em questão
E-ISSN 1808-5245

Mineração de textos científicos: análise de artigos de periódicos científicos brasileiros da área de Ciência da Informação

Márcio Henrique Wanderley Ferreira

Mestre; Universidade Federal de Pernambuco, Recife, PE, Brasil

marcio.ferreira@ufpe.br

Renato Fernandes Correa

Doutor; Universidade Federal de Pernambuco, Recife, PE, Brasil

renato.correa@ufpe.br

Resumo: Este trabalho analisa artigos de periódicos científicos brasileiros da área de Ciência da Informação sobre a mineração de textos e descoberta de conhecimento em textos. Os artigos analisados são indexados pela Brapci ou Scielo e contêm os termos compostos mineração de texto ou descoberta de conhecimento nos metadados, além de abordarem especificamente a aplicação de técnicas de mineração de textos. A metodologia da pesquisa é de natureza exploratória, bibliográfica, e quali-quantitativa, pautada nos procedimentos de estudo métrico e análise de conteúdo. Como resultados, discute-se a distribuição temporal dos trabalhos, as entidades de afiliação dos autores, além de caracterizar os procedimentos metodológicos e resultados dos trabalhos relativos à mineração de textos científicos. Conclui-se que ao longo de 18 anos, 28 trabalhos foram publicados sobre a extração de conhecimento por meio da mineração de textos. Dentre estes, 13 tratam da mineração de textos científicos, dos quais nove tem natureza aplicada, sendo esses analisados mais profundamente.

Palavras-chave*:* Estudos Métricos da Informação. Descoberta de conhecimento em bases de dados. Mineração de texto. Mineração de textos científicos. Ciência da Informação.

1 Introdução

O volume de informação científica cresce de maneira exponencial, fato este que provoca uma série de novas demandas e necessidades. As publicações científicas são disponibilizadas como arquivos digitais, produzidos em larga escala e geralmente nos formatos PDF ou HTML. Entretanto, apesar do acesso

facilitado a estes documentos digitais, a apropriação do seu conteúdo, pelos pesquisadores, continua exigindo muito tempo no processo de leitura textual.

Os grandes volumes de publicações e dados científicos produzidos demandam o desenvolvimento de técnicas capazes de extrair conhecimento de forma rápida e precisa. A exemplo disto, são os cerca de 2,5 milhões de artigos científicos produzidos pelos Estados Unidos entre 2011 e 2016. (CROSS; THOMSON; SINCLAIR, 2016). Neste contexto, Capuano (2009) afirma que a existência de métodos e processos de descoberta de informação são adequadas e tornam-se relevantes quando buscam aplicar essas técnicas em textos em linguagem natural.

Marcondes, Costa e Martins (2016), asseguram que o ato de processar o conhecimento inserido no texto de um artigo, com o objetivo de identificar lacunas, contradições ou concordâncias no conhecimento de uma área, servindo também como instrumento de validação das metodologias desenvolvidas, é bastante dispendioso e trabalhoso, demandando ampla leitura e análise da linguagem por humanos. Os autores apontam que a tarefa base para o processamento automático de textos é a Extração de Informação (EI), que consiste em encontrar informação específica no texto dos documentos, sendo necessário diferenciá-la de outros métodos ou processos:

> Diferente da Recuperação da Informação (RI) que tem por objetivo encontrar textos e documentos relevantes, de acordo com a consulta do usuário, a EI trata de solucionar o problema de achar informações dentro dos textos. Difere, também, da PLN (Processamento de Linguagem Natural) porque é mais específico, visando a extrair determinados tipos de informação (obter informação pré-especificada), geralmente direcionada para extrair características do domínio (termos, objetos, entidades, relações) no qual o texto está inserido. A EI é diferente, ainda, da Extração de Conhecimento (Descoberta de Conhecimento – *Knowledge Discovery in Databases* – KDD), porque não visa deduzir regras. (MARCONDES; COSTA; MARTINS, 2016, p. 186.)

Neste artigo, será explorada a descoberta de conhecimento por meio da mineração de texto, que se mostra como alternativa viável para o tratamento da crescente produção textual disponibilizada em bases de dados científicas. Neste

sentido, o avanço das pesquisas em mineração de texto, torna-se importante para subsidiar a aquisição de conhecimento por especialistas nesses ambientes.

Uma definição pertinente quanto ao conceito de mineração de texto pode ser estabelecida como:

> A mineração de dados significa procurar padrões nos dados. Da mesma forma, a mineração de texto trata da procura de padrões no texto: é o processo de analisar o texto para extrair informações úteis para fins específicos. (WITTEN; FRANK; HALL, 2011, p. 386, tradução nossa)

Para os autores Frawley, Piatetsky-Shapiro e Matheus (1992, p. 58) a descoberta de conhecimento seria "a extração não trivial de implícitos, anteriormente desconhecidos, e descoberta de informações potencialmente úteis dos dados".

Faro, Giordano e Spampinato (2012, p. 62, tradução nossa), declaram que "o principal objetivo da mineração de texto é descobrir o conhecimento oculto no texto em artigos publicados e apresentá-los aos usuários de forma coerente e concisa". Assim como a mineração de dados, a mineração de texto é tarefa específica da descoberta de conhecimento.

Nagarkar e Kumbhar (2015) afirmam que a mineração de texto tem suas técnicas utilizadas em muitas áreas do conhecimento, inclusive na Biblioteconomia e na Ciência da Informação, e que o seu principal objetivo é o de extrair informação em um conjunto de dados não estruturados.

Seguindo essa perspectiva, O'Mara-Eves *et al* (2015) descrevem como as técnicas de mineração de textos podem auxiliar na identificação de citações:

> Quando há um grande número de estudos para triagem manual, identificar rapidamente a maioria dos relevantes permite que alguns membros de uma equipe de revisão comecem a próxima etapa da revisão, enquanto o restante das citações mais irrelevantes, são examinadas por outros membros da equipe. Isto reduz o tempo do início da revisão à conclusão, mesmo que a carga de trabalho total permaneça a mesma. (O'MARA-EVES *et al* 2015, p. 2, tradução nossa)

Desse modo, apesar de continuar existindo um grande esforço na revisão manual de citações, o trabalho evoluí significativamente quando o grupo responsável pela revisão empreende seus esforços na análise das citações mais relevantes identificadas por meio da aplicação das técnicas de mineração de textos.

Apesar da busca pela automatização do processo de mineração de texto, o componente humano ainda é muito importante no que tange à identificação de informação relevante, como afirmou Srinivasan (2004). Em seu estudo, o referido autor ressalta que embora se busque automatizar o processo ao máximo, o envolvimento do usuário em várias decisões é crucial para uma descoberta do conhecimento bem-sucedida.

Neste contexto e buscando contribuir para a consolidação da apropriação do tema pela Ciência da Informação brasileira, o presente artigo busca responder aos seguintes problemas de pesquisa: Quais são os principais indicadores das pesquisas sobre mineração de texto registrados em periódicos científicos brasileiros da área de Ciência da Informação? Quais são as principais tarefas aplicadas na mineração de textos científicos?

