



EIXO TEMÁTICO:
Organização e Representação da Informação e do Conhecimento

MINERAÇÃO DE DADOS: UMA PERSPECTIVA DE ANÁLISE TEXTUAL NA REPRESENTAÇÃO TEMÁTICA DA INFORMAÇÃO

DATA MINING: A PERSPECTIVE OF TEXTUAL ANALYSIS IN THE THEMATIC REPRESENTATION OF INFORMATION

Jaqueline Rodrigues de Jesus (IBICT) - jaque1906@gmail.com
Márcio Bezerra da Silva (UnB) - marciobdsilva@unb.br

Resumo: Apresenta o uso da mineração de dados na representação temática da informação. Discute-se, na fundamentação teórica, a representação temática da informação, pontuando a indexação tradicional e a folksonomia, e a mineração de dados, elencando a versão textual. Objetiva-se analisar o uso da mineração de dados como ação de contribuição à tomada de decisão na representação temática da informação. Constitui-se de um estudo exploratório, de natureza aplicada, que utiliza a técnica de pesquisa bibliográfica, com abordagem de coleta de dados quantitativa e qualitativa, e que adota a ferramenta de mineração de texto *Kitconc* 3.0 como instrumento de coleta de dados para analisar os comentários mais recentes e definidos como úteis pelos usuários em oito livros (aleatórios) disponibilizados no ambiente de comércio eletrônico *Amazon*; em seguida definir as palavras-chaves de cada obra e confrontá-las com os termos indexadores dos mesmos livros no catálogo da Biblioteca Central da Universidade de Brasília (BCE/UnB). Apresenta, como resultados da pesquisa, que a mineração de texto contribuiu de forma significativa e fidedigna à descoberta de padrões para a representação temática da informação a partir dos termos que se apresentaram com frequência e possuíam concordância semântica no texto, tornando-os posteriormente candidatos às palavras-chave que permitirão definir os assuntos das obras. Resulta também, por outro lado, que a mineração de texto não identificou padrões suficientes em duas obras, impossibilitando realizar o confronto entre os dois ambientes de pesquisa.

Palavras-chave: Representação temática da informação. Folksonomia. Mineração de dados. Mineração de texto.

Abstract: Presents the use of data mining in the thematic representation of information. Discusses, in the theoretical foundation, the thematic representation of information, punctuating traditional indexing and folksonomy, and data mining, listing the textual version. The objective of this study is to analyze the use of data mining as a contribution to decision making in the thematic representation of information. It is an exploratory study, of an applied nature, which uses the bibliographic research technique, with a quantitative and qualitative data collection approach, and adopts the text mining tool *Kitconc* 3.0 as a data collection tool to analyze the Most recent reviews and defined as helpful by users in eight books (random) made available in the *Amazon* e-commerce environment; in sequence to define the

keywords of each work and compare them with the indexing terms of the same books in the catalog of the Central Library of the University of Brasília (CL/UnB). Results that the mining of text contributed in a significant and reliable way to the discovery of standards for the thematic representation of the information from the terms that presented frequently and had semantic agreement in the text, becoming later candidates to the keywords that will allow to define the subjects of books. It is also found, on the other hand, that text mining did not identify enough patterns in two books, making it impossible to carry out the confrontation between the two research environments.

Keywords: Thematic representation of information. Folksonomy. Data Mining. Text Mining.

1 INTRODUÇÃO

As recentes mudanças tecnológicas fizeram com que a sociedade alterasse seu comportamento. Um indivíduo ao realizar, de forma cotidiana e instintiva, multitarefas no computador, acessando conteúdos diferentes e em ambientes diversos, insere-se em um contexto de consumo, produção e disseminação de um grande volume de informação, em um curto espaço de tempo. Entretanto, este indivíduo, atualmente, encontra-se conectado digitalmente, fazendo uso não mais apenas dos tradicionais computadores, conhecidos como *personal computer* (PC), mas também dos *tablets* e *smartphones*, promovendo ainda mais a supracitada produção de informação, independentemente da localização física. A citada produção é acessada e analisada pelas empresas e Instituições na perspectiva de dados, para a tomada de decisão, e assim proporcionar vantagens e lucros sobre seus produtos e serviços. Com base na compreensão de Somasundaram e Shrivastava (2011, p. 21), “[...] enormes quantidades de informações digitais são criadas a todo o momento por consumidores individuais e corporativos de TI¹. Esses dados precisam ser armazenados, protegidos, otimizados e gerenciados”.

