

Uma Preliminar Revisão Bibliométrica utilizando Mineração de Dados

Thayanne França¹, Pamella Soares¹, Junior Ferro¹, Luana Brasil¹

¹ Universidade Estadual do Ceará (UECE)
Fortaleza – CE – Brasil

1. Introdução

A mineração de dados é um campo interdisciplinar que combina tecnologia de máquina do conhecimento, reconhecimento de padrões, estatísticas, bancos de dados e visualização para extrair informações e conhecimento de bancos de dados com grandes quantidades de dados [Hirji 1999]. Esta área tem sido de grande importância pois tem proporcionado o desenvolvimento de diferentes domínios, como medicina, finanças, segurança, negócios, dentre outros [Camilo and Silva 2009], nos setores tanto público como privado.

A classificação da mineração de dados pode ser dada a partir das possíveis tarefas que esta tem a capacidade de realizar. Neste caso, a mesma pode ser dividida nas tarefas mais comuns como: descrição, classificação, estimação/regressão, predição, agrupamento e associação. Tais tarefas podem ser aplicadas em problemas reais sendo possível obter diferentes descobertas. Na tarefa de classificação pode-se determinar quando uma transação de cartão de crédito pode ser uma fraude, por exemplo. Em predição, há a possibilidade de prever o valor de uma ação determinados meses adiante. É possível determinar os casos onde um novo medicamento pode apresentar efeitos colaterais, ao aplicar regras de associação, por exemplo [Camilo and Silva 2009].

Assim sendo, tendo em vista a diversidade de estudos envolvendo a área de Mineração de Dados o presente trabalho busca realizar um levantamento de quais técnicas os pesquisadores brasileiros tem utilizado para desenvolver pesquisa na área em questão. Para isso, vislumbra-se realizar uma preliminar “Revisão Bibliométrica” do estado da arte sobre Mineração de Dados no Brasil utilizando técnicas de Mineração de Dados em bases científicas. A presente pesquisa também terá como resultado a geração de um *dataset* com publicações realizadas entre anos de 2016 até 2020, pois sobre o mesmo será realizado o processo de mineração.

O restante do trabalho é organizado da seguinte forma: na Seção 2, apresentam-se resumidamente os conceitos base do trabalho, na Seção 3 apresentam-se os Trabalhos Relacionados. A metodologia e os processos são descritos na Seção 4. Na Seção 5 são apresentados os resultados, discussões, bem como a geração de conhecimento sobre estes. Finalmente, na Seção 6 destacam-se as considerações finais.

2. Fundamentação Teórica

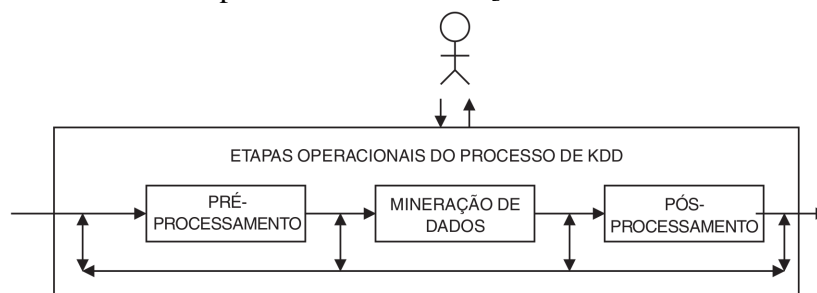
Nesta seção serão apresentadas os conceitos básicos que permeiam o presente trabalho, como o processo de Descoberta de Conhecimento em Base de Dados e algumas das técnicas implementadas.

2.1. *Knowledge Discovery in Databases (KDD)*

A área denominada “Descoberta de Conhecimento em Bases de Dados” (do inglês *Knowledge Discovery in Databases – KDD*) tem sido amplamente estudada tendo em vista

a grande demanda de dados gerados e armazenados em diferentes bases de dados, tais informações advindas dos mais variados domínios e aplicações. Em sua forma mais popular, o termo “Mineração de Dados” é considerado por muitos como sinônimo de KDD, enquanto que para outros consiste em uma das fases do processo do KDD [Han et al. 2011], como mostra resumidamente a Figura 1 a seguir. A fase de mineração dos dados ocorre depois da fase de pré-processamento dos dados, que está relacionada à captação, organização e tratamento dos dados coletados, e antes da fase de pós-processamento, considerado por Goldschmidt e Passos (2005) [Goldschmidt and Passos 2005], o “tratamento do conhecimento obtido na Mineração de Dados”.

Figure 1. Processos Operacionais do KDD [Goldschmidt and Passos 2005].



Em suma, a fase Mineração de Dados pode ser definida como um processo pelo qual determinado conhecimento pode ser extraído de grandes volumes de dados [de Amo 2004]. Hand e Adams (2014) a definem como sendo:

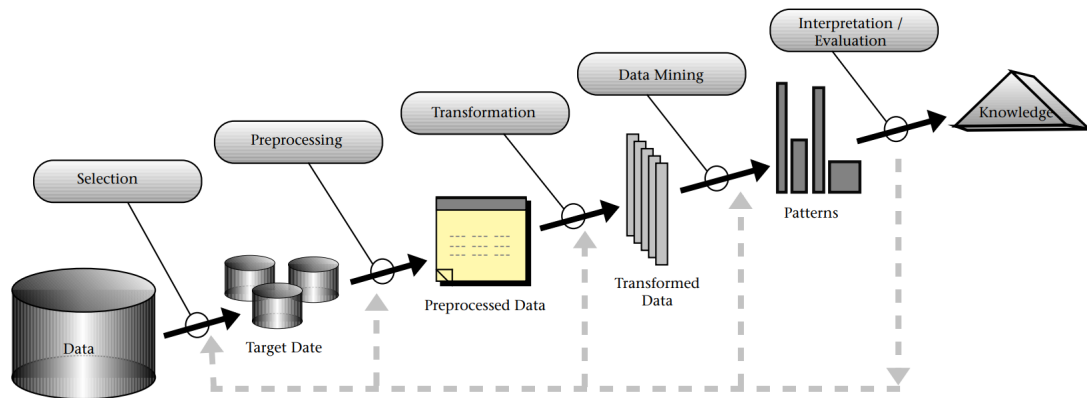
“a tecnologia de descobrir estruturas e padrões em grandes conjuntos de dados. [...] a disciplina tem uma sobreposição substancial com outras disciplinas de análise de dados, especialmente estatística, aprendizado de máquina e reconhecimento de padrões [...] cada uma dessas disciplinas tem sua própria ênfase. A mineração de dados, em particular, se distingue dessas outras disciplinas pelo (grande) tamanho dos conjuntos de dados, frequentemente pela baixa qualidade dos dados e pela amplitude do tipo de estruturas procuradas”