Na busca por obter essas respostas, o objetivo deste trabalho é analisar artigos de periódicos científicos brasileiros da área de Ciência da Informação sobre mineração de textos e descoberta de conhecimento em textos, analisando mais profundamente os que apliquem técnicas de mineração de texto a documentos científicos. A justificativa para realização do presente trabalho parte da busca por compreender os níveis de produção e institucionalização das pesquisas, além de elaborar um panorama da construção do conhecimento sobre o tema.

2 Tarefas de Mineração de Texto

Nesta seção são descritas as principais tarefas de mineração de textos. Tal caracterização serve para qualificar os resultados obtidos na seção 4.

De acordo com Rezende (2005), as tarefas de mineração de texto representam diferentes tipos de extração de conteúdo textual. Cada uma delas tem um propósito específico e um método apropriado no processo de coleta de características textuais.

No Quadro 1 é possível visualizar as diferentes finalidades das tarefas de mineração de textos. Os artigos aplicados de mineração de textos científicos serão analisados quanto à tarefa de mineração de texto realizada, com o propósito de apresentar quais foram as principais tarefas utilizadas em suas metodologias.

Quadro 1 - Tarefas de Mineração de Texto

TAREFA	FINALIDADE DA TAREFA
AGRUPAMENTO	Torna explícito o relacionamento entre documentos, agrupando documentos similares.
CATEGORIZAÇÃO	Identifica os tópicos-chave de um documento associando-o a categorias pré-definidas como áreas, domínios do conhecimento ou temas.
EXTRAÇÃO DE CARACTERÍSTICAS	Extrai padrões de termos com características em comum, também denominada extração de termos.
SUMARIZAÇÃO	Realiza o processo de redução textual no nível de sentenças, mantendo os significados-chave do texto
INDEXAÇÃO	Identifica os tópicos-chave de um documento por meio de padrões pré-existentes que permitem filtrar os termos que são palavras-chave de um documento.
REGRAS DE ASSOCIAÇÃO	Encontra relacionamentos ou padrões frequentes entre documentos, autores, citações, termos, categorias ou outros aspectos dos documentos.
REGRESSÃO	Gera um modelo preditivo da distribuição dos termos ou palavras-chave em um conjunto de documentos.

Fonte: Dados da Pesquisa (2019).

Na próxima seção serão abordados os procedimentos metodológicos necessários para alcançar os objetivos da pesquisa.

3 Procedimentos Metodológicos

A presente pesquisa possui natureza quali-quantitativa e quanto aos objetivos classifica-se como exploratória, pois busca proporcionar um panorama ou mapeamento dos artigos de periódicos científicos brasileiros da Ciência da Informação sobre mineração de textos. Realiza uma pesquisa bibliográfica, pautada nos procedimentos de estudo métrico e análise de conteúdo.

O estudo métrico tem como base teórico-metodológica a bibliometria. Segundo Santos (2003) a bibliometria pode ser considerada como o estudo quantitativo da ciência e da tecnologia. O fato de estabelecer medições ou mensurações é o que permite a quantificação dos elementos presentes nos metadados dos documentos analisados.

Para atingir o objetivo da pesquisa, foram realizadas as seguintes etapas:

1) Efetivou-se uma busca por “Mineração de Texto” nas bases Brapci e Scielo, no dia 12/09/2019, onde foram obtidos 23 trabalhos da área de Ciência da Informação que continham o termo em alguma parte dos metadados. Numa segunda busca, com o objetivo de associar mineração de texto e descoberta de conhecimento em texto, foi realizada uma pesquisa na Brapci e Scielo com o termo composto “Descoberta de Conhecimento”. Essa busca recuperou 14 registros da área de Ciência da Informação que continham o termo em algum dos campos dos metadados. Mesclando os dois conjuntos de documentos, removendo os artigos duplicados e realizando uma filtragem, chegou-se a 28 documentos. Para critério de filtragem, foi adotado o procedimento de leitura dos resumos dos trabalhos e foram escolhidos aqueles que possuíam relação com a temática “mineração de texto” ou “descoberta de conhecimento em texto”.

2) Foram baixados os 28 documentos, para posteriormente extrair os seus metadados e caracterizá-los com relação aos objetivos. Após uma leitura minuciosa foi identificado que dos 28 trabalhos obtidos, 13 tinham relação direta com a mineração de textos científicos. Esses 13 foram selecionados para uma análise mais profunda, sendo tais artigos identificados na Tabela 1. Dos 13 artigos selecionados, 1 foi publicado no Enancib e 12 em artigos de periódicos. Dessa maneira, buscou-se a base de dados BENANCIB com o propósito de investigar se haveria mais trabalhos publicados no Enancib relacionados ao tema. Ao realizar uma busca avançada no repositório, com a seguinte expressão de busca ((resumo: "mineração de texto") OU (título: "mineração de texto") OU (palavras-chave: "mineração de texto")) obteve-se 2 resultados que não se enquadravam na mineração de textos científicos. De forma semelhante, ao substituir o termo composto "mineração de texto" por "descoberta de conhecimento" na expressão de busca, foram obtidos resultados que não se enquadravam com a temática do estudo. Portanto, apenas 1 trabalho publicado no Enancib foi encontrado sobre mineração de textos científicos.

3) Por meio da análise dos dados foram construídas as seguintes visualizações:

    a) Na subseção 4.1 são descritos os resultados da análise de conteúdo de nove trabalhos aplicados dos 13 artigos sobre mineração de textos científicos. Esses nove artigos foram lidos e analisados os principais resultados obtidos por meio dos procedimentos metodológicos adotados.

    b) No Gráfico 1, identificou-se quais as tipologias dos documentos utilizados pelos pesquisadores em suas pesquisas para compor o corpus de análise onde foram aplicadas as técnicas de mineração de texto;

    c) No Gráfico 2, foi realizada uma análise quantitativa dos artigos produzidos ao longo do tempo;

d) No Gráfico 3, procurou-se vincular os autores às respectivas instituições a fim de mapear geograficamente a origem da produção na área;

e) No processo de construção da Tabela 1 foi necessário ler os 13 artigos e identificar em suas metodologias a utilização de software/sistema de mineração de texto. Nesse ponto os artigos foram classificados quanto à categoria e software utilizado;

f) No Gráfico 4, foram obtidas as bases de dados mais utilizadas pelos trabalhos aplicados de mineração de textos científicos;

g) Na tabela 2, buscou-se relacionar cada tarefa de mineração de textos aos respectivos artigos aplicados.

4) Posteriormente, procedeu-se a análise dos resultados com o propósito de gerar as discussões pertinentes ao tema por meio de inferências e deduções.

Os gráficos e tabelas foram gerados com o intuito de facilitar a visualização dos resultados, que são apresentados na seção a seguir.

4 Resultados e discussão

Nesta seção são descritos os resultados alcançados na análise dos artigos de periódicos científicos brasileiros da área de Ciência da Informação sobre mineração de textos, com ênfase na mineração de textos científicos.

