficaz na identificação de padrões e na seleção de palavras-chave que capturam com precisão os temas das obras analisadas, facilitando a busca por elas realizada pelos usuários.

Em relação aos resultados obtidos e estudos realizados na literatura, nossos resultados corroboram observações de trabalhos anteriores que analisam o vocabulário de usuários em

ambientes digitais. Por exemplo, Rocha (2007) e Santini e Souza (2010) já discutem como os termos espontâneos usados em ambientes Web 2.0 tendem a se distanciar da terminologia controlada dos sistemas de indexação tradicionais. No presente estudo, observamos fenômeno similar: os termos extraídos dos comentários frequentemente recorrem a linguagem emocional, informal ou temática (e.g.: "leitura envolvente", "trama pesada"), não prevista nos descritores oficiais do catálogo bibliográfico.

Ainda, no que se refere à natureza metodológica, a estratégia de extrair vocabulário relevante a partir de comentários com base em medidas de ponderação, como o TF-IDF usado neste artigo, já foi aplicada em outros contextos, como recomendação de produtos Liu (2012) e na análise de opinião Pang (2008). Este estudo adapta essa abordagem ao domínio da organização do conhecimento bibliográfico, propondo que os termos extraídos dos comentários possam complementar o vocabulário técnico e ampliar o alcance semântico dos sistemas de busca, ja que, interpretando Rocha (2007) e Santini e Souza (2010), percebe-se uma distância entre a linguagem dos usuários e a dos indexadores formais, o que reforça a importância de estratégias como a nossa.

Em suma, a abordagem proposta se revela como um recurso promissor para os profissionais da informação, mesmo para aqueles sem formação aprofundada em Ciência de Dados. Espera-se que ela estimule avanços e amplie a perspectiva dos Sistemas de Bibliotecas das instituições de ensino federais de ensino em relação às inovações tecnológicas. Ao considerar a mineração de dados como um complemento às práticas existentes, podemos desenvolver um sistema híbrido que aproveita simultaneamente a indexação tradicional e a social (folksonomia), promovendo uma representação da informação mais eficaz e abrangente.

Esta pesquisa não chegou a avaliar os impactos práticos da adoção da terminologia minerada como complemento à original. A pesquisa também não alcançou a etapa de implementação em sistemas reais de busca. Estas, portanto, configuram limitações e caminhos possíveis para investigações futuras.

Notas

(1) 1 zettabyte = 1024 exabytes; 1 exabyte = 1 billion gigabytes.

(2) Estudos reportam que 90% dos dados existentes estão organizados de forma semiestruturada ou não estruturada, o que torna a busca e a representação do conhecimento mais complexas, especialmente quando armazenadas em sistemas de recuperação da informação (Zhang e Gu 2011). Uma das aplicações de AM que possui por natureza uma organização não estruturada dos dados advém de domínios em que dados são apresentados como texto. Na medicina, por exemplo, pode ser necessário a interpretação de texto em laudos médicos para o diagnóstico de doenças; na indústria, pode ser necessário interpretar comandos textuais de especificações esperadas para que se coordene automaticamente sistemas (e sistemas de sistemas); na agricultura, pode ser o caso que usuários e equipamentos se integrem mais naturalmente por meio de comandos de texto; dentre outros exemplos (Ferneda 2006; Pezzini 2017; Coneglian 2020; Srinivas *et al*. 2021).

(3) É importante reconhecer que nem sempre as palavras-chave escolhidas para representar uma obra constituem um vocabulário controlado em sua essência. Linguagens documentárias, como tesauros e sistemas de classificação, são sistemas simbólicos criados para facilitar a comunicação entre o usuário e o sistema. Nesse contexto, integram três elementos básicos: um léxico ou conjunto de descritores, uma rede lógico-semântica e uma rede sintagmática.

(4) A linguagem natural é intrinsicamente ambígua e, por isso, há diferentes tipos de ambiguidade a depender do nível de análise, tais como: ambiguidade lexical, ambiguidade sintática, ambiguidade de escopo e ambiguidade semântica. No contexto de indexação e busca de obras bibliográficas, o nível de análise recai sobre a ambiguidade lexical, isto é, na indeterminação de sentidos ou diferentes interpretações resultantes de um item lexical.

Referências

Abbott, A. "Seven Types of Ambiguity." *Theory and Society*, v. 26, n. 2/3, 1997, p. 357–91.

Aggarwal, C. C., and Zhai, C. Mining Text Data. New York, NY: Springer, 2012.

Aranha, C. N. Uma abordagem de pré-processamento automático para mineração de textos em português: sob o enfoque da inteligência computacional. 2007. Tese (Doutorado) – Universidade Católica do Rio de Janeiro (PUCRio), Rio de Janeiro.

Aranha, C. N. and Passos, E. A. "Tecnologia de Mineração de Textos". *Revista Eletrônica de Sistemas de Informação*, v. 5, n. 2, 2006, p. 1–20.