Pela quantidade de informações que veem sendo geradas e usadas na *web*, torna-se importante a necessidade de organizar essa informação com fins de recuperação e uso. A referida ação, de ordenar a informação, foi sendo moldada conforme a evolução da *web*, ao qual vem permitindo a participação na representação e organização da informação (ROI) pelos usuários. Este cenário vai ao encontro da chamada *web 2.0*, fase em que as páginas da Internet assumem as características “[...] colaborativa por natureza, interativa, dinâmica, e a linha entre criação e consumo de conteúdo nesses ambientes era tênue (usuários criavam o conteúdo nesses sites tanto quanto eles o consumiam)” (MANESS, 2007, p, 43). No

¹ Tecnologia da Informação.

cerne desses predicados está a folksonomia, ou seja, segundo Catarino e Baptista (2007, p. 3, grifos das autoras), “[...] o resultado da atribuição livre e pessoal de etiquetas (*tagging*) a informações ou objetos (qualquer coisa com *URL*²), visando à sua recuperação. A atribuição de etiquetas é feita num ambiente social (compartilhado e aberto a outros)”.

A folksonomia retrata a sabedoria da sociedade, diferentemente do que acontece com a indexação enquanto um ato de representação temática da informação (RTI) na perspectiva das bibliotecas físicas. Enquanto a primeira ocorre a partir dos termos atribuídos pelas pessoas na *web*, a segunda é realizada segundo a subjetividade dos profissionais da informação, como os bibliotecários. Enquanto a folksonomia retrata o saber popular, a indexação volta-se ao conhecimento do profissional.

Ao verificar um grande volume de usuários na *web* torna-se recomendável identificar possíveis padrões quanto às formas que estes produzem, consomem, disseminam e compartilham informações com fins de apresentar melhores sugestões nos processos de busca, como também auxiliar na definição de termos representativos (indexadores) aos mais variados produtos. Chega-se ao discurso de que “[...] é na descoberta desses padrões que se encontra a chave para conhecer melhor as comunidades de usuários dos serviços oferecidos pelas bibliotecas, com o objetivo de melhor adequar estes serviços aos diferentes grupos de usuários” (NICHOLSON, 2004, p. 254).

A mineração de dados afigura-se como uma técnica de análise que pode ser utilizada no universo biblioteconômico, mais especificamente na RTI e na perspectiva textual, confrontando termos atribuídos na folksonomia com palavras-chave definidas na indexação tradicional. Assim, a biblioteconomia pode identificar novas oportunidades de atuação, especialmente nos ambientes digitais, e aproveitá-las a fim de conquistar mais usuários, conhecendo-os e tratando-os de forma singular, ou seja, segmentando-os conforme suas preferências e indicando os materiais que mais se adequem às necessidades informacionais segundo os termos atribuídos pelos próprios ao representar algum objeto, entre livros e demais materiais especiais.

² *Uniform Resource Locator*.

Na eminência de um problema de pesquisa centrado em uma possível aproximação entre a mineração de dados e a indexação, o presente trabalho objetivou analisar o uso da mineração de dados como ação de contribuição à tomada de decisão na RTI, a partir de um percurso metodológico caracterizado como um estudo exploratório, de natureza aplicada, e que utiliza a técnica de pesquisa bibliográfica, com abordagem de coleta de dados quantitativa e qualitativa, e fazendo uso da ferramenta de mineração de texto *Kitconc 3.0*³ como instrumento de coleta de dados, que por sua vez foram extraídos de oito livros (aleatórios) presentes no comércio eletrônico *Amazon* e no catálogo da Biblioteca Central da Universidade de Brasília (BCE/UnB).