4.1 Descrição dos trabalhos aplicados de mineração de textos científicos

Foram identificados nove artigos de natureza aplicada, entre os 13 artigos que tratam de mineração em textos científicos. Seus objetivos e métodos são descritos a seguir, em ordem cronológica.

A primeira pesquisa é a de Maia e Souza (2010), intitulada “Uso de Sintagmas nominais na classificação automática de documentos eletrônicos”. O estudo buscou verificar se ocorreria o aprimoramento na classificação de

documentos eletrônicos com o uso de técnicas de mineração de texto. Nesta pesquisa os autores utilizaram dois softwares: OGMA e WEKA. O OGMA foi desenvolvido pelos autores para a extração dos sintagmas nominais, já o WEKA foi utilizado pelos autores para analisar os resultados obtidos por meio da aplicação de algoritmos de agrupamento e de classificação aos documentos, os quais foram representados pelo conjunto de sintagmas nominais extraídos.

A segunda pesquisa a ser mencionada é a de Trucolo e Digiampietri (2014), "Análise de tendências da produção científica nacional na área de Ciência da Informação: estudo exploratório de mineração de textos". Os autores buscaram identificar tendências das produções científicas de artigos de periódicos dos doutores da área de Ciência da Informação no Brasil. Eles utilizaram como fonte a plataforma Lattes e 34.289 títulos de artigos publicados entre 1991 e 2012. O método consistiu em determinar os termos mais importantes a partir da extração automática de termos compostos inseridos nos títulos das publicações. Os principais resultados apontaram os temas mais trabalhados entre 1991 e 2012, e por meio de um cálculo de regressão foram projetadas tendências de frequência de ocorrência para os principais temas a serem trabalhados em 2013, 2015 e 2020.

A terceira pesquisa a ser mencionada é a de Bezerra e Guimarães (2014), o estudo intitulado "Mineração de texto aplicada às publicações sobre gestão do conhecimento no período de 2003 a 2012" buscou apresentar os termos mais empregados sobre trabalhos na área de Gestão do Conhecimento ao longo de 10 anos. O estudo empregou 3.457 resumos de língua inglesa e 380 resumos em língua portuguesa como corpus de análise. Os principais resultados apontaram associações temáticas entre os termos presentes nos resumos dos artigos e identificaram os agrupamentos de termos mais frequentes e mais ocorrentes em relação ao total de documentos. Além disso, apresentaram termos mais associados aos aspectos pragmáticos e termos associados aos elementos mais abstratos da Gestão do Conhecimento.

A quarta pesquisa a ser ressaltada é a de Carvalho, Escobar e Tsunoda (2014), com o título "Pontos de Atenção para o Uso da Mineração de Dados na

Saúde". Os autores buscaram apresentar uma pesquisa exploratória em bases de dados, das quais foram adotados 18 artigos no corpus. Eles utilizaram o algoritmo Apriori do ambiente Weka e o algoritmo DRE (Descobre Regras de Exceção) de onde obtiveram 345 regras gerais de associação.

O quinto estudo teve como autores Araújo e outros (2016), intitulado "Descoberta de Conhecimentos sobre Esquistossomose a partir de documentos científicos utilizando técnicas de Mineração de Textos". A pesquisa propôs a aplicação de técnicas de mineração de texto para a descoberta de conhecimento sobre esquistossomose a partir de um acervo científico do Instituto Oswaldo Cruz. Nesse estudo, os pesquisadores coletaram 179 resumos de artigos publicados entre os anos de 2005 e 2015 selecionados a partir do descritor "schisto". Dessa maneira, foram geradas listas de termos mais relevantes, documentos classificados por categorização e agrupamento sem suporte de especialista.

A sexta pesquisa foi desenvolvida pelos autores Sérgio, Silva e Gonçalves (2016), com o título "Descoberta de Conhecimento a partir de informações não estruturadas por meio de técnicas de correlação e associação". Esse trabalho propôs a apresentação de um modelo para descoberta de Conhecimento em textos com base nas técnicas de correlação e associação temporal entre termos de indexação. Para isso, o corpus foi composto por artigos coletados na Science Direct e utilizados para revelar as relações entre "biotecnologia" e "engenharia genética", e "nanotecnologia" e "medicina". Foram coletados 551 artigos no ano de 2013 compreendendo 2 períodos respectivos para a busca, 1993 a 2002 e 1984 a 1993. Os resultados apontaram a evolução temporal dos relacionamentos entre os termos, possibilitando a identificação de padrões e tendências via descoberta de conhecimento. A pesquisa propôs o desenvolvimento de um modelo que permitiu analisar a evolução da relação temporal de associação e correlação entre pares de termos, bem como construiu mapas de tópicos com os termos mais relacionados a um termo alvo.

O sétimo trabalho desenvolvido, intitulado "Análise de dados em artigos recuperados da *Web of Science* (WoS)" de autoria de Carvalho e Tsunoda

(2018). Nesse trabalho os autores buscaram a identificação de padrões em publicações sobre mineração de textos, mais especificamente em 1.193 registros de artigos na plataforma da WoS. Para isso foi desenvolvida uma aplicação para inserção dos dados no formato BibTex num banco de dados MySQL, seguida da utilização das ferramentas R e Apriori para extração e análise dos dados. Os principais resultados apontaram periódicos, autores, países, palavras-chave e termos de indexação utilizados nos registros dos artigos, bem como 13 regras de associação envolvendo palavras isoladas co-ocorrentes nos termos de indexação.

O oitavo estudo foi desenvolvido por Ferreira e Corrêa (2018) intitulado "Estudo Métrico sobre biblioteca digital: uso do software iramuteq". A pesquisa envolveu o levantamento e visualização dos termos em formas de grafos de redes, tabela de termos mais frequentes e nuvem de palavras. A metodologia adotada utilizou análise de frequência e co-ocorrência de termos nas palavras-chave, resumos e títulos. Foram analisados 82 trabalhos por meio do software Iramuteq, e foi apontado que o termo composto Biblioteca digital têm proximidade com outras temáticas como informação, tecnologia, artigo, digital, biblioteca, acesso, serviço, teses e dissertações (FERREIRA, CORRÊA, 2018).

Posteriormente, destaca-se a nona pesquisa, de Braga (2018). A autora propôs a criação de um modelo associado a uma representação de documentos por conceitos e aplicação de um método de agrupamento hierárquico de documentos. Tal processo metodológico baseou-se na frequência da co-ocorrência dos conceitos com o objetivo de produzir uma taxonomia de conceitos. A pesquisadora utilizou o algoritmo Apriori de Agrawal e Srikant (1994). O corpus analisado utilizou 1.841 trabalhos científicos da biblioteca digital da Comissão Nacional de Energia Nuclear (CNEN), utilizando o texto completo. O principal resultado obtido foi uma árvore de taxonomia com a representação dos principais conceitos presentes nos trabalhos analisados.

A próxima subseção sistematiza o resultado da análise dos trabalhos sobre mineração de textos e mineração de textos científicos.