Como já mencionado, o KDD consiste em uma sequência de passos além da aplicação das técnicas de mineração dos dados. A Figura 2 [Fayyad et al. 1996] apresenta que, após a coleta dos dados em diferentes banco de dados, ocorrem as fases necessárias antes da extração do conhecimento. Ressalta-se que o processo KDD é interativo e iterativo, dependendo das decisões feitas pelo usuário, de maneira que seja possível repetir passos anteriores do processo.

Han et al. (2011) elenca os passos desse processo como descrito a seguir. Entre os Passos 1 e 4 ocorre o pré-processamento das informações coletadas de diferentes maneiras.

1. **Limpeza de dados:** a fim de remover ruídos e dados inconsistentes e irrelevantes.
2. **Integração de dados:** quando os dados coletados de diferentes fontes podem ser combinados.
3. **Seleção de dados:** quando os dados relevantes para a tarefa de análise são recuperados do banco de dados.

Figure 2. Uma visão geral das etapas que compõem o processo KDD [Goldschmidt and Passos 2005].



4. **Transformação de dados:** quando os dados são transformados e consolidados em formas apropriadas para a aplicação dos algoritmos de mineração.
5. **Mineração de dados:** um processo essencial onde métodos inteligentes são aplicados para extrair padrões de dados.
6. **Avaliação de padrões:** para identificar os padrões verdadeiramente interessantes que representam o conhecimento com base em medidas de interesse.
7. **Apresentação do conhecimento:** quando técnicas de visualização e representação do conhecimento são usadas para apresentar o conhecimento extraído aos usuários.

2.2. Técnicas de Mineração de Dados

As técnicas de mineração consistem em um conjunto de passos a serem realizados a fim de descobrir padrões que interessam em determinado estudo. Dentre as principais técnicas utilizadas em mineração de dados, tem-se técnicas estatísticas, técnicas de aprendizado de máquina e técnicas baseadas em *crescimento-poda-validação* [de Amo 2004]. A seguir algumas das técnicas/áreas que serão utilizados no presente trabalho.

Processamento de Linguagem Natural (PLN). Este campo inclui a construção de modelos computacionais para realizar tarefas que dependem de informações expressas em uma determinada linguagem natural (por exemplo, tradução e interpretação de texto, busca de informações em documentos e interfaces homem-máquina, dentre outros)¹.

Locality-Sensitive Hashing (LSH). Funções de Hashing Sensíveis à Localidade mapeiam dados similares com alta probabilidade de ter o mesmo código *hash* nos mesmos *buckets*, reduzindo assim a dimensionalidade. Uma família de funções LSH é um conjunto F de funções hash tal que para dois objetos x e y , $\Pr_{h \in F}[h(x) = h(y)] = \text{sim}(x, y)$, onde $\text{sim}(x, y) \in [0, 1]$ é uma função de similaridade sobre universo de objetos [Mourão et al. 2016]. O LSH é usado para realizar pesquisas por vizinhos mais próximos com base em um conceito simples de *similaridade*. Pode-se dizer que dois itens são semelhantes se a interseção de seus conjuntos for suficientemente grande. Tal descrição advém da noção de Similaridade de Conjuntos de Jaccard [Real and Vargas 1996], onde similaridade de Jaccard é definida como a interseção

¹<https://www.ime.usp.br/siago/IA-pln.pdf>

de dois conjuntos dividida pela união de ambos os conjuntos. Em suma, a aplicação do LSH passa por três etapas: (i) *Slingling* – extração de um conjunto de strings de comprimento k que advindas de um documento; (ii) *Min-Hashing* – que tem o objetivo de substituir um grande conjunto por uma “assinatura” menor que preserve a métrica de similaridade, e por fim (iii) o LSH – que extrai todos os pares de documentos semelhantes.

3. Trabalhos Relacionados

Alguns trabalhos correlatos sobre mineração em comunidades de pesquisas científicas serão apresentados a seguir, alguns dos artigos abrangem domínios mais gerais em comparação ao presente estudo.

Martino et al. (2009) introduzem o sistema KAIROS, capaz de coletar automaticamente dados sobre publicações científicas, teses e patentes para realizar a prospecção tecnológica e identificação de especialistas em determinadas áreas. Assim, o sistema é capaz de inferir o comportamento das áreas ao longo do tempo de forma a perceber àquelas que são promissoras. Os dados das fontes Curriculum Lattes, DBLP, Base de Teses e Dissertações da Capes e Base de Patentes da *Derwent* foram filtrados e estruturados para realização do processo de mineração. Araujo et al. (2015) propuseram um método de análise de comunidades científicas com base em publicações científicas, as quais são analisadas por meio de técnicas de mineração de texto para identificar o contexto da publicação e gerar sua rede de colaboração. Os autores procuram entender a cooperação e parceria entre pesquisadores da área, bem como temas de interesse comum, e o campo que aparenta estar convergindo. Para isso, a mineração do repositório ocorre de forma a gerar dois tipos de rede de colaboração: (i) colaboração entre autores e (ii) colaboração entre as instituições as quais os autores estão vinculados.

Torres (2011) propôs um método de sumarização de artigos científicos em engenharia de software, com foco na localização e identificação de termos que representem os resultados do trabalho científico analisado, de forma a reduzir o tempo despendido pelos pesquisadores em revisões sistemáticas. Tal trabalho identificou padrões estruturais em artigos científicos, avaliou técnicas de comparação em relação aos termos e avaliou a precisão do método proposto. Por sua vez, com o propósito semelhante ao presente estudo, Bezerra e Guimarães (2017) realizaram também a mineração de textos mas, dessa vez, buscando nas bases de dados o tema ‘Gestão de Conhecimento’ (GC). Os autores tiveram como objetivos principais a busca pelos termos mais empregados nesse campo de estudo e a avaliação das diferenças entre os termos encontrados em publicações brasileiras e estrangeiras. Para a mineração dos textos, os autores utilizaram uma ferramenta denominada PreText 2.