2 DESENVOLVIMENTO TEÓRICO

Apresentando os termos de indexação como seus produtos, a RTI objetiva-se ao “[...] conteúdo informacional dos documentos e permite à identificação do tema ou do assunto a que se refere através das ações de indexação, elaboração de resumos, classificação, disseminação, busca e recuperação” (DA SILVA; NEVES, 2013, p. 32). A referida ação, realizada em ambientes digitais da *web 2.0*, tem o seu rótulo transcrito para indexação colaborativa, enquanto uma ação proporcionada pelos usuários desses ambientes, ou seja, a prática da folksonomia.

As resultantes desta representação podem ser usadas pelos bibliotecários na definição de palavras-chave, chamada na *web 2.0* de tagueamento (etiquetagem), nas obras presentes nos acervos das bibliotecas, por exemplo. Neste caso, a indexação social não deve ser vista como uma forma de exclusão do tratamento temático da informação tradicional, mas uma ação complementar, agregando valor à linguagem dos usuários conforme características culturais, sociais e temporais. Realiza-se então a folksonomia, termo formado pelas palavras em inglês *folk* (povo) e *taxonomy* (taxonomia), criada por Thomas Vander Wal no ano de 2004, em associação a um grupo de arquitetura da informação (AI) e correspondente a forma de representar e organizar a informação na *web 2.0*. A expressão folksonomia refere-se ao processo de representar e organizar a informação como um todo, ou seja, “[...] um sistema de indexação de informações que permite a adição de *tags* (etiquetas) que descrevem o conteúdo dos documentos armazenados” (AQUINO,

³ Versão de demonstração (DEMO): <http://www.corpuslg.org/>

2007, p. 3).

Sobre a folksonomia, Gouvêa e Loh (2007, p. 3) entendem que é possível “comparar padrões de uso individual e coletivo das *tags*. Uma suposição é que há diferença, ou seja, *tags* usadas por várias pessoas (coletivamente) seguem um padrão diferente das *tags* individuais (usadas por somente uma pessoa)”. Neste sentido faz-se necessário estudar a forma de representação supracitada por meio da identificação de padrões de uso, pois o usuário, ao representar algo conforme seu cognitivo, teoricamente, garantirá a linguagem do público do ambiente.

A identificação de padrões trata de uma ação que nos permite conhecer mais sobre os gostos, preferências e ações dos usuários no sentido de quantificar, analisar e tomar decisões frente a linguagem natural inerente aos ambientes colaborativos. Entre as ferramentas que podem ser usadas para observar as preferências dos usuários, com fins de análise e tomada de decisão, encontra-se a mineração, comumente conhecida como mineração de dados (*data mining*).

A mineração de dados surgiu na área empresarial, especialmente nos Estados Unidos da América (EUA), por volta dos anos 1990, quando os repositórios de dados se popularizaram, armazenando grandes volumes de dados, que por sua vez eram ferramentas para o desenvolvimento de estratégias competitivas. Para Sferra e Corrêa (2003, p. 20), “[...] Data Mining define o processo automatizado de captura e análise de grandes conjuntos de dados para extrair um significado, sendo usado tanto para descrever características do passado como para prever tendências para o futuro”.

Além da abordagem tradicional da mineração, com enfoque nos dados, existe também uma análise que leva em consideração os textos presentes nas realidades em estudo. Cada vez mais usado nos espaços da *web*, a mineração de textos, de acordo com Aranha e Passos (2006, p. 2), é “[...] um conjunto de métodos usados para navegar, organizar, achar e descobrir informações em bases textuais”. Na *web*, por exemplo, os textos provenientes dos usuários, em linguagem natural, são analisados para recomendar opções de busca além da realizada pelo próprio, apresentando uma abordagem de mineração de dados na *web* (*web mining*).

Também chamada por mineração de dados textuais ou descoberta de conhecimento de dados textuais, a mineração em questão busca investigar textos a partir de sua semântica ou estatisticamente (ocorrência de palavras), que podem ser utilizadas juntas, ou não, dependendo do objetivo da análise. Enquanto a análise

semântica busca o conhecimento a partir do exame dos termos quanto aos seus significados, morfologia, sintática, a própria semântica, a pragmática e o contexto que envolve o texto, a análise estatística quantifica os números de ocorrência dos termos no texto sem se preocupar com as características da análise semântica.