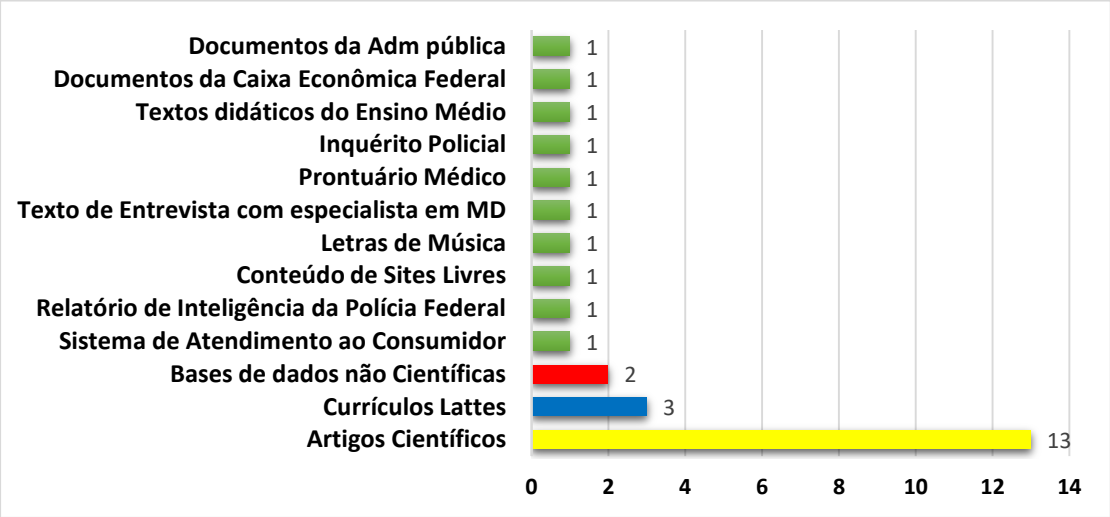
4.2 Análise dos trabalhos sobre mineração de textos

Analisando a tipologia documental dos 28 trabalhos foi construído o Gráfico 1, que quantifica as tipologias de documentos no corpus de análise utilizado pelas publicações sobre mineração de texto.

Ressalta-se que as publicações sobre mineração de textos científicos, que são objetos de análise nesta pesquisa, estão representadas na barra em amarelo. Esses documentos discutem o uso de técnicas de mineração de texto para extrair conhecimento de publicações científicas, embora nem todos tenham caráter aplicado, como será discutido mais adiante.

No Gráfico 1 observa-se que, dos 28 trabalhos sobre mineração de texto 10 analisaram uma única vez determinada tipologia documental. Aponta-se que a repetição na análise de determinada tipologia documental não é ruim, e a repetição é necessária para amadurecer os aspectos metodológicos da aplicação da mineração de texto a determinada tipologia documental. A análise desse aspecto foi importante por permitir a seleção dos trabalhos que analisaram artigos científicos. Tais artigos possuem em comum o objetivo de discutir a aplicação de mineração de texto em artigos científicos.

Gráfico 1 – Caracterização do Corpus

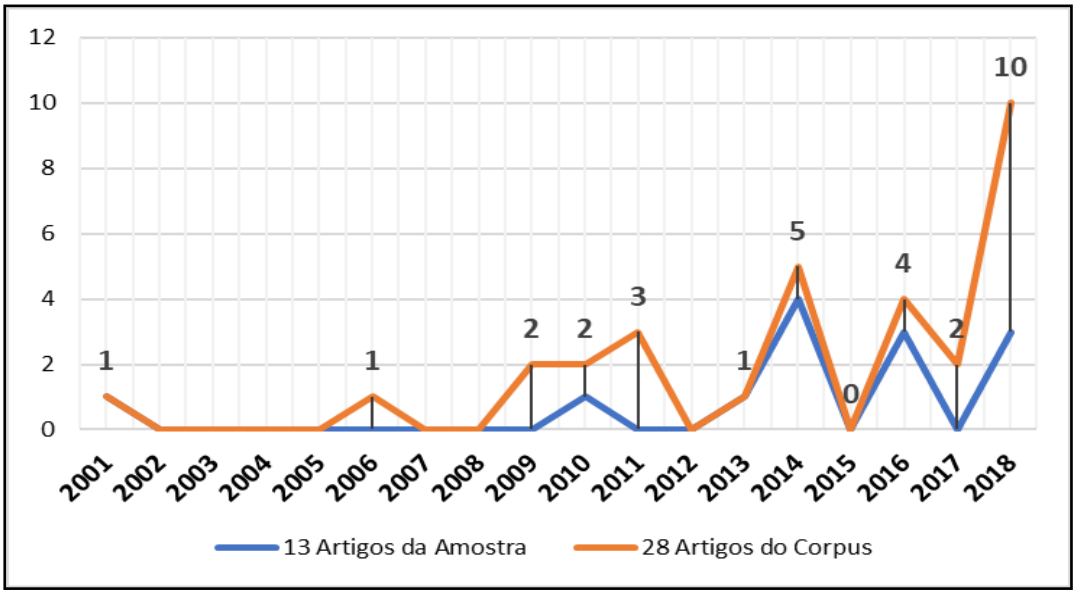


Fonte: Dados da Pesquisa (2019)

Dando prosseguimento as análises, com o objetivo de apresentar os anos de publicação dos artigos, foi elaborado o Gráfico 2. Nele é possível identificar os anos de publicação dos artigos sobre mineração de textos científicos, assim como compará-los com o total de trabalhos sobre mineração de texto.

É importante destacar que a maioria dos trabalhos sobre mineração de texto foram publicados a partir do ano de 2009. Dos 28 trabalhos, apenas um foi publicado em 2001 e outro em 2006. As publicações sobre o tema tornaram-se mais frequentes a partir do ano de 2014, apresentando assim uma tendência de crescimento do número de publicações nos últimos anos. Percebe-se também que em 2014, 2016 e 2018 ocorrem picos na quantidade de artigos publicados sobre o tema (5, 4 e 10 respectivamente).

