ARTIGO

Recebido em:
28/04/2017

Aceito em:
09/04/2018

Encontros Bibli: revista eletrônica de biblioteconomia e ciência da informação, v. 23, n. esp., p. 112-125, 2018.
ISSN 1518-2924. DOI: 10.5007/1518-2924.2018v23nespp112

Análise de dados em artigos recuperados da Web of Science (WoS)

Data analysis on articles retrieved from Web of Science (WOS)

Marcelo Batista de CARVALHO (carvalhomarcelob@gmail.com)*
Denise Fukumi TSUNODA (dtsunoda@ufpr.br)**
* Bacharel em Gestão da Informação pela Universidade Federal do Paraná – UFPR.
** Professor(a) da Universidade Federal do Paraná – UFPR.

Resumo
Dado o contexto da Mineração de Dados e da Mineração de Textos, objetiva-se analisar dados recuperados da *Web of Science* (WoS). Pretende-se identificar padrões nos estudos sobre Mineração de Textos voltados a escolha de ferramentas a serem utilizadas na aplicação de método de mineração de dados. Recuperaram-se referências de artigos no formato BibTeX na plataforma WoS. Desenvolveu-se uma aplicação para inserção de dados do formato BibTeX para um banco de dados MySQL. Com base nas características encontradas, elegeram-se a ferramenta R e algoritmo Apriori para utilização em parte dos dados. Extraíram-se dados de ferramentas, métodos, palavras-chave, termos, periódicos, países e autores presentes nos registros. A aplicação do Apriori resultou em treze regras de associação. A exploração dos dados de artigos provenientes da WoS revelou características dos estudos da área de Mineração de Textos. Trabalhos futuros podem adaptar a aplicação usada neste estudo e aplicar outros métodos de mineração no conjunto de dados.
Palavras-chave: Recuperação da informação. Descoberta de conhecimento em base de dados. Mineração de texto.

Abstract
In Data mining and Text mining context, the goal is to analyze data retrieved from Web of Science (WoS). This paper intends to identify patterns in Text mining researches on selection of tools to be used on datamining application. References in BibTeX format were retrieved from articles existing in WoS platform. An application imported data from BibTeX to a MySQL database. The found characteristics led to choose the R programming language and the Apriori algorithm on a subset of data. Data about tools, methods, keywords, indexing terms, journals, countries, and authors were identified in records. Apriori resulted on thirteen association rules. The exploration of data from WoS articles revealed characteristics of Data mining researches. Future works can adapt the application used on this study and use other datamining methods on the dataset.
Keywords: Information Retrieval. Knowledge Discovery in Databases. Text Mining.

v. 23, n. esp., 2018
p. 112-125
ISSN 1518-2924

1 INTRODUÇÃO

Para Fayyad, Piatetsky-Shapiro e Smyth (1996), a abordagem clássica de análise de dados, baseada no analista de dados como interface entre os dados e os usuários, é lenta, cara e altamente subjetiva. Os autores indicam a Descoberta de Conhecimento em Bases de Dados (*Knowledge Discovery in Databases* - KDD) como uma alternativa para ampliar as capacidades humanas de análise e lidar com a quantidade de informação coletada.

O KDD é um processo de procura de padrões a partir de dados, composto de várias etapas, dentre estas a Mineração de Dados (*Data Mining*). Esta é a principal fase do KDD, quando de fato é realizada a busca por conhecimentos úteis (GOLDSCHMIDT; PASSOS, 2005; AMARAL, 2001).

A Mineração de Textos pode ser entendida como uma variante da Mineração de Dados ou do KDD, por isso é também conhecida por Mineração de Dados em Textos ou Descoberta de Conhecimento em Textos (*Knowledge Discovery in Texts* - KDT). Enquanto a Mineração de Dados destina-se a lidar com dados estruturados (de bancos de dados), a Mineração de Textos suporta conjuntos de dados não-estruturados ou semiestruturados, como correios eletrônicos, arquivos de Linguagem de Marcação de Hipertextos (HTML) e documentos de texto em geral (VIJAYARANI; MUTHULAKSHMI, 2013).

A escolha de uma ferramenta para realizar qualquer pesquisa que envolva a Mineração de Textos não é tarefa trivial na maioria dos casos uma vez que estão disponíveis diversas ferramentas e análises (conforme apresentando no decorrer do artigo) sobre as mesmas sob os mais diversos pontos de vista: interface, linguagem, plataforma (software e hardware), compatibilidade com banco de dados etc. Assim, este trabalho objetiva explorar artigos provenientes da Web of Science (WoS), identificando-se as ferramentas, autores, países, palavras-chave, termos de indexação e métodos nos estudos sobre mineração de textos no período de 2010 a 2016 publicados na WoS.