4. Metodologia

Nesta seção é descrito como foi realizada a criação do *dataset*, juntamente com o pré-processamento das informações coletadas e, por fim, aplicação dos Algoritmos de Mineração.

4.1. Coleta dos Dados

O processo iniciou-se com a fase de Coleta dos Dados na qual foram realizadas as buscas dos artigos nas bases de dados científicas. Nesta etapa, as buscas pelos artigos foram

realizadas de forma semi-automática, aplicando as strings de buscas. Como o objetivo é analisar o contexto de pesquisa sobre mineração de dados no Brasil, foram pré-definidas algumas restrições para a busca dos artigos. Dessa forma, foram definidos dois *datasets*: (i) o *Dataset 1* no qual contém resultado da busca apenas da string de busca “data mining” e o (ii) *Dataset 2* onde foram adicionados artigos advindos de uma busca mais específica incluindo palavras chaves relacionadas ao Brasil, como será mostrado a seguir.

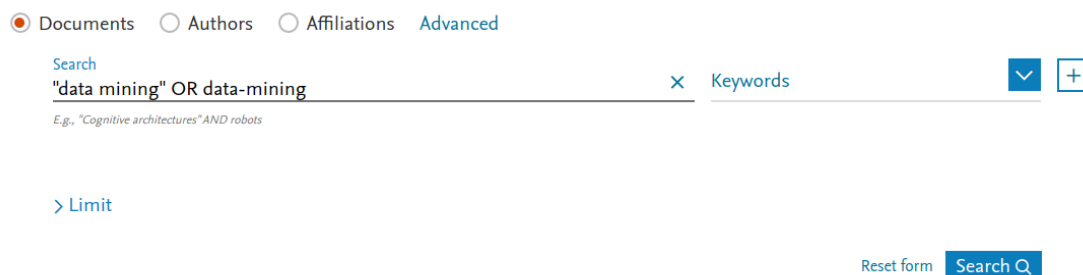
Adicionalmente, dentre as bases científicas existentes, a equipe escolheu somente duas, visto que foi observado que apenas estas traziam em lote e de forma automática informações essenciais para a presente pesquisa. Tais informações estão relacionadas à afiliações/universidades, país dos autores, dentre outras. Em um contexto geral, pretendeu-se também realizar uma análise da pesquisa de mineração de dados em relação à outros países. Para isso, uma amostra de artigos envolvendo outros países também foi coletada, sendo esta última adicionada ao *Dataset 1*.

As duas bases de dados científicas utilizadas para a coleta de dados foram:

- **Scopus Elvevier:** um banco de dados de resumos e citações de artigos para jornais/revistas acadêmicos, e
- **IEEE Xplore:** um banco de dados de pesquisa para descoberta e acesso a artigos de periódicos, anais de conferências, normas técnicas e materiais relacionados em ciência da computação, engenharia elétrica e eletrônica e campos afins.

Na busca para composição do *Dataset 1* foi utilizada a string de busca ```data mining`` OR data-mining` especificamente no campo *Keywords* e *Author Keywords* para Scopus e IEEE, respectivamente, como apresentado nas Figuras 3 e 5. Ou seja, os artigos retornados deveriam ter um desses termos em suas *keywords*. Ambas as bases científicas retornam um lote de contendo os 2000 primeiros artigos para cada busca. Com objetivo de aumentar o número da amostra essa mesma busca foi feita para os anos entre 2010 e 2020. Visto que para cada ano foram retornados 2000 trabalhos, o total foi de 44000 retornos, 22000 da base Scopus e 22000 da base IEEE.

Figure 3. Busca na Scopus nas Keywords.



The image shows the Scopus search interface. At the top, there are radio buttons for 'Documents' (selected), 'Authors', and 'Affiliations', followed by a link to 'Advanced'. Below this is a search bar with the text 'Search' and a dropdown menu set to 'Keywords'. The search query entered is '"data mining" OR data-mining'. Below the search bar, there is a hint: 'E.g., "Cognitive architectures" AND robots'. To the left of the search bar is a link '> Limit'. To the right of the search bar are two buttons: 'Reset form' and 'Search Q'.

Na segunda pesquisa para composição do *Dataset 2* foi utilizada a string de busca ```data mining`` OR data-mining) AND (brazil OR brasil OR brazilian)`, sendo os primeiros termos no campo *Keywords* e *Author Keywords* para Scopus e IEEE, respectivamente (Figuras 3 e 5), e o segundo termo para as afiliações dos autores dos artigos encontrados. Neste segundo termo, a busca está considerando que

Figure 4. Busca na IEEE nas Keywords.

The image shows a search interface with three rows of search criteria. The first row has a search term "data mining" OR data-mining in the field "Author Keywords". The second and third rows have "AND" in a dropdown menu, a "Search Term" input field, and "All Metadata" in the field dropdown. To the right of each row are navigation icons: a question mark, up/down arrows, and a close button for the first row; and up/down arrows, a close button, and a plus button for the second and third rows.

os artigos retornados correspondem a autores advindos de instituições brasileiras. Em relação ao retorno das buscas, foi retornado 3208 trabalhos da Scopus e 1227 da IEEE.

Cada autor do presente trabalho ficou responsável por anos específicos para as buscas nas bases apresentadas acima. Os artigos retornados das bases foram baixados no formato de arquivo *.csv*.