Gráfico 2 – Quantidade de Artigos publicados por ano



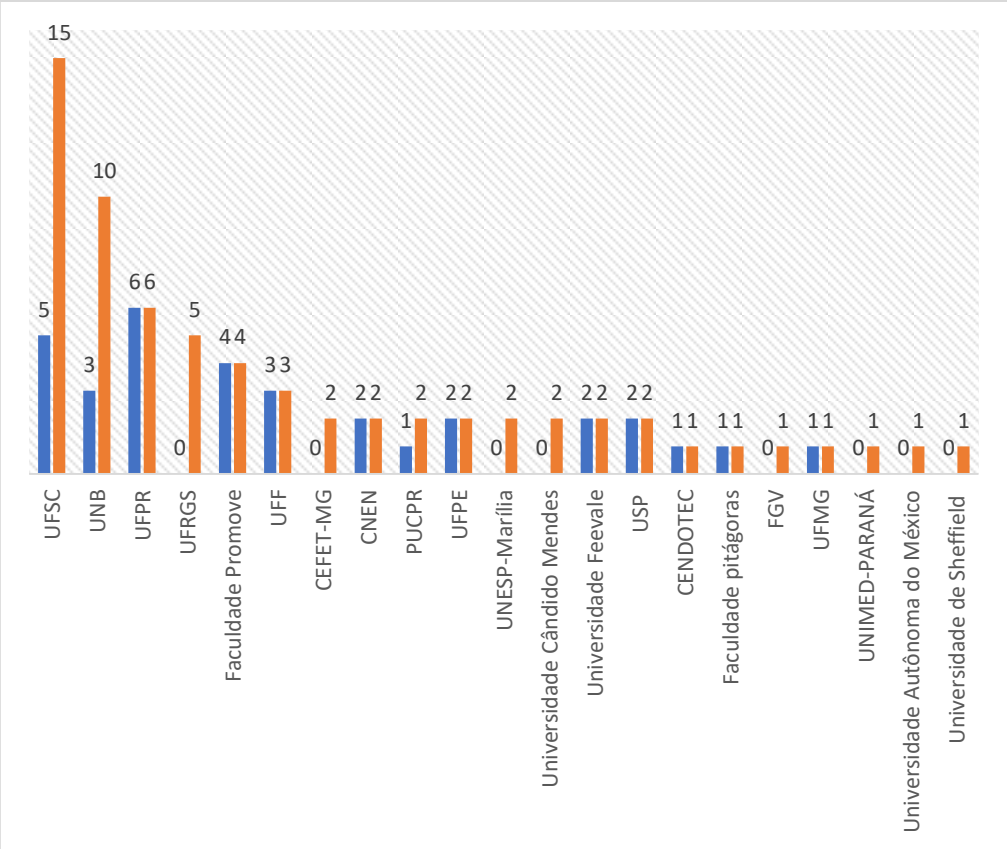
Fonte: Dados da pesquisa (2019)

Outro fator importante diz respeito ao período em que não se produziu publicações sobre o tema a partir de 2001, mais especificamente entre os anos 2002 e 2005, e nos anos de 2007, 2008, 2012 e 2015. As publicações que tratam da mineração de textos científicos têm início em 2001, sendo retomada em 2010 e 2013, tendo quatro publicações em 2014, três em 2016 e três em 2018. Percebe-se, assim, uma quase estabilização do número de trabalhos sobre o

tema, e um repentino crescimento exponencial em 2018, indicando que a mineração de textos científicos está sendo apropriada pela Ciência da Informação no Brasil.

Já no Gráfico 3, observam-se as instituições nas quais os autores dos trabalhos eram vinculados na época da publicação dos artigos. Destacam-se 2 cores para a legenda: 1) em azul visualizam-se as instituições vinculadas aos autores dos 13 artigos sobre mineração de textos científicos; 2) já em laranja, as instituições vinculadas aos autores dos 28 trabalhos sobre mineração de textos. O objetivo de compará-los foi o de facilitar a identificação do núcleo dominante de instituições que realizam estudos na área. Percebe-se que a maioria das instituições pertence ao eixo sul-sudeste do país, com destaque para Universidades renomadas como a UFSC, UFPR, UFRGS e UFF. No centro-oeste, tem-se em destaque a UNB. Dentre as instituições privadas, destacam-se a Faculdade Promove, a PUC do Paraná e as Universidades Cândido Mendes e Feevale como instituições de vínculo dos autores. Dentre as instituições públicas que se destacam, ressalta-se a UFSC com 15 autores vinculados que produziram trabalhos sobre mineração de textos, sendo que desse total 6 autores produziram trabalhos sobre mineração de textos científicos. A segunda colocada em evidência é a UNB com 10 autores vinculados que produziram trabalhos sobre mineração de textos, considerando que 5 destes autores produziram trabalhos sobre mineração de textos científicos. Regiões como a Norte e o Nordeste possuem poucos pesquisadores publicando sobre temas que envolvem mineração de textos científicos (apenas 2 publicaram na UFPE). De forma geral, pode-se concluir que existe uma necessidade de maior apropriação do tema por cursos de graduação e pós-graduação em Ciência da Informação no Brasil.

Gráfico 3 – Quantidades de Autores Vinculados às Instituições



Fonte: Dados da pesquisa (2019)

Realizando um estudo mais aprofundado das origens das pesquisas desenvolvidas, visualiza-se que a maioria dos estudos tem origem em parcerias entre mestrandos, doutorandos e docentes em diferentes programas de pós-graduação. Esse cenário demonstra a capacidade produtiva das pós-graduações em estimular seus alunos e pesquisadores a publicar trabalhos em coautoria. Deste modo, é importante reconhecer o esforço conjunto em prol da pesquisa sobre mineração de textos científicos realizado por essas instituições e os seus respectivos programas de pós-graduação.

Já na Tabela 1 é interessante destacar a caracterização do método dos artigos sobre mineração de texto científicos quanto à natureza teórica ou prática. Verifica-se que 4 dos 13 artigos não utilizam softwares/sistemas em suas metodologias para realizar a mineração de textos científicos. As propostas desenvolvidas nesses 4 trabalhos buscaram realizar um estudo teórico

contemplando os elementos constituintes dos estudos sobre mineração de texto em publicações científicas de domínios especializados.

Ainda na Tabela 1 é possível identificar os softwares/ferramentas utilizados nas pesquisas pelos respectivos autores dos trabalhos. Inicialmente constata-se que o software livre Weka é o mais utilizado na mineração de texto de publicações científicas. Uma das principais justificativas pela sua utilização deve-se ao fato de ser um software livre e por possuir uma coleção de algoritmos que utilizam a aprendizagem de máquina com o objetivo de realizar a mineração de dados e textos.

Outro software livre citado na Tabela 1 é o R. O R também é um software livre e multiplataforma, podendo ser utilizado em Unix, Windows ou MacOS. Sua principal vantagem é possuir uma vasta capacidade de realizar análises estatísticas com uso de testes e análises já consolidadas. Outro software livre aplicado é o Iramuteq, também baseado na linguagem R, porém específico para análises multidimensionais em textos. Além dos softwares identificados nesta tabela, também foram utilizados *EndNot*e, *Infotrans* e *Dataview* como softwares para realização de estudos métricos, neste caso, foram utilizados para apresentar dados de revisão bibliográfica.

Outro ponto importante identificado na Tabela 1 foi a ausência de identificação do software no artigo de Trucolo e Digiampietri (2014), os autores mencionaram as fórmulas utilizadas para o cálculo de tendências de termos na área da CI, mas não descreveram a ferramenta utiliza para extrair os termos dos currículos Lattes.