Em busca realizada no dia quatorze de novembro de dois mil e dezessete pelos termos “Mineração de Dados” e “Web of Science” nas bases: Periódicos CAPES, na EBSCOhost e Banco de Dados de Teses e Dissertações, foram retornados, 4 (quatro), 0 (nenhum) e 14 (catorze) documentos, respectivamente. Após a análise de todos os 18 (dezoito) documentos retornados, observou-se que nenhum deles realizou pesquisa semelhante a relatada neste artigo.

A WoS é uma plataforma que oferece uma interface única de pesquisa a sete bases de dados, incluindo: Science Citation Index (SCI), Social Science Citation Index (SSCI) e Arts and Humanities Citation Index. Produzida pelo Institute for Scientific Information (ISI), a WoS é exaustivamente utilizada em estudo de citações (YONG-HAK, 2013; MUGNAINI; STREHL, 2008).

As próximas seções contemplam a revisão de literatura de suporte, principais encaminhamentos metodológicos, resultados (e análises) e as considerações finais com sugestões de continuidade do estudo.

2 REVISÃO DA LITERATURA

Abordam-se a seguir os temas e conceitos tomados como base para a pesquisa, a saber: banco de dados, Descoberta de Conhecimento em Bases de Dados, Mineração de Dados, Mineração de Textos, Descoberta de Associação e Apriori.

2.1 Banco de dados

Um banco de dados é uma "[...] coleção de dados que geralmente descrevem as atividades de uma ou mais organizações relacionadas" (RAMAKRISHNAN; GEHRKE, 2000, p. 3, tradução do autor)[1]. Num sentido mais amplo, um banco de dados é um conjunto de dados integrados que atendem a um conjunto de sistemas ou uma comunidade de usuários (HEUSER, 1998).

Para manter um banco de dados são utilizados os Sistemas de Gerenciamento de Banco de Dados (SGBDs). Esses sistemas reúnem funções que são comuns a diversas aplicações de forma a facilitar o desenvolvimento destas. Segundo Silberschatz, Korth e Sudarshan (2006, p. 1), "[u]m Sistema de Gerenciamento de Banco de Dados [...] é uma coleção de dados inter-relacionados e um conjunto de programas para acessar esses dados". Heuser (1998, p. 4) define um SGBD como um "software que incorpora as funções de definição, recuperação e alteração de dados em um banco de dados".

2.2 Descoberta de conhecimento em bases de dados

O termo Descoberta de Conhecimento em Bases de Dados refere-se ao processo de busca e extração de conhecimento de grandes bases de dados. Fayyad, Piatetsky-Shapiro e Smyth (1996 apud GOLDSCHMIDT, 2005) definem o KDD como "[...] um processo, de várias etapas, não trivial, interativo e iterativo, para a identificação de padrões compreensíveis, válidos, novos, e potencialmente úteis a partir de grande conjunto de dados". Para Frawley, Piatetsky-Shapiro e Matheus (1992, p. 58, tradução do autor), "a descoberta de conhecimento é a extração não-trivial de informações implícitas, previamente desconhecidas e potencialmente úteis de dados"[2]. As informações extraídas são utilizadas "[...] para aumentar os ganhos, reduzir os custos ou melhorar o desempenho do negócio [...]", por exemplo, (THOMÉ, s.d., p. 11).

De acordo com Fayyad, Piatetsky-Shapiro e Smyth (1996), o KDD possui os seguintes passos (também apresentados na Figura 1):

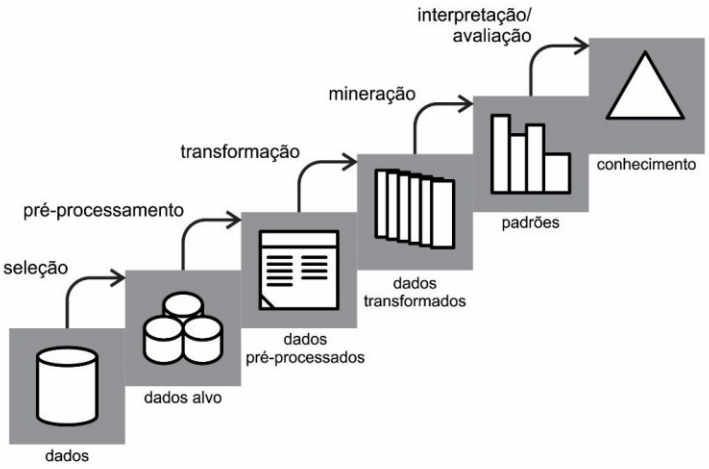

Figura 1: Visão geral dos passos que compõem o processo de KDD
Fonte: Adaptado de Fayyad, Piatetsky-Shapiro e Smyth (1996).