Brandt, M., and Medeiros, M. B. B. "Folksonomia: esquema de representação do conhecimento?". *Transinformação*, v. 22, no 2, 2010, p. 111–121.

Boccato, V. R. C. Avaliação de linguagem documentária em Fonoaudiologia na perspectiva do usuário: estudo de observação da recuperação da informação com protocolo verbal. 2005. Tese (Doutorado) – Faculdade de Filosofia e Ciências, Universidade Estadual Paulista, Marília.

Campos, L. F. B. "Web 2.0, biblioteca 2.0 e ciência da informação: um protótipo para disseminação seletiva de informação na Web utilizando mashups e feeds RSS". *In*: *Encontro Nacional de Pesquisa em Ciência da Informação*. Salvador: [s. n.], 2007. p. 232–245.

Catarino, M. E. and Baptista, A. A. "Folksonomia: um novo conceito para a organização dos recursos digitais na web". *DataGramaZero*, v. 8, n. 3, 2007.

Chierchia, G. Semântica. Editora da Unicamp e Editora da UEL: Campinas e Londrina, 2003.

Coneglian, C. S. *Recuperação semântica da informação com linguagem natural: a inteligência artificial na Ciência da Informação*. 2020. Tese (Doutorado em Ciência da Informação) – Universidade Estadual Paulista Júlio de Mesquita Filho, Marília, SP.

Feldman, R. and Sanger, J. The Text Mining Handbook: Advanced Approaches in Analyzing Unstructured Data. Cambridge, UK: Cambridge University Press, 2007.

Ferneda, E. "Redes neurais e sua aplicação em sistemas de recuperação de informação". *Ciência da Informação*, v. 35, n. 1, 2006.

Golub, K. "Automated Subject Indexing: An Overview". *Cataloging & Classification Quarterly*, v. 59, no 8, 2021, p. 702–719. https://doi.org/10.1080/01639374.2021.2012311. Acessado 25 jun. 2025.

Greenberg, R. and Bar-Ilan, J. "Library metrics – studying academic users' information retrieval behavior: A case study of an Israeli university library". *Journal of Librarianship and Information Science*, v. 43, n. 5, 2017, p. 662–675.

Guedes, R. M.; Moura, M. A. and Dias, E. J. W. "Indexação social e pensamento dialógico: reflexões teóricas". *Informação & Informação*, v. 16, n. 3, 2011, p. 40–59.

Jo, T. Text Mining: Concepts, Implementation, and Big Data Challenges. Cham, Switzerland: Springer Cham, 2024.

Kao, A. Natural Language Processing and Text Mining. Berlin, Germany: Springer-Verlag, 2007.

Kreutz, C. K., Blum, M., Schaer, P., Schenkel, R., and Weyers, B. "Evaluating digital library search systems by using formal process modelling". *arXiv preprint*, 2023.

Kuhar, M. and Mercun, T. "Exploring user experience in digital libraries through questionnaire and eye-tracking data". *Library & Information Science Research*, v. 44, n. 3, 2022. https://doi.org/10.1016/j.lisr.2022.101175. Acessado 25 jun. 2025.

Labov, W. Padrões sociolinguísticos. São Paulo: Parábola, 2008 [1972].

Labov, W. Principles of linguistic change: cognitive and cultural factors. Oxford: Wiley Blackwell, 2010.

Lancaster, F. W. Indexação e resumos: teoria e prática. 2. ed. [S. l.]: Briquet de Lemos, 2004.

Leskovec, J., Rajaraman, A., and Ullman, J. D. Mining of Massive Datasets. Cambridge, 2020.

Lima, G. A. B. O. Organização e representação do conhecimento e da informação na web: teorias e técnicas. Perspectivas em Ciência da Informação, v. 25, n. esp, 2020, p.1–25.

Liu, B. "Sentiment Analysis and Opinion Mining". In *Synthesis Lectures on Human Language Technologies*. Morgan & Claypool Publishers, 2012.

Li, S., Liu, F., Zhang, Y., Zhu, H., and Yu, Z. "Text mining of user-generated content (UGC) for business applications in e-commerce: A systematic review". *Mathematics*, v. 10, n. 19, 2022.

Leite, N. and Martins, D. "Automatic document indexing using terms extracted from highlights". *Ciência da Informação em Revista*, v. 11, n. 1, 2024.

Levy, S., Meaney, G., Wade, K., and Greene, D. *Curatr: A Platform for Semantic Analysis and Curation of Historical Literary Texts*. Springer International Publishing, p. 354–366, 2019.

Levy, S., O'Sullivan, D., and O'Connor, N. "CuratR: A platform for exploring and curating historical text corpora". In *Proceedings of the Digital Humanities Conference*, Riga: Latvia, 2023.

Lopes, M. V. and Albuquerque, A. C. "Folksonomias na organização e representação do conhecimento: um estudo na literatura da área". *Revista Brasileira de Biblioteconomia e Documentação*, v. 18, 2022, p. 1–25.