4.2. Pré-Processamento

Nesta fase iniciou-se a codificação do presente trabalho para as fases de pré-processamento e mineração dos dados, onde foi utilizada a linguagem *Python* para implementação por meio do *Jupyter Notebook*. O projeto pode ser encontrado no repositório do Github². Inicialmente, foram importadas as seguintes bibliotecas:

Figure 5. Bibliotecas.

```
import pandas as pd
from wordcloud import WordCloud
import matplotlib.pyplot as plt
from nltk import tokenize
import nltk
import seaborn as sns
from nltk import word_tokenize
from string import punctuation
from sklearn.preprocessing import StandardScaler
from sklearn.cluster import KMeans
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.neighbors import NearestNeighbors
import re
import time
from datasketch import MinHash, MinHashLSHForest
from string import punctuation
```

Para execução desta fase, foram aplicados os Passos do 1 ao 4 apresentados na Seção 2.1. Em **Limpeza dos Dados** a leitura dos dados de todos os arquivos *.csv* exportados está sendo realizada pelo comando `pd.read_csv()`. Nesta fase foram removidos os ruídos, dados inconsistentes e irrelevantes. As colunas disponibilizadas pela IEEE e Scopus foram verificadas a fim de excluir aquelas que não são necessárias para o presente estudo e selecionar as que são relevantes e comuns entre as duas bases. O arquivo *.csv* da Scopus retorna 33 colunas, e o da IEEE 30 colunas. Destas colunas foram selecionadas 9 colunas em comum para compor os *datasets* criados. Visto que alguns títulos das colunas eram diferentes entre Scopus e IEEE, mesmo apresentando o mesmo significado, estes foram renomeadas para posterior integração dos resultados de ambas as bases científicas em um único *dataframe*. Por exemplo, na IEEE o título *Doc-*

²<https://github.com/pamellasds/trabalho-final-mineracao>

ument Title foi renomeado para Title, assim como está na Scopus, por meio do comando: `df.rename(columns='Document Title': 'Title')`.

Com a remoção das colunas irrelevantes para o estudo, os atributos que irão compor os *datasets* inicial serão:

```
[ 'Authors' | 'Title', 'Year' | 'DOI' | 'Link', 'Authors with
affiliations', 'Abstract' | 'Author Keywords' | 'Publisher' ]
```

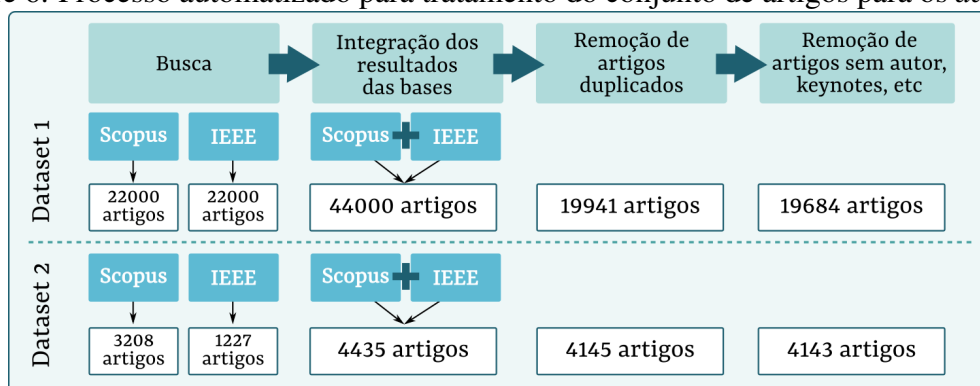
Na fase de **Integração dos Dados** todas as informações das diferentes bases de dados foram integradas em um único *dataset* por meio do comando `pd.concat()`, após os atributos terem sido padronizados, formando um *dataframe*. Haja vista que existe a possibilidade de duas bases diferentes retornarem artigos em comum, foi aplicada a remoção das duplicatas com o comando: `df.drop_duplicates(subset='Title', keep='first')`.

A **Transformação dos Dados** ocorre em dois processos: (i) transformação de todas as “Author Keywords” em palavras minúsculas, e (ii) remoção das pontuações. Na fase de **Seleção dos Dados**, o presente trabalho visa selecionar os dados para o processo de mineração a fim de extrair o conhecimento em três critérios:

1. Artigos em que o primeiro autor seja de instituição brasileira;
2. Artigos em que pelo menos um autor seja de instituição brasileira;
3. Todos os artigos do *dataset*.

A fim de atender cada um desses critérios, os devidos procedimentos de seleção foram codificados. Para o primeiro critério foi adicionada uma coluna denominada ‘First Author Country’, a partir da coluna ‘Authors and Affiliations’, de onde foram filtrados todos os artigos em que o primeiro autor era de uma instituição brasileira. Em relação ao segundo critério, o atributo ‘there are Brazilian authors?’ foi adicionada ao *dataframe*. Se tiver pelo menos uma ocorrência em que o autor é de uma instituição brasileira, o valor é “sim”, caso contrário, “não”. Por fim, para o último critério foram utilizados todos os artigos coletados.

Figure 6. Processo automatizado para tratamento do conjunto de artigos para os *datasets*.



4.3. Mineração de Dados

A etapa de mineração dos dados foi realizada em várias partes: (i) utilização de PLN sobre o atributo Keywords de cada dataset para observar quais temas os pesquisadores

brasileiros mais tem publicado e os temas que mais são publicados no âmbito internacional, e (ii) sistema de recomendação utilizando LSH sobre o atributo `Title` para apresentar artigos referentes às palavras-chave mais utilizadas pelos autores brasileiros detectadas na primeira parte. (iii) aplicação do algoritmo para encontrar regras de associação nas `Keywords`. (iv) utilização de PLN sobre o atributo `Authors`, construção de um grafo para representar a conexão entre autores de artigos e aplicação de algoritmos para identificar grupos e comunidades de autores.

As subseções a seguir detalha cada etapa de mineração de dados.

4.3.1. Pré-Processamento

Inicialmente foi realizado o pré-processamento utilizando o PLN. O primeiro pré-processamento foi retirar as duplicadas que poderia existir entre as duas base de dados escolhidas, aplicando sobre o atributo `Title`. Depois, foi convertido todas as palavras dos atributos `Keywords` e `Authors` para letra minúsculas. Por fim, foi retirado todas as pontuações contidas nas `Keywords`. Esse pré-processamento foi necessário para evitar que a mesma palavra seja contada como diferente e também, no caso das duplicadas, evitar que a mesma informação seja contada duas vezes ocasionando em resultados errôneos.