[1] "A database is a collection of data, typically describing the activities of one or more related organizations." (RAMAKRISHNAN; GEHRKE, 2000)
[2] "Knowledge discovery is the nontrivial extraction of implicit, previously unknown, and potentially useful information from data." (FRAWLEY; PIATETSKY-SHAPIRO; MATHEUS, 1992).

a) *Seleção:* é a definição do objetivo do processo do ponto de vista do usuário. Envolve a compreensão do domínio da aplicação e de conhecimentos prévios relevantes;
b) *Dados alvo:* a criação de um conjunto de dados alvo inclui a seleção de um conjunto de dados ou o foco num subconjunto de variáveis ou amostra de dados, no qual a descoberta será realizada;
c) *Pré-processamento:* pode incluir remoção de ruído, reunião de informações necessárias para a modelagem ou correção de ruído e decisão de estratégias para lidar com dados ausentes;
d) *Redução e projeção de dados*: consiste em encontrar características úteis para representar os dados dependendo do objetivo da tarefa. Com métodos de redução ou transformação da dimensionalidade, o número de variáveis consideradas pode ser reduzido ou representações invariantes para os dados podem ser encontradas;
e) *Decisão de método de mineração de dados:* escolha de um método específico (sumarização, classificação, regressão, agrupamento, etc.) que condiga com os objetivos do processo;
f) *Análise e modelo exploratório e seleção de hipótese:* escolha do algoritmo de mineração de dados e do método de seleção a ser usado na busca de padrões. Inclui a decisão de modelos e parâmetros apropriados.
g) *Mineração de dados:* é a procura de padrões de interesse numa forma representacional específica ou num conjunto dessas representações. Inclui regras ou árvores de classificação, regressão e agrupamento. Esse passo depende significantemente do bom desempenho dos anteriores;
h) *Interpretação/avaliação:* interpretação dos padrões minerados, com possível retorno a algum passo anterior para interação adicional. Pode envolver a visualização dos padrões e modelos extraídos ou a visualização dos dados tendo em conta os modelos extraídos;
i) *Ação sobre o conhecimento descoberto*: é o uso direto do conhecimento, incorporando-o em outro sistema para ação futura ou documentando-o e relatando para os interessados. Envolve a verificação e resolução de potenciais conflitos com conhecimentos anteriores.

2.3 Mineração de dados

Conforme dito acima, a Mineração de Dados é a principal etapa do KDD, quando ocorre a busca por padrões de interesse.

Para Silberschatz, Korth e Sudarshan (2006, p. 496), o termo mineração de dados "[...] refere-se, em geral, ao processo de analisar grandes bancos de dados de forma semiautomática para encontrar padrões úteis". Segundo Castro e Ferrari (2016), o termo mineração de dados diz respeito à exploração de uma base de dados por meio de algoritmos para obtenção do conhecimento, em uma alusão à extração de minerais preciosos (conhecimento) em uma mina (base de dados).

Na Mineração de Dados definem-se algoritmos e técnicas a serem aplicados no problema em questão. Redes neurais, algoritmos genéticos, modelos estatísticos e probabilísticos e árvores de decisão são exemplos de técnicas utilizadas (GOLDSCHMIDT; PASSOS, 2005; CASTRO; FERRARI, 2016).

Os dois principais objetivos da Mineração de Dados são a predição e a descrição. O primeiro envolve o uso de variáveis ou campos do banco de dados para predizer valores desconhecidos ou futuros de outras variáveis. O segundo trata da descoberta de padrões que descrevam os dados e sejam interpretáveis por humanos (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996).

Fayyad, Piatetsky-Shapiro e Smyth (1996) listam as seguintes tarefas de Mineração de Dados:

a) *Classificação:* descoberta de uma função que mapeie (classifique) um item de dados em um conjunto de classes pré-definidas;
b) *Regressão:* descoberta de uma função que mapeie um item de dados em uma variável de predição de valor real;
c) *Agrupamento (clusterização):* identificação de um conjunto finito de categorias (*clusters*) que descrevam os dados;
d) *Sumarização:* busca de uma descrição compacta para um subconjunto de dados;
e) *Modelagem de dependência:* busca de um modelo que descreva as dependências significativas entre as variáveis;
f) *Detecção de mudança e desvio:* descoberta das mudanças mais significativas nos dados a partir de valores normativos ou previamente medidos.

Goldschmidt e Passos (2005, p. 13) acrescentam ainda dois itens a essa lista: a descoberta de associação, que consiste na "[...] busca de itens que frequentemente ocorram de forma simultânea em transações do banco de dados"; e a descoberta de sequências, uma extensão da anterior, "[...] em que são buscados itens frequentes considerando-se várias transações ocorridas ao longo de um período".

A Mineração de Dados pode ser usada em análises de crédito, análise de concorrentes, descoberta de produtos que são comprados em conjunto, descobertas de causas etc. (SILBERSCHATZ; KORTH; SUDARSHAN, 2006). Carvalho (2005, p. 3) indica o uso para "[...] descobrir relações entre produtos, classificar consumidores, prever vendas, localizar áreas geográficas potencialmente lucrativas para novas filiais, inferir necessidades, entre outras".