Tabela 1 - Caracterização dos artigos sobre mineração de textos científicos

CITAÇÃO	TÍTULO	CATEGORIA	SOFTWARE
(QUONIAM; TARAPANOFF; ARAÚJO JÚNIOR; ALVARES, 2001)	Inteligência obtida pela aplicação de data mining em base de teses francesas sobre o Brasil	**TEÓRICO**	**NÃO UTILIZA**
(MAIA; SOUZA, 2010)	Uso de Sintagmas nominais na classificação	**APLICADO**	**Ogma; Weka**

	automática de documentos eletrônicos		
(WOSZEZENKI; GONÇALVES, 2013)	Mineração de textos biomédicos: uma revisão bibliométrica	**TEÓRICO**	**NÃO UTILIZA**
(BEZERRA; GUIMARÃES, 2014)	Mineração de texto aplicada às publicações sobre gestão do conhecimento no período de 2003 a 2012	**APLICADO**	**EndNote; Aplicativo PHP; PreText2**
(CARVALHO; ESCOBAR; TSUNODA, 2014)	Pontos de Atenção para o Uso da Mineração de Dados na Saúde	**APLICADO**	**Apriori; Weka**
(PINHEIRO; BARTH, 2014)	Produção Científica na Base de Dados SCOPUS: uma análise sobre indústria criativa	**TEÓRICO**	**NÃO UTILIZA**
(TRUCOLO; DIGIAMPIETRI, 2014)	Análise de tendências da produção científica nacional na área de Ciência da Informação: estudo exploratório de mineração de textos.	**APLICADO**	**NÃO INDENTIFICADO**
(ARAÚJO; DAVID; RIOS; VELOSO, 2016)	Descoberta de Conhecimentos sobre a esquistossomose a partir de Documentos Científicos Utilizando Técnicas de Mineração de Textos	**APLICADO**	**Weka**
(MARCONDES; COSTA; MARTINS, 2016)	Descoberta de Conhecimento em Artigos Digitais em Ciências Biomédicas	**TEÓRICO**	**NÃO UTILIZA**
(SÉRGIO; SILVIA; GONÇALVES, 2016)	Descoberta de Conhecimento a partir de informações não estruturadas por meio de técnicas de correlação e associação	**APLICADO**	**GridGain**
(BRAGA, 2018)	Extração semiautomática de taxonomia para domínios especializados usando técnicas de mineração	**APLICADO**	**Software R**

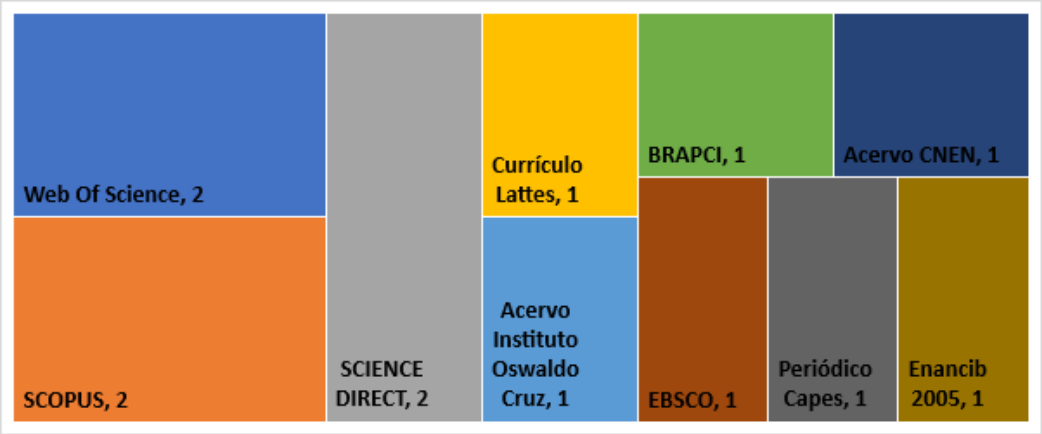
	de textos		
(CARVALHO; TSUNODA, 2018)	Análise dos dados Recuperados da Web of Science (WoS)	**APLICADO**	**MinGNU; Weka; MySQL; Apriori**
(FERREIRA; CORRÊA, 2018)	Estudo Métrico Sobre Biblioteca Digital: Uso do software Iramuteq	**APLICADO**	**Iramuteq**

Fonte: Dados da pesquisa (2019)

Já o Gráfico 4 apresenta um mapa de árvore onde é possível observar os valores absolutos da quantidade de vezes em que as fontes de informações ou bases de dados foram utilizadas na aquisição de documentos para aplicação das técnicas de mineração de texto. Ele apresenta as bases de dados utilizadas nos artigos aplicados sobre mineração de texto científico. A leitura do grafo ocorre pelo tamanho da área ocupada pelo polígono quadrilátero correspondente à frequência de ocorrência de cada valor.

Dessa forma, ele indica as bases Scopus, *Web of Science* e a *Science Direct* como as mais utilizadas (2 vezes), enquanto as demais foram utilizadas uma vez cada para obter os trabalhos analisados. Esse gráfico apresenta a predominância da busca de fontes de informação já consolidadas e com ampla cobertura de periódicos que possam aumentar, significativamente, a quantidade de artigos recuperados sobre determinado tema. A Scopus é uma das maiores bases de dados de resumos e citações da literatura revisada por especialistas, já a *Web of Science* indexa mais de 33 mil periódicos no mundo todo. Nessa mesma linha, a *Science Direct* possui mais de 250 mil artigos no formato de arquivos abertos com amplitude mundial. É justificável a escolha de tais bases por sua amplitude de cobertura, reconhecimento mundial, pela confiabilidade e segurança das informações pesquisadas.

Gráfico 4 – Fontes de Informação dos documentos analisados pelos artigos aplicados



Fonte: Dados da Pesquisa (2019)

Por fim, na Tabela 2, buscou-se descrever as tarefas de mineração de textos realizadas pelos trabalhos aplicados de mineração de textos científicos. As tarefas de mineração de texto estão descritas no Quadro 1 da seção 2.

Tabela 2 - Artigos aplicados e respectivas tarefas de Mineração de Textos

ARTIGO	TAREFA DE MINERAÇÃO	DETALHAMENTO DE TAREFA
(MAIA; SOUZA, 2010)	**CATEGORIZAÇÃO**	Extraiu sintagmas nominais para representar o conteúdo dos documentos e depois realizou a categorização automática dos documentos.
(BEZERRA; GUIMARÃES, 2014)	**EXTRAÇÃO DE CARACTERÍSTICAS**	Realizou a identificação de termos por meio de n-gramas de radicais de palavras recorrentes nos resumos.
(CARVALHO; ESCOBAR; TSUNODA, 2014)	**REGRAS DE ASSOCIAÇÃO**	Obteve pontos de atenção da área de Saúde por meio da extração de regras de associação. A Associação entre os pontos de atenção identificados serviu como base para ampliar as estratégias que ampliam a mineração de dados na área da Saúde.
(TRUCOLO; DIGIAMPIETRI, 2014)	**EXTRAÇÃO DE CARACTERÍSTICAS**	Utilizou técnica de extração automática de termos para determinar os termos mais importantes num conjunto de