A mineração de textos utiliza-se de várias áreas do conhecimento: a informática, incluindo a recuperação de informação, o aprendizado de máquina e a inteligência computacional; a estatística; a linguística; e a ciência cognitiva. O conjunto destas áreas subsidiará, na mineração, extrair dados; resumir textos; recuperar informações; descobrir padrões, associações e regras; e analisar, de maneira qualitativa e quantitativa, um documento em texto.

Como substância à extração de elementos na mineração de dados, o método de reconhecimento de entidades nomeadas (REN) pode contribuir na análise ao identificar entidades do tipo: pessoa, lugar, organizações, quantidades etc. Em um levantamento teórico dissertativo, Amaral (2013) exemplifica os tipos mais comuns de entidades a partir da seguinte categorização de elementos, a saber:

- Substantivos próprios: nomes de pessoas e organizações/entidades;
- Substantivos temporais: datas, tempo, dia, ano e mês;
- Entidades numéricas: medições, percentagens e valores monetários.

Quanto ao objetivo, REN é um “tipo de extração de informação que visa identificar regiões do texto (menções) correspondentes a entidades e categorizá-las numa lista pré-determinada de tipos (entidades de interesse)” Para (SILVA; CASELI, 2015). Em suma, o método REN objetiva “[...] identificar as entidades nomeadas bem como sua posterior classificação, atribuindo uma categoria semântica para essas entidades” (AMARAL, 2013, p. 34).

A dificuldade na identificação e classificação de nomes próprios em textos como, por exemplo, Rio de Janeiro como uma localização, Miriam como uma pessoa e Ministério da Educação e Ministério da Cultura como organizações podem ser minimizados pela técnica de REN (SUTTON; MCCALLUM, 2006), pois ajudaria a descobrir se o texto aborda a capital ou a cidade do Rio de Janeiro, quem é a pessoa chamada Miriam e que os citados Ministérios são a mesma organização, ou seja, Ministério da Educação e Cultura (MEC).

A partir dos conceitos levantados, a mineração de textos pode contribuir em áreas como a do direito e da medicina, na geração de resumos e

relatórios/prontuários, pois agilizará a indexação e a recuperação destes documentos (CARRILHO JUNIOR, 2003). Essas características vão ao encontro da CI, principalmente na RTI, enquanto um campo que trabalha com dados não estruturados e, por vezes, na linguagem natural, presentes nos documentos e apresentados na forma textual como em resumos, informativos, relatos, notas etc.

3 RESULTADOS: APRESENTAÇÃO E DISCUSSÃO

Foram escolhidas oito obras aleatórias e disponíveis no catálogo da *Amazon* e que possuísem, no mínimo, três comentários de usuários em cada. Em seguida, conferimos se os mesmos livros se encontravam também no catálogo da BCE/UnB. Realizada esta verificação, os três comentários mais recentes e definidos (marcados) como úteis pelos próprios usuários foram salvos em um arquivo no formato texto (txt)⁴ para cada livro. Para tanto, usamos o programa intitulado bloco de notas, um processador de textos do sistema operacional (SO) *Windows*, enquanto que para extrair as palavras-chaves dos arquivos e realizar a mineração de textos utilizamos o *software Kitconc*, na versão 3.0.

Os livros escolhidos no catálogo da *Amazon* e analisados foram:

- OE1 – Os cinco porquinhos / Agatha Christie / 1984;
- OE2 – O iluminado / Stephen King / 1984;
- OE3 – Mentres perigosas / Ana Beatriz Barbosa Silva / 2003;
- OE4 – O mundo de Sofia / Jostein Gaarder / 1991;
- OE5 – Memórias póstumas de Brás Cubas / Machado de Assis / 1999;
- OE6 – Orgulho e preconceito / Jane Austen / 1997;
- OE7 – O pequeno príncipe / Antoine de Saint-Exupéry / 2003;
- OE8 – Por quem os sinos dobram / Autor: Ernest Hemingway / 1969.