2.4 Mineração de textos

A Mineração de Textos, também chamada de Mineração de Dados em Textos ou Descoberta de Conhecimento em Textos (*Knowledge Discovery in Texts* - KDT), é uma variante da Mineração de Dados. É o processo de extração de padrões ou conhecimento não-trivial, inesperado, útil e de interesse de documentos de texto não-estruturado (TAN, 1999).

Diferentemente da Mineração de Dados, a Mineração de Textos suporta conjuntos de dados não-estruturados e semiestruturados, como correios eletrônicos, arquivos HTML e documentos de texto em geral (VIJAYARANI; MUTHULAKSHMI, 2013). Oliveira et al. (2004) indicam usos da Mineração de Textos em estudos econômicos, descobrindo associações entre países e organizações ou realizando previsões sobre tecnologias, e na Internet, revelando associações, autoridades e eixos (*hubs*) de alguma área, por exemplo.

Com o aumento do uso da Internet e a necessidade de técnicas especializadas para atender a este domínio, surgiram variações tais como: mineração web (LIU, 2007) e mineração de opiniões em redes sociais (JAVA, 2007; LIU, 2010; LIU, 2012).

2.5 Descoberta de regras de associação

A descoberta de associação é uma das tarefas de mineração de dados. De acordo com Castro e Ferrari (2016, p. 235), "[a] *mineração de regras de* associação é uma técnica usada na construção de relações sob a forma de regras entre itens de uma base de dados transacional". Uma regra de associação possui o formato $X \rightarrow Y$, onde X e Y são conjuntos de itens de um conjunto de transações e não compartilham itens em comum (CASTRO; FERRARI, 2016; ZHENG; KOHAVI; MASON, 2001). Uma regra de associação

pode ser lida como "se *X* então *Y*" ou "*X* implica em *Y*" (BOOK, ZHENG; KOHAVI; MASON, 2001).

Como nem toda regra de associação descoberta num conjunto de transações é relevante, costuma-se indicar ao algoritmo de mineração certas limitações, como suporte ou confiança mínimos.

O suporte de um conjunto de itens é a porcentagem de transações no conjunto de transações que contém o conjunto de itens. Da mesma forma, o suporte de uma regra é a porcentagem de transações que a satisfazem. Um conjunto de itens ou uma regra é frequente quando seu suporte é maior que o suporte mínimo (ZHENG; KOHAVI; MASON, 2001).

A confiança de uma regra é "o número de transações que ela prediz corretamente proporcionalmente às transações para as quais ela se aplica" (CASTRO; FERRARI, 2016, p. 236). Equivale à probabilidade de uma transação conter *Y* se ela contém X, para uma regra $X \rightarrow Y$ (ZHENG; KOHAVI; MASON, 2001). Uma regra é segura se sua confiança é mais alta que uma confiança mínima dada (ZHENG; KOHAVI; MASON, 2001).

2.6 O algoritmo 'APRIORI'

De acordo com Castro e Ferrari (2016), a descoberta de regras de associação pode ser executada por diferentes algoritmos, muitos deles variações do pioneiro Apriori.

O Apriori foi proposto como algoritmo de descoberta de regras de associação por Agrawal e Srikant em 1994 (SUMITHRA; PAUL, 2010). Ele emprega busca em profundidade e gera conjuntos de itens candidatos de k elementos a partir de conjuntos com $k - 1$ elementos (CASTRO; FERRARI, 2016). O algoritmo é composto de duas etapas: geração de conjunto de itens frequentes e geração das regras (CASTRO; FERRARI, 2016; ZHENG; KOHAVI; MASON, 2001).

O Apriori segue o princípio de que, se um conjunto de itens é frequente, qualquer subconjunto seu também é frequente. De outra maneira, se um conjunto de itens não é frequente, qualquer sucessor seu também não o é (CASTRO; FERRARI, 2016; ZHENG; KOHAVI; MASON, 2001). Assim, o algoritmo utiliza apenas os conjuntos itens frequentes de $k - 1$ elementos para encontrar os conjuntos frequentes de k elementos, aumentando sua eficiência (CASTRO; FERRARI, 2016; SUMITHRA; PAUL, 2010).

3 METODOLOGIA

A metodologia de trabalho foi composta pela criação de um banco de dados de artigos sobre mineração de texto recuperados da Web of Science, identificação de características dos dados e aplicação do algoritmo APRIORI de mineração de dados.

3.1 Criação do banco de dados

A criação do banco de dados realizou-se em quatro etapas: pesquisa, exportação, seleção e inserção, conforme apresentado na Figura 2 e explicado na sequência.

Figura 2: Metodologia adotada para criação do banco de dados
Fonte: Elaborado pelos autores.