4.3.2. Temas mais usados em Mineração de Dados

Esta etapa foi dividida em três filtros para realizar a mineração de dados.

Inicialmente foi filtrado os artigo na qual o primeiro era de instituição brasileira. Para isto, foi utilizado o PLN no atributo `Authors with affiliations`, pois este atributo possui os países das instituições do autores. Foi gerado uma nova coluna, `First Author Country`, com o país do primeiro autor. Depois, foi processado palavra por palavra com o PLN para gerar uma lista de palavras, sendo palavras simples ou compostas, nos artigos que o atributo `First Author Country` possuía a palavra “Brasil” ou “Brasil”. Através da função `CountFrequency` foi verificado a frequência das palavras nesta lista e ordenadas em ordem decrescente. Assim, foi selecionado as onze primeiras palavras, pois a primeira palavra de cada lista era justamente o “data mining” que é a condicional do artigo ser selecionado para nossa base de dados.

Após essa primeira filtragem, foi filtrado os artigos na qual deveria existir pelo menos um autor de instituição brasileira, seja primeiro autor ou não. O processo foi semelhante ao descrito na abordagem com o primeiro autor do artigo ser de instituição brasileira. O diferencial foi a geração da coluna `There Are Brazilian Authors?`, que assinala “sim” para os países cuja a análise detectou a palavra “Brasil” ou “Brazil”.

Por fim, nesta etapa, foi feito a mineração de dados em todos os artigos contidos no nosso banco de dados.

4.3.3. Regras de Associação nas Keywords

Para tentar identificar regra de associação dentro do campo `Keywords` de cada dataset, foi criado um dataset auxiliar contendo uma entrada para cada conjunto de palavras-chave usada em cada artigo. A partir desse dataset auxiliar foi aplicado o algoritmo *frequent pattern growth* para tentar encontrar as regras de associação, os parâmetros dessa etapa vão ser discutidos na seção de resultados.

4.3.4. Sistema de Recomendação de Artigos

Nesta etapa, nós construímos um mecanismo de recomendação com Hash Sensível à Localidade (LSH do inglês *Locality-Sensitive Hashing*) em Python. Para isso, foram utilizados os unigramas extraídos do título do artigo.

Inicialmente, os títulos dos artigos foram convertidos em *tokens* (ou *shingles*). Por exemplo, o título “Experimental training in engineering” após o processo de “tokenização” fica [‘experimental’, ‘training’, ‘in’, ‘engineering’], assim como para todos os artigos da base. Em relação à configuração dos parâmetros, o número padrão de permutações foi equivalente à 128 e o número de artigos recomendados igual a 10.

Depois, aplicamos o MinHash e LSH ao conjunto, que o mapeia para um *hash*. O objetivo do MinHash é substituir um grande conjunto por uma “assinatura” menor que ainda preserva a métrica de similaridade subjacente. O LSH é usado para realizar pesquisas por vizinhos mais próximos com base em um conceito simples de “similaridade”. Pode-se dizer que dois itens são semelhantes se a interseção de seus conjuntos for suficientemente grande.

Assim, na aplicação do MinHash e LSH ao conjunto, os parâmetros de permutação e número de artigos recomendados foram usados como entradas da função. Cada título foi *tokenizado*, calculado o seu respectivo MinHash e armazenado para a construção da *Forest* de todos os MinHashs das strings. Além disso, a *Forest* foi indexada para se tornar pesquisável na próxima etapa. A partir disso, foi realizada uma pesquisa de similaridade entre o item de consulta, isto é, as dez primeiras palavras-chave dos temas mais pesquisados com primeiro autor de instituição brasileira, com pelos menos um autor de instituição brasileira e todos os artigos do *dataset* criado, e os outros itens no *hash* que foram encontrados na execução dos processos mencionados na Seção 4.3.2.

4.3.5. Mineração de Comunidades de Autores

Nesta etapa, o campo `Authors` (já pré-processado) de cada entrada de cada dataset foi copiado e transformado em um novo dataset, onde cada entrada desse dataset representa a relação de dois autores. Para realizar isso, os autores de cada artigo eram extraídos e eram feitas as combinações dois-a-dois dos autores e cada combinação se tornava uma nova entrada do novo dataset auxiliar. Os dados redundantes desse dataset foram então removidos (relações do tipo A-B/B-A) e então a partir desse dataset foi montado um grafo, onde cada vértice representa um autor e as arestas indicam se aqueles autores já

escreveram algum artigo juntos. Então foi aplicado o algoritmo de Girvan–Newman para encontrar comunidades no grafo.

5. Resultados e Discussão

Os resultados sobre a execução dos códigos serão apresentados nesta seção. Os resultados serão sumarizados a seguir, partir da aplicação de NPL mencionada anteriormente, e o uso de LHS para recomendação de artigos. As análises estão sendo realizadas para os dois *datasets* diferentes (Dataset 1, Dataset 2 mencionados na Seção 4).

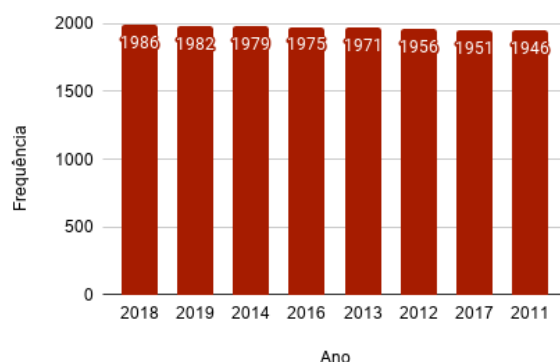
5.1. Sumarização dos Resultados

Esta análise é realizada a fim de descobrir a frequência dos temas que estão sendo mais abordados entre os pesquisadores brasileiros e, de uma maneira geral, considerando também em termos internacionais.