Lv, Z.; Song, H.; Basanta-Val, P.; Steed, A. and Jo, M. "Next-Generation Big Data Analytics: State of the Art, Challenges, and Future Research Topics". *IEEE Transactions on Industrial Informatics*, v. 13, n. 4, 2017, p. 1891–1899.

Madureira, J. C. M.; Santana, R. D. and Santos, E. D. J. dos. "Vocabulário controlado sobre bairros de Salvador: uma busca pela visibilidade afrocentrada na biblioteca da Faculdade de Arquitetura da UFBA". *Revista Fontes Documentais*, v. 4, 2022, p. 159–176.

Maimone, G. D. and Gracioso, L. S. "Representação temática de imagens: perspectivas metodológicas". *Informação & Informação*, v. 12, n. 1, 2007, p. 130–141.

Meira, S. R. de L.; Silva, E. M.; Costa, R. A. and Jucá, P. M. Folksonomia. https://sistemascolaborativos.uniriotec.br/folksonomia/. Acessado 23 jun. 2025.

Mitchell, T. M. Machine Learning. New York, NY: McGraw-Hill, 1997.

Brasil, Associação Brasileira de Normas Técnicas. NBR 12676: Métodos para análise de documentos – Determinação de seus assuntos e seleção de termos de indexação. [S. l.]: ABNT, 1992. p. 1–4.

Nunes, C. O. "Algumas considerações acerca da ausência de políticas de indexação em bibliotecas brasileiras". *Biblos*, Rio Grande, v. 16, 2004, p. 55–61.

Pang, B. and Lee, L. "Opinion mining and sentiment analysis". *Foundations and Trends® in Information Retrieval*, v. 2, no 1, 2008, p. 1–135.

Pezzini, A. "Mineração de textos: conceito, processo e aplicações". *Revista Brasileira de Contabilidade e Gestão*, v. 5, n. 10, 2017, p. 58–61.

Rocha, P. R. "Folksonomia e a web 2.0: um estudo sobre o etiquetamento social". In *Anais do VI Seminário Nacional de Informação e Comunicação,* Brasília, DF, 2007.

Santini, M. L. and Souza, A. C. "Folksonomia: uma reflexão sobre a organização da informação na Web 2.0." *Perspectivas em Ciência da Informação,* v. 15, n. 2, 2010, p. 5-19.

Santos, H. S.; Oliveira, J. R. and Lima, J. S. "Folksonomia: representação da informação na web". *Revista Bibliomar*, v. 16, n. 1, 2017, p. 105–114.

Santos, L. B. P. Política de indexação em bibliotecas universitárias: estudo diagnóstico na região de Marília. 2011.

Santos, R. F. and Corrêa, R. F. "Análise das definições de Folksonomia: em busca de uma síntese". *Perspectivas em Ciência da Informação*, v. 23, n. 2, p. 1–32. 2018.

Santos, R. R. et al. Fundamentos de Big Data. Porto Alegre: Grupo A e Editora Sagah, 2021.

Srinivas, M; Sucharitha, G.; Matta, A. and Chatterjee, P. Machine Learning Algorithms and Applications. Hoboken: John Wiley & Sons Ltd, 2021.

Tabosa, H. R.; Souza, O.; Candido, J. C. S.; Melo, A. C. A. U.; Reis, K. G. B. and Souza, O. "Avaliação do desempenho de um software de sumarização automática de textos". *Informação & Informação*, v. 25, n. 1, 2020, p. 189–210.

Tuffery, S. "Natural Language Processing". In: *DEEP Learning:* From Big Data to Artificial Intelligence with R. Hoboken, NJ: Wiley, 2023. p. 275–299.

Vaidya, P. and Harinarayana, N. S. "The role of social tags in web resource discovery: an evaluation of ER-generated keywords". *Annals of Library and Information Studies* v. 63, 2016. p. 289-297.

Vajjala, S., Majumder, B., Gupta, A. and Surana, H. Practical Natural Language Processing: A Comprehensive Guide to Building Real-World NLP Systems. O'Reilly Media, 2020.

Vignoli, R. G., Almeida, P. O. P. and Catarino, M. E. "Folksonomias como ferramenta da organização e representação da informação". *Revista digital de biblioteconomia & ciência da informação,* v. 12, n. 2, 2014

Wal, T. W. Folksonomy. 2007. http://www.vanderwal.net/folksonomy.html. Acessado em: 23 jun. 2025.

Zhang, Y. and Gu, H. "Text Mining with Application to Academic Libraries". In: Yu, Y.; Yu, Z.; Zhao, J. (ed.). *Computer Science for Environmental Engineering and EcoInformatics*. Berlin, Heidelberg: Springer, 2011.

Zhang, T., Niu, X., and Promann, M. "Assessing the User Experience of E-Books in Academic Libraries." *College & Research Libraries*, v. 78, n.5, 2017. https://doi.org/10.5860/crl.78.5.578. Acessado 25 jun. 2025.

Submetido: 12/05/2025 Aceito: 05/07/2025