Na primeira etapa, buscou-se por artigos na coleção principal da WoS com a sintaxe *"text mining" OR "text based" AND "information retrieval" NOT sentiment* AND tool* NOT emotion* NOT opinion** no dia 1 de março de 2016 e obtiveram-se 2.829 resultados. Após o uso de filtros (Quadro 1), recuperaram-se 1.193 referências condizentes com o escopo da pesquisa.

Quadro 1: Filtros utilizados na pesquisa

<table>
<tr><th>Categorias do Web of Science</th><th>Áreas de pesquisa</th></tr>
<tr><td rowspan="4">Business
Communication
Computer Science Artifical Intelligence
Computer Science Cybernetics
Computer Science Information Systems
Computer Science Interdisciplinary Applications
Computer Science Software Engineering
Computer Science Theory Methods
Engineering Multidisciplinary
Ergonomics
Information Science Library Science
Management
Mathematical Computational Biology
Mathematics Applied
Mathematics Interdisciplinary Applications
Multidisciplinary Sciences
Operations Research Management Science
Social Sciences Interdisciplinary
Social Sciences Mathematical Methods
Statistics Probability</td><td>Business Economics
Communication
Computer Science
Engineering
Information Science Library Science
Mathematical Computational Biology
Mathematical Methods In Social Sciences
Mathematics
Medical Informatics
Operations Research Management Science
Science Technology Other Topics
Social Sciences Other Topics</td></tr>
<tr><td>Índices</td></tr>
<tr><td>SCI-EXPANDED
SSCI
CPCI-S
CPCI-SSH
ESCI</td></tr>
<tr><td>Idiomas</td><td>Tempo estipulado</td></tr>
<tr><td>English
Portuguese
Spanish</td><td>2010-2016</td></tr>
</table>

Fonte: Elaborado pelos autores, com base nos filtros disponíveis na WoS em 01 mar. 2016.

Na etapa de exportação, as referências foram salvas em 3 arquivos formato

BibTeX (bib), uma vez que a interface WoS possui um limite de 500 referências por exportação. O BibTeX foi escolhido devido a familiaridade do bolsista com a linguagem.

Na seleção, os dados foram tratados por uma aplicação desenvolvida em linguagem C, utilizando-se o compilador do ambiente Minimalist GNU for Windows (MinGW). A aplicação tem o objetivo de recuperar os dados (pelos campos selecionados) no arquivo BibTeX, identificando-os pelas marcações da linguagem, com o objetivo de alimentar um banco de dados que será utilizado para a Mineração de Dados.

As ferramentas e métodos adotados em cada trabalho (documento recuperado) não são apontados pelo BibTeX em um campo exclusivo. Em tais casos, a aplicação pesquisou no campo *abstract* por termos de listas (de ferramentas e de métodos) elaboradas previamente. A lista de ferramentas baseou-se nos sites KDnuggets e Predictive Analytics Today, em um total de 138 ferramentas pesquisadas em 15 de março de 2016. A lista de métodos foi concebida usando os métodos disponíveis no programa Weka (versão 3.8), em um total de 66 métodos. A busca por ferramentas também foi feita no campo *funding-text.*

A inserção dos dados recuperados pela aplicação proposta ocorreu no banco de dados que havia sito previamente elaborado utilizando o MySQL Server versão 5.7.9 para sistema Windows.

3.2 Identificação de características

Após a concepção, o banco de dados foi explorado a fim de identificar características dos registros. As análises foram realizadas por meio do uso de comandos SQL utilizando o MySQL Command Line Client 5.7. As listas resultantes foram copiadas para arquivos CSV para posterior análise em linguagem R.

Além disso, preparou-se uma planilha para relacionar os periódicos e termos de indexação. A elaboração da planilha baseou-se em lista de termos utilizados por periódico recuperada do banco de dados. Os dados dessa planilha foram utilizados como alvo da descoberta de associação com o Apriori.

3.3 Descoberta de associação com APRIORI

Com base nas características levantadas, o método escolhido para a mineração do banco de dados foi a descoberta de associação usando o algoritmo Apriori na linguagem R (versão 3.3.1). O método foi aplicado na planilha de periódicos e termos criada anteriormente.

No ambiente da linguagem R, a biblioteca *arules* foi utilizada para a aplicação do Apriori com suporte mínimo de 5% e confiança mínima de 90%, conforme Figura 3. O valor do suporte mínimo foi escolhido considerando a diversidade dos dados da planilha. São 1881 termos de indexação e 280 periódicos; o maior suporte entre os termos é de 23,5714% e o menor é de 0,3571%; o suporte médio é de 0,9087%.

```
pt = read.csv("periodicos-termos.csv", sep=";")
library("arules")
regras = apriori(as.matrix(pt), parameter = list(supp = 0.05, conf = 0.9, minlen = 2))
inspect(regras)’
```

Figura 3: Aplicação do apriori na ferramenta R
Fonte: Elaborado pelos autores.

Foram obtidas 13 regras de associação, conforme indicação e análise na seção de resultados.