5.1.1. Dataset 1

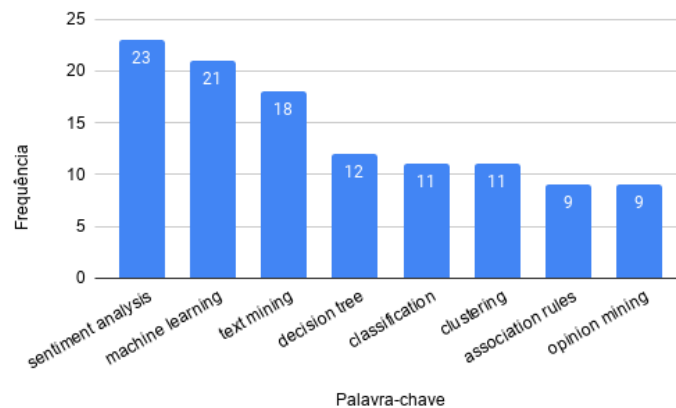
Neste *dataset* estão contidos os resultados de uma busca geral contendo os 2000 primeiros artigos de cada base para cada ano (2010-2020). A Figura 7 apresenta a frequência dos artigos publicados considerando a amostra por ano. Ressalta-se que para cada ano foram retornados os 2000 primeiros artigos, por isso que no gráfico não passa de 2000 por ano. Levando em conta essa amostragem por ano, percebe-se graficamente que as publicações ocorrem com uma certa constância.

Figure 7. Frequência da amostra de artigos entre os anos 2010 e 2020.



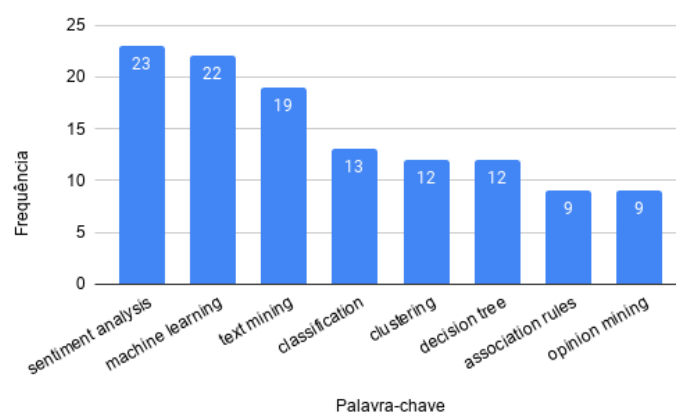
O gráfico apresentado na Figura 8 mostra a frequência dos oito primeiros temas mais pesquisados em relação aos artigos em que o primeiro autor é de alguma instituição brasileira. Assuntos relacionados à *Sentiment Analysis* (23 artigos) em geral é o tema que vem sendo mais estudado em pesquisas brasileiras considerando a amostra de resultados em que os 2000 artigos em cada ano podem tanto conter artigos de brasileiros, como de estrangeiros. Portanto, naturalmente esse número realmente pode ser baixo, considerando que não foi realizada a coleta de toda população. Em seguida tem-se os respectivos temas: *machine learning* (21), *text mining* (18), *decision tree* (12), *classification* (11), *clustering* (11), *association rules* (9), *opinion mining* (9).

Figure 8. Oito primeiros temas de artigos em que o primeiro autor é de instituição brasileira.



Por sua vez, o gráfico da Figura 9 mostra a frequência dos oito primeiros temas mais pesquisados em relação aos artigos em que pelo menos 1 autor seja de alguma instituição brasileira. Similarmente, assuntos relacionados à *Sentiment Analysis* (23 artigos) em geral é o tema que vem sendo mais estudado em pesquisas brasileiras. O que pode-se constatar que todos os artigos que utilizaram essa palavra-chave tem o primeiro autor de alguma instituição brasileira. Em seguida tem-se os respectivos temas: *machine learning* (22), *text mining* (19), *classification* (13), *clustering* (12), *decision tree* (12), *association rules* (9), *opinion mining* (9). Em “*Machine Learning*”, por exemplo, dos 22 artigos que tinham pelo menos um autor brasileiro, 21 tem como primeiro autor um pesquisador brasileiro.

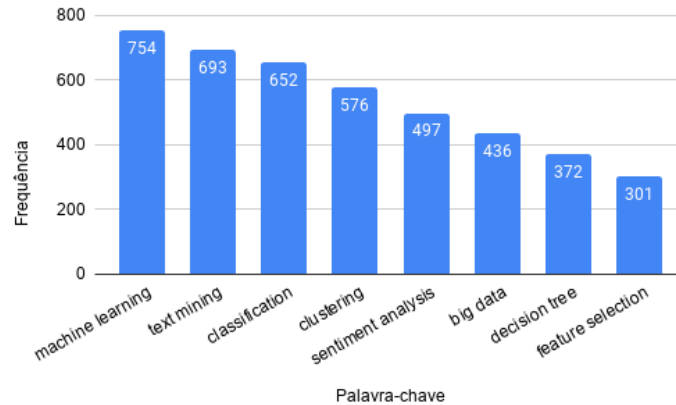
Figure 9. Oito primeiros temas de artigos em que pelo menos 1 autor é de instituição brasileira.



Por fim, o gráfico da Figura 10 mostra a frequência dos oito primeiros temas mais pesquisados considerando todos os autores, tanto os de instituição brasileira como artigos que contenham apenas autores estrangeiros. Assim, temas relacionados à *Machine Learning* (754 artigos) em geral é o tema que vem sendo mais estudado em pesquisas no geral. Em seguida tem-se os respectivos temas: *text mining* (693), *classification* (652), *clustering* (576), *sentiment analysis* (497), *big data* (436), *decision tree* (372) e *feature*

selection (301).

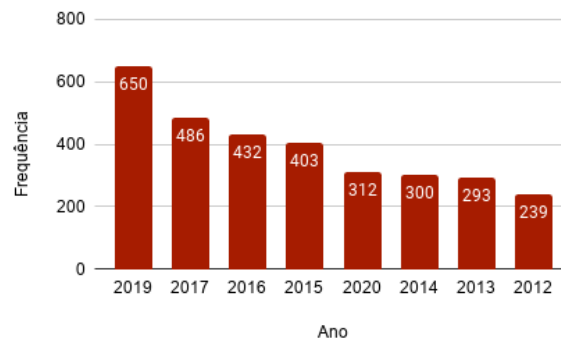
Figure 10. Oito primeiros temas de todos os artigos retornados da busca.