Conforme o critério metodológico adotado, os textos coletados são especificamente os comentários referentes a avaliação “muito bom” de cada obra. Em seguida, foram realizadas as seguintes etapas, constituintes do software *Kitconc*:

1. **Formação do “corpus”:** os (três) comentários selecionados de cada livro foram transformados em arquivos txt, salvo com o nome da obra. A partir

⁴ Foram gerados oito arquivos txt.

da aba “corpus”, o *software* lista as palavras presentes no arquivo para que seja iniciada a mineração;

2. **Formação da lista de palavras:** para identificar os termos encontrados nos textos inseridos na etapa anterior, acessamos a aba “lista de palavras”, onde foram classificadas todas as palavras quanto a sua frequência de apresentação e um valor em porcentagem (peso) referente à quantidade da citada frequência. A aba em discussão, após coletar e analisar os textos, identifica a presença de palavras distintas, incluindo as *stopwords*⁵;
3. **Identificação das palavras consideradas mais importantes:** na aba “palavras-chave” é encontrado o total de palavras e identificadas as mais importantes a partir da classificação de “chavicidade”, ao qual gera um ranking de palavras-chave. Porém salientamos que nem todas as palavras-chave podem ser classificadas como um assunto, independentemente de a “chavicidade” ser alta, pois na aba “concordância” foi possível analisar os contextos em que a palavra no “corpus” aparece, além de pontos como traduções idiomáticas, significados e fonéticas distintas;
4. **Recorte final:** ao final realizamos uma série de recortes para definir os assuntos:
 - a. Entre as palavras-chave com valor de “chavicidade” de um a oito;
 - b. Foram definidas até cinco (1 a 5) palavras-chave com maior concordância ao assunto de cada livro;
 - c. Por fim, selecionamos as três (1 a 3) palavras-chaves identificadas como as mais importantes nos comentários de cada livro.

Os resultados de cada etapa da análise de mineração, a partir do uso do *software* Kitconc, podem ser observadas no quadro 1:

Quadro 1: Análise de corpus de cada livro.

| Obras | Palavras-chaves (chavicidade de 1 a 8) | Concordâncias (1 a 5) | Palavras-chaves (1 a 3) | Assuntos |
|-------|---|--------------------------|----------------------------|----------|
|-------|---|--------------------------|----------------------------|----------|

⁵ Em uma tradução livre significa palavras paradas, ou seja, que não contenham significado para substantivos. Por exemplo: artigos, até, “tbm”, “vc” etc.

| | | | | |
|-----|--|----------------------------|------------------|-----------|
| OE1 | recomendo; christie; livro; sempre; você; final. 15 palavras | -- | - | - |
| OE2 | livro; gostar; king; adaptação; filme; realmente; opinião; obra. 24 palavras | -- | - | - |
| OE3 | mente; livro; interessante; bem; eu; muito; para; não. 17 palavras | mente | mente | mente |
| OE4 | filosofia; história; inicio; mundo; 11 palavras | história; inicio; mundo | filosofia; mundo | filosofia |
| OE5 | cubas; machado; brás; sterne; póstumas; casmurro; ironista; jamais. 58 palavras | póstumas; ironista | ironista | ironia |
| OE6 | austen; clássico; jane; livro. 19 palavras | clássico | clássico | clássico |
| OE7 | desperta; obra; olhar; através. 16 palavras | desperta; obra; olhar | - | - |
| OE8 | hemingway; livro; ganharia; fortes; romance. 30 palavras | romance; livro; fortes | romance | romance |

Fonte: Da pesquisa, 2017.

Apesar dos termos oriundos dos itens “chavicidade” e “concordância”, as definições das palavras-chave ocorreram conforme uma possível representação temática da obra, listados na coluna “Palavras-chave (1 a 3)” do quadro dois (1). Neste sentido, entre os resultados obtidos pela mineração de texto, realizada pelo software *Kitconc*, inferimos que o resultado da análise da obra OE1 não foi positivo ao apresentar somente 15 palavras entre elas, termos “recomendo”, “christie”, “livro” como palavra-chave do livro e assim não permitindo a identificação do assunto, pois trata-se de uma ficção policial. Os livros OE2 e OE7 não obtiveram resultados positivos, pois não restaram palavras que pudessem representar o assunto principal da obra. No caso da obra OE3 e OE4 obtivemos “filosofia” e “mente” como palavras-chaves, respectivamente, sendo estas as que possuem mais referência ao conteúdo da obra, tratando-se de uma literatura baseada na filosofia e de indivíduos que possuem psicopatologias, sendo este um transtorno mental. Quanto ao livro OE5,