4 RESULTADOS: APRESENTAÇÃO E DISCUSSÃO

O banco de dados foi criado contendo 12 tabelas, sendo 5 agregadoras. Para que fossem atingidos os objetivos definidos nesta etapa da pesquisa, as tabelas referentes às fontes de dados citados nos documentos recuperados não foram alimentadas, assim como os atributos relacionados a objetivos, limitações, considerações finais e sugestões de trabalhos futuros da tabela *artigo*.

Dentre as 138 ferramentas pesquisadas, foram encontradas 11, conforme Tabela 1. A linguagem R foi a mais citada, mais que o triplo das referências ao Matlab e ao SAS, ambos em segundo lugar.

Tabela 1: Ferramentas citadas nos artigos

Ferramenta	Frequência	Ferramenta	Frequência
R	29	Leximancer	4
Matlab	7	Gate	3
SAS	7	Knime	2
Fraunhofer Idea	5	Outros	4

Fonte: Elaborado pelos autores.

Foram encontrados 24 métodos dentre os 66 pesquisados (Tabela 2). A classificação foi o mais encontrado, com mais que o dobro de citações que a associação, segundo lugar. Agrupamento aparece em terceiro.

Para fins desta pesquisa, a tarefa de Associação é a mais adequada, uma vez que está se realizado uma tarefa descritiva da base de dados. Desta forma, o algoritmo usado foi o Apriori.

Tabela 2: Métodos citados nos artigos

Método	Frequência	Método	Frequência
Classification	176	Logistic Regression	13
Association	81	Markov	11
Cluster	63	SVM	11
Bayes	30	Decision Tree	8
Naive Bayes	26	C4.5	6
Regression	23	C4	6
Suport Vector Machine	22	Boosting	4
Kmeans	20	Outros	13

Fonte: Elaborado pelos autores.

Os artigos no banco de dados utilizaram 2602 palavras-chave diferentes. Na Tabela 3 são listadas as mais frequentes. A palavra-chave mais encontrada foi *text mining*, por ter sido um dos termos utilizados na pesquisa ao WoS. Em seguida encontra-se *natural language processing* e *data mining*.

Tabela 3: Palavras-chave mais utilizadas

Palavra-chave	Frequência	Palavra-chave	Frequência
text mining	747	sentiment analysis	42
natural language processing	68	information retrieval	41
data mining	65	clustering	41
machine learning	60	classification	39
information extraction	58	text classification	36

Fonte: Elaborado pelos autores.

Os registros salvos no banco de dados contêm 1881 termos de indexação (*keywords-plus*) diferentes. Na Tabela 4 são listados os mais frequentes. O termo *information* foi o mais encontrado, seguido de *text* e *classification*.

Tabela 4: Termos mais utilizados

Termo	Frequência	Termo	Frequência
information	147	extraction	72
text	136	knowledge	63
classification	100	model	62
system	90	retrieval	57
database	78	web	56

Fonte: Elaborado pelos autores.

O total de periódicos no banco de dados é de 280. A Tabela 5 mostra aqueles com maior número de artigos. BMC Bioinformatics foi o periódico com mais artigos no banco de dados, seguido de Expert Systems with Applications e Journal of Biomedical Informatics.

Tabela 5: Periódicos com maior número de artigos

Periódico	Frequência	Periódico	Frequência
BMC Bioinformatics	83	Bioinformatics	44
Expert Systems with Applications	71	Scientometrics	34
Journal of Biomedical Informatics	54	IEEE Transactions on Knowledge and Data Engineering	31
Plos	53	Decision Support Systems	25
Database - The Journal of Biological Databases and Curation	48	Journal of the American Medical Informatics Association	24

Fonte: Elaborado pelos autores.

Os artigos no banco de dados foram publicados em 24 países diferentes (Tabela 6). O país com maior número de publicações é a Inglaterra, seguida pelos Estados Unidos da América e pelos Países Baixos.

Tabela 6: Países de publicação dos artigos

País	Frequência	País	Frequência
England	450	Austria	12
USA	423	Singapore	11
Netherlands	198	Japan	9
Switzerland	28	People R China	7
Germany	19	Outros	36

Fonte: Elaborado pelos autores.

Foram identificados 3.472 autores diferentes para os artigos. A Tabela 7 lista os mais frequentes, dentre os quais destaca-se Sophia Ananiadou, diretora do The National Centre for Text Mining (NaCTeM), professora pesquisadora na Ciência da Computação na Universidade de Manchester e uma das principais pesquisadoras sobre o assunto Text Mining no mundo, foi a autora com maior número de artigos no banco de dados, seguida por Zhiyong Lu (professor pesquisador do Biomedical Text Mining Group do National Center for Biotechnology Information (NCBI)), Karin Versoor (professora pesquisadora na Universidade de Melbourne sobre os temas mineração de textos em publicações científicas e textos clínicos da área de Biomédica) e Cathy H. Wu (professora pesquisadora na Universidade de Delaware do departamento de Ciências Biológicas na área de mineração de textos de publicações científicas na área Biológica, dentre outros interesses).