5.1.2. Dataset 2

Neste *dataset* estão contidos os resultados de uma busca pelos artigos que contém “data mining” nas *keywords* e “brasil”, “brazil” ou “brazilian” nas *affiliations*. Tal estratégia foi utilizada para aumentar a amostra de artigos brasileiros dos anos de 2010 a 2020, visto que para a versão deste trabalho não houve automatização para coleta de toda a população. A distribuição por ano, considerando esta restrição, está sendo mostrada na Figura 11. Observa-se que o ano de 2019 foi o ano que tiveram mais artigos publicados comparados aos anteriores.

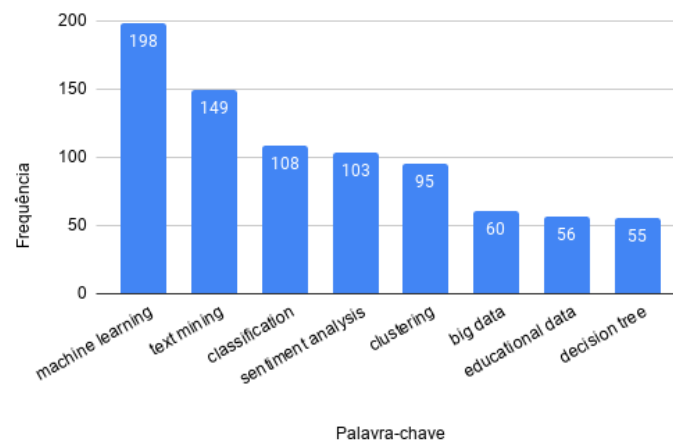
Figure 11. Frequência da amostra de artigos brasileiros entre os anos 2010 e 2020.



O gráfico apresentado na Figura 12 mostra a frequência dos oito primeiros temas mais pesquisados em relação aos artigos em que o primeiro autor é de alguma instituição brasileira. Assuntos relacionados à *Machine Learning* (198 artigos) em geral é o tema que vem sendo mais estudado em pesquisas brasileiras. Em seguida tem-se os respectivos temas: *text mining* (149), *classification* (108), *sentiment analysis* (103), *clustering* (95), *big data* (60), *educational data mining* (56), *decision tree* (55).

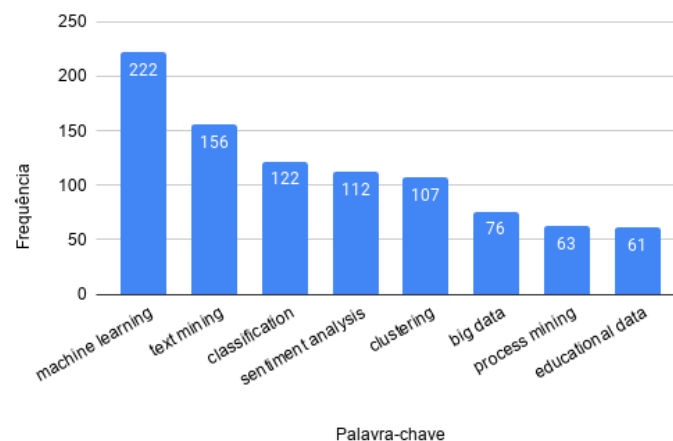
Por sua vez, o gráfico da Figura 13 mostra a frequência dos oito primeiros temas mais pesquisados em relação aos artigos em que pelo menos 1 autor seja de alguma

Figure 12. Oito primeiros temas de artigos em que o primeiro autor é de instituição brasileira.



instituição brasileira. Assuntos relacionados à *Machine Learning* (222 artigos) em geral é o tema que vem sendo mais estudado em pesquisas brasileiras. Em seguida tem-se os respectivos temas: *text mining* (156), *classification* (122), *sentiment analysis* (112), *clustering* (107), *big data* (76), *process mining* (63) *educational data mining* (61).

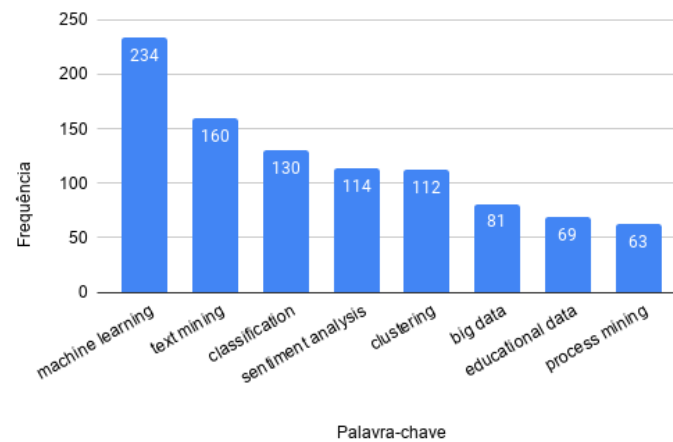
Figure 13. Oito primeiros temas de artigos em que pelo menos 1 autor é de instituição brasileira.



Por fim, a frequência dos oito primeiros temas mais pesquisados considerando todas as buscas retornadas, nesse caso, podendo também conter artigos que tenham apenas autores estrangeiros. Observa-se, na Figura 14, que assuntos relacionados à *Machine Learning* (234 artigos) em geral continua sendo o tema que mais estudado em pesquisas brasileiras. Em seguida tem-se os respectivos temas: *text mining* (160), *classification* (130), *sentiment analysis* (114), *clustering* (112), *big data* (81), *educational data mining* (63), *process mining* (69).

O *Dataset 2* está mais preciso em relação ao *Dataset 1* considerando os trabalhos de autoria brasileira, visto que aumentou-se a quantidade da amostra. Ou seja, contém

Figure 14. Oito primeiros temas de todos os artigos retornados da busca.



mais trabalhos brasileiros na análise. Apesar disso, o comportamento deste *dataset* segue parecido com o *Dataset 1* quanto às palavras-chave mais utilizadas.