apesar do “corpus” ser grande, não obtivemos uma análise positiva quanto ao descobrimento de palavras-chave que representassem o conteúdo do livro, mesmo com as “chavidades” altas, apresentando somente uma característica da literatura “ironia”. Na análise do “corpus” da obra OE6 encontramos um resultado que não havíamos previsto a priori, pois obtivemos algumas palavras que contêm características (qualificadores) da obra como o termo “clássico”, mas que não pode ser adotada como palavra-chave representativa dos assuntos do livro. O conteúdo da OE8 trata-se de um romance que acontece em um período de guerra indo ao encontro da análise realizada, ao qual foi possível verificar a palavra-chave “romance”, como representante do assunto da obra. Em suma, a partir das palavras-chave definidas para cada obra foi possível perceber que os materiais analisados tratam de temáticas relacionadas ao assunto literatura em consonância com os resultados dos livros consultados no catálogo da BCE/UnB (quadro 2).

Analisados os livros supracitados e obedecendo aos objetivos da presente pesquisa foi elaborado um quadro comparativo entre as palavras-chave (*tags*) adquiridas no *website* da *Amazon*, representando o ambiente da *web 2.0* e que faz uso da folksonomia, com os descritores apresentados no catálogo da BCE/UnB para representar os livros rotulados pela indexação tradicional, comumente adotada em bibliotecas físicas.

Quadro 2: Comparação de palavras-chave entre os ambientes.

| Obras | Amazon | BCE/UnB |
|-------|-----------|---|
| OE1 | - | Literatura inglesa; Ficção inglesa |
| OE2 | - | Literatura americana; Ficção americana |
| OE3 | Mente | Psicologia – doentes mentais; Psicologia – comportamento humano; Psicopatologia |
| OE4 | Filosofia | Literatura norueguesa |
| OE5 | -- | Literatura brasileira; Romance |
| OE6 | Clássico | Literatura inglesa; Romance |
| OE7 | - | Literatura francesa; ficção francesa; literatura infanto-juvenil |
| OE8 | Romance | Literatura americana |

Fonte: Da pesquisa, 2017.

No quadro 2, podemos perceber que a maioria dos termos elencados pelo *software*, a partir da sabedoria dos usuários nos comentários das obras no ambiente *Amazon*, não são adotadas como palavras-chave representativas dos assuntos de cada obra. Entretanto estas *tags* possuem as qualidades (características) de apresentarem os significados a respeito do conteúdo dos livros, permitirem uma navegação intuitiva entre os materiais segundo as *tags* oriundas da indexação social, oferecerem recomendações de outros livros semelhantes ao pesquisado e subsidiarem os usuários no momento da escolha do livro a partir de avaliações/comentários feitos pela coletividade, diferentemente dos catálogos tradicionais.

Neste sentido, apesar de alguns termos não serem definidos como palavras-chave, os mesmos são fundamentais para a descoberta do assunto de cada obra, situação que se deflagra na obra OE8, pois a indexação tradicional apresentou somente um único descritor, “literatura americana”, enquanto que a mineração resultou em duas palavras que representam o conteúdo da respectiva obra, ou seja, aferimos que trata-se de uma obra de literatura americana, especificamente um romance em meio ao decorrer de uma guerra, nos permitindo inferir que o conhecimento sobre o conteúdo da obra pode ser descoberto a partir de palavras identificadas na mineração de texto. Algumas palavras-chave definidas na mineração de texto se aproximam de termos presentes no catálogo da BCE/UnB, como é o caso da obra OE3, ao verificarmos que, na indexação social, o termo que se aproxima de um assunto é “mente”, enquanto que na versão tradicional da indexação, o descritor adotado é “psicopatologia”. Versão tradicional da indexação, o descritor adotado é “psicopatologia”.