		documentos pelo cálculo da frequência adjacente das palavras que compõem esses termos.
(ARAÚJO; DAVID; RIOS; VELOSO, 2016)	**EXTRAÇÃO DE CARACTERÍSTICAS; CATEGORIZAÇÃO; E AGRUPAMENTO**	Objetivou descobrir conhecimento em textos científicos, por meio de termos extraídos e categorizados. Os documentos foram indexados por um vetor multidimensional e cada dimensão foi representada por um termo da coleção.
(SÉRGIO; SILVIA; GONÇALVES, 2016)	**REGRAS DE ASSOCIAÇÃO**	Propôs um modelo computacional baseado em técnicas de correlação e associação entre termos para gerar as regras de associação. A associação foi gerada por meio da utilização de um modelo vetorial com o objetivo de gerar um coeficiente a partir de uma lista de termos em diferentes datas.
(BRAGA, 2018)	**EXTRAÇÃO DE CARACTERÍSTICAS**	Utilizou metodologia para a extração semiautomática de taxonomia de conceitos.
(CARVALHO; TSUNODA, 2018)	**REGRAS DE ASSOCIAÇÃO**	Procurou explorar dados de artigos provenientes do Web of Science revelando características dos estudos sobre mineração de textos publicados entre 2010 e 2016. As 13 regras de associação observadas indicaram a possibilidade de um periódico ter utilizado um determinado termo de indexação.
(FERREIRA; CORRÊA, 2018)	**AGRUPAMENTO**	Aplicou a técnica de agrupamento de termos por meio da técnica de correlação temática e similitude.

Fonte: Dados da pesquisa (2019)

Na Tabela 2 observam-se as principais tarefas de mineração de textos realizadas nos artigos aplicados sobre mineração de textos científicos. Assim, é possível identificar as tarefas mais frequentes e as principais abordagens de aplicação referentes a cada tarefa. Ressalta-se, ainda, que a sumarização e a indexação automática não foram objetivo principal de nenhum dos trabalhos, mas que podem vir a ser em estudos vindouros.

5 Considerações finais

O presente trabalho analisou os artigos de periódicos científicos brasileiros na área de Ciência da Informação sobre mineração de textos. Os resultados apresentados revelam o conhecimento apropriado e registrado por pesquisadores na área sobre a temática.

Percebe-se que, ao longo de 18 anos, apenas 28 trabalhos foram identificados como pesquisas que sinalizam o propósito de investigar a descoberta de conhecimento por meio da mineração de textos. Entretanto, dentre esses, 13 trabalharam a temática da mineração de textos científicos.

Outra questão importante foi a identificação de que a maioria dos artigos foram publicados a partir de 2014, indicando ser um tema de interesse recente e crescente de pesquisa na Ciência da Informação brasileira. Além disso, a presença de 13 artigos sobre mineração de textos científicos, sendo nove de natureza aplicada, comprova o interesse no tema, assim como enfatiza a existência de uma massa crítica capaz de se apropriar do conceito e aplicá-lo.

Deduz-se que os estudos produzidos por pesquisadores ainda não alcançaram um grau de maturidade capaz de afirmar a existência de um núcleo de especialistas produzindo publicações em revistas da área sobre o tema. Os trabalhos aplicados parecem ser desdobramentos do atendimento a demandas específicas e pontuais de pesquisa.

Neste sentido, as discussões e pesquisas no campo da mineração de textos científicos podem ser ampliadas, analisando trabalhos e experiências estrangeiras, adaptando e reproduzindo experimentos bem-sucedidos em documentos de outros idiomas para documentos em língua portuguesa.

Estudos futuros almejam a análise da literatura internacional de Ciência da Informação, visando a ampliação do corpus de análise. Tal prática pode fornecer maior amplitude nas análises e permitir levantar um panorama do estado da arte no assunto para a área.

Referências

AGRAWAL, R.; SRIKANT, R. Fast Algorithms for Mining Association Rules in Large Databases. In: INTERNATIONAL CONFERENCE ON VERY LARGE DATA BASES, 20., 1994, Santiago de Chile. **Proceedings**... Santiago de Chile: Morgan Kaufmann, 1994. p. 487-499.

ARAÚJO, D. A. O.; DAVID, L. R. S.; RIOS, R. S. H.; VELOSO, R. R. Descoberta de Conhecimentos sobre a esquistossomose a partir de Documentos Científicos Utilizando Técnicas de Mineração de Textos. **Pesq. Bras. em Ci. da Inf. e Bi.**, João Pessoa, v.11, n.2, p. 173-186. 2016. Disponível em: www.periodicos.ufpb.br/ojs2/index.php/pbcib/article/view/31846. Acesso em: 10 abr. 2019.

BEZERRA, C. A.; GUIMARÃES, A. J. R. Mineração de texto aplicada às publicações científicas sobre gestão do conhecimento no período de 2003 a 2012. **Perspectivas em Ciência da Informação**, Belo Horizonte, v.19, n.2, p.131-146, abr./jun. 2014. Disponível em: http://www.scielo.br/pdf/pci/v19n2/10.pdf . Acesso em: 16 out. 2018.

BRAGA, F. R. Extração semiautomática de taxonomia para domínios especializados usando técnicas de mineração de textos. **Ciência da Informação**, Brasília, v.45, n.3, p.175-186, set./dez. 2016. Disponível em: http://revista.ibict.br/ciinf/article/view/4056. Acesso em: 16 out. 2018.

CAPUANO, E. A. O poder cognitivo das redes neurais artificiais modelo ART1 na recuperação da informação. **Ciência da Informação**, Brasília, v.38, n.1, p.9-30, jan./abr. 2009. Disponível em: http://www.scielo.br/pdf/ci/v38n1/01.pdf. Acesso em: 05 mar. 2019.

CARVALHO, D. R.; ESCOBAR, L. F. A.; TSUNODA, D. Pontos de Atenção para o Uso da Mineração de Dados na Saúde. **Informação e Informação**, Londrina, v.19, n.1, p.249-273, jan./abr. 2014. Disponível em: http://www.uel.br/revistas/uel/index.php/informacao/article/view/16532. Acesso em: 15 jul. 2019.

CARVALHO, M. B. de; TSUNODA, D. F. Análise de dados em artigos recuperados da Web of Science (WoS). **Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação**, Florianópolis, v.23, n. esp., p. 112-125, 2018. Disponível em: https://periodicos.ufsc.br/index.php/eb/article/view/1518-2924.2018v23nespp112. Acesso em: 07 jun. 2019.

CROSS, D.; THOMSOM, S.; SINCLAIR, A. **Research in Brazil: A report for Capes by Clarivate Analytics**. Brasília: Clarivate Analytics, 2016. 73 p.

FARO, A.; GIORDANO, D.; SPAMPINATO, C. Combining literature text mining with microarray data: advances for system biology modeling. **Briefings in Bioinformatics**, Oxford, v.13, n.1, p. 61-82, 2011. Disponível em: https://academic.oup.com/bib/article/13/1/61/219461. Acesso em: 16 out. 2018.

FERREIRA, M. H. W.; CORRÊA, R. F. Estudo Métrico Sobre Biblioteca Digital: uso do software Iramuteq. *In*: XIX ENCONTRO NACIONAL DE PESQUISA EM CIÊNCIA DA INFORMAÇÃO – ENANCIB 2018, 19., 2018, Londrina. **Anais...** Londrina: Ancib. 2018. 4436-4454. Disponível em: http://enancib.marilia.unesp.br/index.php/XIXENANCIB/xixenancib/paper/view/1238. Acesso em: 04 jun. 2019.