Tabela 7: Autores com maior número de artigos

Autor	Frequência	Autor	Frequência
Ananiadou, Sophia	23	Nenadic, Goran	12
Lu, Zhiyong	20	Arighi, Cecilia N.	12
Verspoor, Karin	13	Wei, Chihhsuan	11
Wu, Cathy H.	13	Kao, Hungyu	11
Rinaldi, Fabio	12	Wilbur, W. John	11

Fonte: Elaborado pelos autores.

Após a aplicação do Apriori na planilha de periódicos e termos, foram obtidas 13 regras de associação, conforme Tabela 8. Essas regras indicam a possibilidade de um periódico ter utilizado um termo de indexação *Y*, dado que utilizou um termo *X*, considerando-se os critérios da pesquisa feita ao WoS (categorias, índices, áreas de pesquisa, idiomas e data de publicação).

A regra com o maior suporte prevê que quando um periódico utiliza os termos *text* e *classification*, também utilizará o termo *information*. Essa regra tem 6,4516% de suporte. Ou seja, ocorreu em 6,4516% dos periódicos. Sua confiança é de 90%. Em outras palavras, 90% dos periódicos que usaram *text* e *classification* também usaram *information*.

Tabela 8: Regras de associação resultantes do apriori

Antecedente	Consequente	Suporte (%)	Confiança (%)
text, classification	information	6,4516	90
text, knowledge	information	5,7348	94,1177
classification, knowledge	information	5,7348	94,1177
text, system, classification	information	5,7348	94,1177
system, information, classification	text	5,7348	94,1177
retrieval, knowledge	information	5,3763	93,75
extraction, database	system	5,0179	100
extraction, system, information	text	5,0179	100
classification, retrieval, knowledge	information	5,0179	93,3333
disease	information	5,0179	93,3333
identification, retrieval	information	5,0179	93,3333
text, extraction, information	system	5,0179	93,3333
information, retrieval, knowledge	classification	5,0179	93,3333

Fonte: Elaborado pelos autores.

5 CONSIDERAÇÕES FINAIS

A exploração dos dados de artigos provenientes do WoS revelou características dos estudos sobre Mineração de Textos publicados entre 2010 e 2016. Identificaram-se periódicos, autores e países com publicações na área; palavras-chave e termos de indexação utilizados nos artigos; e ferramentas e métodos citados neles. A aplicação do algoritmo Apriori indicou associações entre os termos de indexação utilizados por cada periódico.

A aplicação desenvolvida para o estudo poderá ser utilizada em trabalhos semelhantes que necessitem do levantamento de dados bibliográficos. Porém, no caso de exploração de dados não-estruturados pode ser necessária a utilização de outro método. Por exemplo, os campos *fontedados* e *artigo_fontedados* e os atributos *objetivogeral*, *limitacoes*, *consideracoesfinais* e *trabalhosfuturos* da tabela *artigo* do banco de dados não foram alimentados por dificuldades na identificação de tais dados.

Trabalhos futuros podem aplicar outros métodos de mineração em dados provenientes do WoS. Considerando-se as indicações recuperadas do banco de dados, métodos de classificação, como o Bayes, ou de agrupamento, como o K-médias, podem ser empregados. Ainda, para a continuidade do estudo, as tabelas referentes às fontes de dados citados nos documentos recuperados, atributos relacionados a objetivos, limitações, considerações finais e sugestões de trabalhos futuros serão recuperados (alimentação da base de dados) para posterior análise.

REFERÊNCIAS

AMO, S. de. **Técnicas de mineração de dados.** s.l.: Universidade Federal de Uberlândia, s.d. Disponível em: <https://www.researchgate.net/profile/Sandra_Amo/publication/260300816_Tcnicas_de_Minerao_de_Dados/links/54230bd80cf290c9e3ae25e3.pdf>. Acesso em: 26 jan. 2016.

CARVALHO, L. A. V. de. **Datamining**: a mineração de dados no marketing, medicina, economia, engenharia e administração. Rio de Janeiro: Ciência Moderna, 2005.

CASTRO, L. N. de; FERRARI, D. G. **Introdução à mineração de dados**: conceitos básicos, algoritmos e aplicações. São Paulo: Saraiva, 2016.

COSTA, C. N. et al. Descoberta de Conhecimento em Bases de Dados**. Revista Eletrônica: Faculdade Santos Dumont**, 2 ed., s.d. Disponível em: <http://fsd.edu.br/revistaeletronica/arquivos/2Edicao/artigo9.pdf>. Acesso em: 26 jan. 2016.

FAYYAD, U.; PIATETSKY-SHAPIRO, G. SMITH, P. From datamining to knowledge discovery in databases. **AI Magazine**, v. 17, n. 3, p. 37-54, 1996.