Por outro lado, os descritores das obras selecionadas no catálogo da BCE/UnB aparentemente foram retirados de uma taxonomia e/ou sistema de classificação, como a CDD, instrumentos de representação e organização que apresentam-se como um conjunto de descritores com o objetivo de recuperação, visando somente à informação registrada em seu ambiente físico, ou seja, representando a classe que indica o endereço das obras na estante e, por vezes, não indicando o assunto com riqueza de detalhes. Nos comentários da obra OE8, por exemplo, foram encontradas, na mineração de textos, palavras-chave que são assuntos do conteúdo do livro, mas que não foram contempladas na indexação do

catálogo da BCE/UnB, nos permitindo compreender que consultar unicamente a taxonomias, por vezes, pode ser uma ação limitada no processo de decisão quanto aos termos representativos ao conteúdo da obra.

4 CONSIDERAÇÕES FINAIS OU PARCIAIS

Com fins de descobrir de que forma a mineração de dados pode contribuir na representação temática da informação, partindo do pressuposto de que a tecnologia de mineração é utilizada para analisar e descobrir conhecimentos que possam agregar valores estrategicamente para os objetivos estabelecidos, a priori, em grandes quantidades de dados armazenados em BD. Neste sentido, no interesse em confrontar a indexação social, realizada no *website* da *Amazon*, com a indexação tradicional, contemplada no catálogo da BCE/UnB, uma amostragem foi constituída. Percebemos que os resultados obtidos na mineração de texto foram mais fidedignos do que os apresentados no catálogo da BCE/UnB. Direcionando os resultados para as grandezas inversamente proporcionais da indexação, nos parece que os ambientes da *web 2.0* promulgam pela revocação, especialmente pelas iniciativas de navegação e recomendação, enquanto que o catálogo da BCE/UnB adota a precisão como grandeza, ao levar em consideração o tempo do leitor na busca pela informação necessitada, mesmo que as ocorrências não sejam satisfatórias.

Apesar de vislumbramos certa contribuição da mineração na representação temática da informação, vale destacar uma dificuldade/limitação pertinente ao uso do *software Kitconc* por não disponibilizar alguns recursos em sua versão gratuita (DEMO), neste caso, a aba “dispersão”, que nos permitiria visualizar imagetivamente as posições dos termos com mais frequência ao longo do texto.

Conclui-se que os comentários obtidos na *web 2.0* apresentam o conhecimento dos usuários que usufruíram da obra e explicitaram sua sabedoria em postagens, aqui transformadas em palavras-chave, mostrando que os mesmos podem participar efetivamente na representação temática da informação e contribuir posteriormente no sucesso dos processos de busca, nos permitindo inferir que para uma possível representação fidedigna do conteúdo da obra e melhor recuperação, a indexação social pode subsidiar/complementar a indexação tradicional, ou seja, as *tags* dos usuários, analisadas pela mineração, podem ser comparadas aos

descritores definidos pelos bibliotecários para a tomada de decisão quanto aos termos representativos das obras.

Esperamos que este trabalho fomente novas pesquisas e amplie o olhar do bibliotecário frente às recentes inovações tecnológicas, a partir do uso de métodos de análise, como o REN, na mineração de dados, por exemplo, não as estudando como substitutivas, mas como um aperfeiçoamento/complemento às práticas consolidadas, como um sistema híbrido que faz uso concomitante das indexações tradicional e social (folksonomia).