FRAWLEY, W. J., PIATETSKY-SHAPIRO, G., MATHEUS, C. J. Knowledge Discovery in Databases: An Overview. **AI Magazine**, S. L., v.13, n.3, p.57-70, 1992. Disponível em: https://pdfs.semanticscholar.org/7a7b/51b86e22d0077215287980c7ba793b09e4cd.pdf. Acesso em: 19 jun. 2019.

MAIA, L. C.; SOUZA, R. R. Uso de sintagmas nominais na classificação automática de documentos eletrônicos. **Perspectivas em Ciência da Informação**, Belo Horizonte, v.15, n.1, p.154-172, abr., 2010. Disponível em: http://www.scielo.br/scielo.php?script=sci_arttext&pid=S1413-99362010000100009&lng=en&nrm=iso. Acesso em: 14 maio 2020.

MARCONDES, C. A.; COSTA, L. C. da; MARTINS, S. de C.; Descoberta de Conhecimento em Artigos Digitais em Ciências Biomédicas. **Informação e Informação**, Londrina, v.21, n.2, p.170-216, maio/ago., 2016. Disponível em: http://www.uel.br/revistas/uel/index.php/informacao/article/view/27941. Acesso em: 12 jun. 2019.

NAGARKAR, S. P.; KUMBHAR, R.; Text mining: An analysis of research published under the subject category 'Information Science Library Science' in Web of Science Database during 1999-2013. **Library Review**, v.64, n.3, 2015. Disponível em: https://doi.org/10.1108/LR-08-2014-0091 . Acesso em: 25 maio 2020.

O'MARA-EVES, A. *et al*. Using text mining for study identification in systematic reviews: a systematic review of current approaches. **Systematic**

reviews, v. 4, n. 1, p. 1-22, 2015. Disponível em: https://doi.org/10.1186/2046-4053-4-5 . Acesso em: 25 maio 2020.

PINHEIRO, C. M. P.; BARTH, M. Produção Científica na Base de Dados SCOPUS: uma análise sobre a indústria criativa. **Pesq. Bras. em CI. Da Inf. e Bib.**, João Pessoa, v. 9, n. 2, p. 048-061, 2014. Disponível em: http://www.periodicos.ufpb.br/ojs/index.php/pbcib/article/view/19990. Acesso em: 22 jul. 2019.

QUONIAM, L.; TARAPANOFF, K.; ARAÚJO JÚNIOR, R. H.; ALVARES, L. Inteligência obtida pela aplicação de data mining em base de teses francesas sobre o Brasil. **Ciência da Informação**, Brasília, v. 30, n. 2, p. 20-28, maio/ago. 2001. Disponível em: http://www.scielo.br/pdf/ci/v30n2/6208.pdf. Acesso em: 08 jun. 2019.

REZENDE, S. O. (org.) **Sistemas Inteligentes:** fundamentos e aplicações. 1. ed. Editora Manole: 2005.

SANTOS, R. N. M. Indicadores estratégicos em ciência e tecnologia: refletindo a sua prática como dispositivo de inclusão/exclusão. **Transinformação**, v. 15, n. 3 esp., p. 129-140, 2003. Disponível em: http://www.brapci.inf.br/v/a/452. Acesso em: 15 jul. 2018.

SÉRGIO, M. C.; SILVA, T. N. da.; GONÇALVES, A. L. Descoberta de conhecimento a partir de informações não estruturadas por meio de técnicas de correlação e associação. **Em questão**, Porto Alegre, v. 22, n. 2, p. 87-113, mai/ago. 2016. Disponível em: https://seer.ufrgs.br/EmQuestao/article/view/59514/37953. Acesso em: 01 nov. 2018.

SRINIVASAN, P. Text Mining: Generating Hypotheses From MEDLINE. **Journal of The American Society for Information Science and Technology,** v. 55, n. 5, p. 396-413, 2004. Disponível em: https://onlinelibrary.wiley.com/doi/abs/10.1002/asi.10389 . Acesso em: 22 maio 2020.

TRUCOLO, C. C.; DIGIAMPIETRI, L. A. Análise de tendências da produção científica nacional na área de ciência da informação: estudo exploratório de mineração de textos. **AtoZ: Novas Práticas em Informação e Conhecimento**, Curitiba, v. 3, n. 2, p. 87-94, 2014. Disponível em: https://revistas.ufpr.br/atoz/article/view/41341/25335. Acesso em: 01 nov. 2018.

WITTEN, I. H.; FRANK, E.; HALL, M. A. **Data Mining**: practical machine learning tools and techniques. 3. Ed. Burlington: Elsevier, 2011. 629 p.

WOSZEZENKI, C. R.; GONÇALVES, A. L. Mineração de textos biomédicos: uma revisão bibliométrica. **Perspectivas em Ciência da Informação**, Belo Horizonte, v. 18, n. 3, p. 24-44, jul./set. 2013. Disponível em: http://portaldeperiodicos.eci.ufmg.br/index.php/pci/article/view/1733/1189. Acesso em: 23. abr. 2019.

Mining scientific texts: analysis of articles from Brazilian Scientific Journals in the Information Science area

Abstract: This work analyzes articles from Brazilian Scientific Journals of the Information Science area about text mining and knowledge discovery in texts. It analyses articles indexed by Brapci or Scielo that contain the compound terms text mining or knowledge discovery in the metadata, and specifically address the application of text mining techniques. The research methodology is exploratory, bibliographic, and qualitative and quantitative, based on the procedures of metric study and content analysis. As results, we discuss the temporal distribution of the works, the authors' affiliation entities, and characterize the methodological procedures and results of the works related to the mining of scientific texts. It concludes that over 18 years, 28 publications discusses about forms of knowledge discovery through text mining. Among these, 13 deal with the mining of scientific texts, of which nine have an applied nature, and these are analyzed more deeply.

Keywords: Metric Studies of Information. Knowledge Discovery in Databases. Text mining. Scientific Text Mining. Information Science.

Recebido: 20/01/2020
Aceito: 28/05/2020

Declaração de autoria
Concepção e elaboração do estudo: Márcio Henrique Wanderley Ferreira, Renato Fernandes Correa
Coleta de dados: Márcio Henrique Wanderley Ferreira, Renato Fernandes Correa
Análise e discussão de dados: Márcio Henrique Wanderley Ferreira, Renato Fernandes Correa
Redação e revisão do manuscrito: Márcio Henrique Wanderley Ferreira, Renato Fernandes Correa

Como citar

FERREIRA, Márcio Henrique Wanderley; CORREA, Renato Fernandes. Mineração de textos científicos: Análise de artigos de periódicos científicos brasileiros da área de Ciência da Informação. **Em Questão**, Porto Alegre, v.27, n. 1, p. 237-262, 2021. Doi: http://dx.doi.org/10.19132/1808-5245271.237-262