FRAWLEY, W. J.; PIATETSKY-SHAPIRO, G.; MATHEUS, C. J. Knowledge discovery in databases: an overview. **AI Magazine**, v. 13, n. 3, p. 57-70, 1992. Disponível em: <http://www.aaai.org/ojs/index.php/aimagazine/article/viewFile/1011/929>. Aceso em: 21 jan. 2016.

GOLDSCHIMIDT, R.; PASSOS, E. **Data mining**: um guia prático. Rio de Janeiro: Elsevier, 2005.

HEUSER, C. A. **Projeto de banco de dados**. 4 ed. [Porto Alegre]: Sagra Luzzatto, 1998. Disponível em: <http://www.julianoribeiro.com.br/troca/banco_de_dados/material_der.pdf>. Acesso em: 05 fev. 2016.

JAVA, A. et al. Why we twitter: understanding microblogging usage and communities. In: WORKSHOP ON WEB MINING AND SOCIAL NETWORK ANALYSIS, 9., 2007, Estados Unidos. **Proceedings of** ... Estados Unidos: San Jose, 2007.

LIU, Bing. **Web data mining**: exploring hyperlinks, contents, and usage data. Springer Science & Business Media, 2007.

_____. Sentiment analysis and subjectivity. **Handbook of Natural Language Processing**, v. 2, p. 627-666, 2010.

_____. Sentiment analysis and opinion mining. **Synthesis Lectures on Human Language Technologies**, v. 5, n. 1, p. 1-167, 2012.

MUGNAINI, R.; STREHL, L. Recuperação e impacto da produção científica na era Google: uma análise comparativa entre o Google Acadêmico e a Web of Science. **Encontros Bibli**, Florianópolis, n. esp., 1º sem. 2008.

OLIVEIRA, J. P. M. de et al. Applying Text Mining on electronic messages for Competitive Inteligence. In: INTERNATIONAL CONFERENCE ON ELECTRONIC COMMERCE AND WEB TECHNOLOGIES, 5., 2004, Spain. **Proceedings** ... Spain: Zaragoza, 2004. Disponível em: <https://www.researchgate.net/profile/Leandro_Wives/publication/221017413_Applying_Text_Mining_on_Electronic_Messages_for_Competitive_Intelligence/links/09e41510bbc3323c4100000.pdf>. Acesso em: 27 jan. 2016.

RAMAKRISHNAN, R; GEHRKE, J. **Database management systems**. s.l.: s.n., [2000]. Disponível em: <http://dspace.utamu.ac.ug:8080/xmlui/bitstream/handle/123456789/85/%5BRamakrishnan_R.,_Gehrke_J.%5D_Database_Management_S(BookFi.org).pdf>. Acesso em: 15 fev. 2016.

SILBERSCHATZ, A.; KORTH, H. F.; SUDARSHAN, S. **Sistema de banco de dados**. Rio de Janeiro: Elsevier: 2006. (tradução de Daniel Vieira)

SUMITHRA, R.; PAUL, S. Using distributed apriori association rule and classical apriori mining algorithms for grid based knowledge discovery. In: SECOND INTERNATIONAL CONFERENCE ON COMPUTING, COMMUNICATION, AND NETWORKING TECHNOLOGIES, 2010, India. **Proceedings of**... Índia, 2010. Disponível em: <http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5591577>. Acesso em: 03 ago. 2016.

TAN, A-H. Text Mining: the state of the art and the challenges. In: WORKSHOP ON KNOWLEDGE DISCOVERY FROM ADVANCED DATABASES, 1999. **Proceedings of** ... 1999. Disponível em: <http://www.ntu.edu.sg/home/asahtan/papers/tm_pakdd99.pdf>. Acesso em: 21 jan. 2016.

THOMÉ, A. C. G. **Redes neurais:** uma ferramenta para KDD e Data Mining. s.l.: [Universidade Federal do Rio de Janeiro], s.d. (Apostila). Disponível em: <http://equipe.nce.ufrj.br/thome/grad/nn/mat_didatico/apostila_kdd_mbi.pdf>. Acesso em: 26 jan. 2016.

VIJAYARANI, S.; MUTHULAKSHMI, M. Comparative analysis of Bayes and Lazy classification algorithms. **International Journal of Advanced Research in Computer and Communication Engineering**, v. 2, n. 8, ago. 2013.

YONG-HAK, J. Web of Science. **Thomson Reuters**, 2013.

ZHENG, Z.; KOHAVI, R.; MASON, L. Real world performance of association rule algorithms. In: INTERNATIONAL CONFERENCE ON KNOWLEDGE DISCOVERY AND DATA MINING, 2010, New York. **Proceedings of** … New York: ACM, 2001. Disponível em: <http://robotics.stanford.edu/users/ronnyk.link/realWorldAssocLongPaper.pdf> Acesso em: 03 ago. 2016.