REFERÊNCIAS

- AMARAL, D. O. F. **O reconhecimento de entidades nomeadas por meio de conditional random fields para a língua portuguesa**. 2013. 100 f. Dissertação (Mestrado) - Curso de Ciência da Computação, Pontifícia Universidade Católica do Rio Grande do Sul, Rio Grande do Sul, 2013. Disponível em: <<http://tede2.pucrs.br/tede2/bitstream/tede/5246/1/457280.pdf>>. Acesso em: 06 jul. 2017.
- AQUINO, M. C. Hipertexto 2.0, folksonomia e memória coletiva: um estudo das tags na organização da web. **E-Compós**, Brasília, DF, v. 9, ago. 2007. Disponível em: <http://www.compos.org.br/files/15ecompos09_MariaClaraAquino.pdf>. Acesso em: 13 ago. 2014.
- ARANHA, C.; PASSOS, E. A tecnologia de mineração de textos. **Revista Eletrônica de Sistemas de Informação**, v. 5, n. 2, 2006. Disponível em: <<http://189.16.45.2/ojs/index.php/reinfo/article/view/171>>. Acesso em: 28 dez. 2014.
- CARRILHO JUNIOR, J. R. **Desenvolvimento de uma metodologia para mineração de textos**. Orientador: Emmanuel Piseces Lopes Passos. – 2007. 96f.; 30 cm. Dissertação (Mestrado em Engenharia Elétrica) – Pontifícia Universidade Católica do Rio de Janeiro, Rio de Janeiro, 2007. Disponível em: <http://www.maxwell.vrac.puc-rio.br/Busca_etds.php?strSecao=resultado&nrSeq=11675@1&msg=28#>. Acesso em: 15 dez. 2014.
- CATARINO, M. E., BAPTISTA, A. A. Folksonomia: um novo conceito para a organização dos recursos digitais na Web. **DataGramZero**, v. 8, n. 3, 2007. Disponível em: <<http://basessibi.c3sl.ufpr.br/brapci/index.php/article/download/7548>>. Acesso em: 31 maio 2017.
- DA SILVA, M.; NEVES, D. A. B. Estudo sobre o uso da teoria da classificação facetada em banco de dados. In: ENCONTRO NACIONAL DE PESQUISA EM PÓS-GRADUAÇÃO EM CIÊNCIA DA INFORMAÇÃO, 11. 2010, **Anais...** Florianópolis: UFSC, 2010. p.1 - 20. Disponível em:

<<http://repositorios.questoesemrede.uff.br/repositorios/handle/123456789/870>>.
Acesso em: 21 nov. 2014.

GOUVÊA, C.; LOH, S. Folksonomias: identificação de padrões na seleção de tags para descrever conteúdos. **Revista Eletrônica de Sistemas de Informação**, v. 6, n. 2, 2007. Disponível em: <<http://189.16.45.2/ojs/index.php/reinfo/article/view/214>>.
Acesso em: 11 out. 2014.

MANESS, J. Teoria da biblioteca 2.0: web 2.0 e suas implicações para as bibliotecas. **Informação & Sociedade: Estudos**, v. 17, n. 1, 2007. Disponível em: <<http://periodicos.ufpb.br/ojs/index.php/ies/article/view/831>>. Acesso em: 25 nov. 2016.

NICHOLSON, S.. O processo da bibliomineração: repositório de dados e mineração de dados para tomada de decisão em bibliotecas. **Transinformação**, v. 16, n. 3, 2012. Disponível em: <<http://periodicos.puc-campinas.edu.br/seer/index.php/transinfo/article/view/712>>. Acesso em: 12 dez. 2016.

PRIMO, A. O aspecto relacional das interações na Web 2.0. **E-Compós (Brasília)**, v. 9, p. 1-21, 2007. Disponível em: <<http://www.compos.org.br/seer/index.php/e-compos/article/viewArticle/153>>. Acesso em: 23 nov. 2016.

SFERRA, H. H.; CORRÊA, A. M. C. J. Conceitos e aplicações de data mining. **Revista de ciência & tecnologia**, v. 11, n. 22, 2003. Disponível em: <<http://www.unimep.br/phpg/editora/revistaspdf/rct22art02.pdf>>. Acesso em: 10 jan. 2016.

SILVA, L. H.; CASELI, H. M. **Reconhecimento de entidades nomeadas em textos em português do Brasil no domínio do e-commerce**. 2015. p.1-7. Disponível em: <<http://www.lbd.dcc.ufmg.br/colecoes/tilic/2015/010.pdf>>. Acesso em: 06 jul. 2017.

SOMASUNDARAM, G; SHRIVASTAVA, A. **Armazenamento e Gerenciamento de Informações: Como armazenar, gerenciar e proteger informações digitais**. Bookman, 2011.

SUTTON, C.; MCCALLUM, A. An introduction to conditional random fields for relational learning. 2006. In: GETOOR, L.; TASKAR, B. **Introduction to Statistical Relational Learning**. MIT Press, 2006.