5.2. Cálculo de Regras de Associação

Após analisarmos as frequências de ocorrência das palavras chaves, foram aplicados os algoritmos FP Growth e Apriori com o intuito de identificar relações entre as palavras-chaves definidas pelos autores. Inicialmente para encontrar a frequência dos termos e aplicar os algoritmos foi usado um suporte mínimo de 20%, porém ambos os *datasets* apenas o termo '*data mining*' foi reportado, o que já era esperado por ser o termo central dos *datasets*.

Ao diminuir o suporte para 3%, os termos *classification*, *text mining*, *clustering*, *sentiment analysis* e *machine learning* também foram identificados no *dataset 1* e os termos *big data*, *classification*, *clustering*, *data mining*, *machine learning*, *sentiment analysis* e *text mining* foram identificados para o *dataset 2*. Com o intuito de investigar esse comportamento, aplicamos os algoritmos mencionados. Apenas regras de associação “triviais” foram encontradas, como (*data mining*) - (*classification*) ou (*clustering*) - (*data mining*), além disso todas com valores de confiança baixos. Esses resultados foram, a princípio, inesperados então os valores de suporte foram diminuídos ainda mais para 0.01% apenas com o intuito de investigar esse comportamento. Com tais alterações outras regras surgiram, como (*social network*) - (*data mining*, *quality of life*, *facebook*) ou (*educational data mining*) - (*academic performance*), mas novamente com métricas muito baixas para serem consideradas relevantes.

Para investigar mais fundo esse comportamento, foi extraída a lista de palavras únicas que aparecem nas palavras-chave definidas pelos autores e uma análise manual foi feita. A partir dessa análise foi teorizado que a variedade de palavras e a forma como os autores reportam elas podem estar prejudicando o uso dos algoritmos. Foram encontrados casos em que os autores colocaram frases completas como uma única palavra-chave, ou que colocaram palavras ou siglas específicas do domínio de aplicação deles. Também foi possível observar repetidos casos de palavras citadas apenas uma vez em cada conjunto de dados.

Para testar essa hipótese o mesmo procedimento para calcular as regras de associação foi realizado apenas com as 20, 50, 100 e 200 palavras-chaves mais frequentes, todas as palavras-chaves que não faziam parte do conjunto foram removidas. Nestes testes foi possível observar o surgimento de novas regras de associação, como (*data reduction*) - (*big data, instance selection*) ou um aumento na confiança das relações previamente encontradas, porém novamente os valores de suporte tiveram que ser colocados em valores abaixo de 5%. Ao continuar a analisar o conjunto de palavras-chaves filtrados, foi observado que após a filtragem uma porção majoritária dos artigos ficava com apenas uma ou duas palavras chaves, o que poderia prejudicar a construção das regras. Apesar dos testes, a causa desse comportamento ainda não foi identificada.

5.3. Recomendação de Artigos Usando LHS

A partir das palavras-chaves mais pesquisadas resultantes do uso de NPL e mineração de texto, foi possível utilizá-las na recomendação de artigos implementados com LSH.

Por exemplo, na Tabela 1, mostra apenas alguns dos resultados que a execução do LSH retornou. O algoritmo teve como entrada o *Dataset 1*, onde os dados armazenados advêm da busca geral, através da qual as bases científicas retornaram os 2000 primeiros artigos para cada ano, podendo conter ou não artigos brasileiros. Como são muitos artigos e este trabalho pretende apenas apresentar como o algoritmo comportou-se, foi escolhido apenas dois artigos do total retornado para cada palavra-chave pesquisada, sendo estas as oito mais utilizadas.

Table 1. Amostra de Artigos Recomendados para cada Palavra-chave

Palavra-chave	Artigo Recomendado	Autores
machine learning	Machine Learning	Grosan, C., Department of Computer Science, Babes-Bolyai University, Kogalniceanu 1, 400084 Cluj - Napoca, Romania; Abraham, A., Machine Intelligence Research Labs (MIR Labs), Scientific Network for Innovation and Research Excellence, 2259, 98071-2259 Auburn, WA, United States
	Machine learning with templates	Fiske, M.S., Aemea Institute, San Francisco, CA, United States
text mining	Social innovation activities in Japanese firms: A pilot study with text mining	Zhao, W.L., Economic Research Center, Fujitsu Research Institute, Tokyo, Japan; Ouchi, N., Department of Industrial and Systems Engineering, Aoyama Gakuin University, Tokyo, Japan; Watanabe, C., Department of Mathematical Information Technology, University of Jyväskylä, Jyväskylä, Finland
	Lexico-syntactic causal pattern text mining	Joshi, S., Applied Artificial Intelligence Laboratory, University of Cincinnati, Cincinnati, OH 45221, United States; Pangaonkar, M., Applied Artificial Intelligence Laboratory, University of Cincinnati, Cincinnati, OH 45221, United States; Seethakkagari, S., Applied Artificial Intelligence Laboratory, University of Cincinnati, Cincinnati, OH 45221, United States; Mazlack, L.J., Applied Artificial Intelligence Laboratory, University of Cincinnati, Cincinnati, OH 45221, United States
classification	Discriminatory decision policy aware classification	Mancuhan, K., Department of Computer Science, Purdue University, United States; Clifton, C., Department of Computer Science, Purdue University, United States
	Functional classification of websites	Gali, N., University of Eastern Finland, P.O. Box 111, Finland; Istodor, R.M., University of Eastern Finland, P.O. Box 111, Finland; Fränti, P., University of Eastern Finland, P.O. Box 111, Finland
clustering	Density-based clustering	Kriegel, H.-P., Ludwig-Maximilians-Universität München, Munich, Germany; Kröger, P., Ludwig-Maximilians-Universität München, Munich, Germany; Sander, J., University of Alberta, Edmonton, AB, Canada; Zimek, A., Ludwig-Maximilians-Universität München, Munich, Germany
	Clustering by shift	Chehreghani, M.H., NAVER LABS Europe, Xerox Research Centre Europe - XRCE, France
sentiment analysis	Automatic sentiment analysis of user reviews	Abinaya, R., Department of Computer Science and Engineering, Easwari Engineering College, Chennai, India; Aishwaryaa, P., Department of Computer Science and Engineering, Easwari Engineering College, Chennai, India; Baavana, S., Department of Computer Science and Engineering, Easwari Engineering College, Chennai, India; Selvi, N.D.T., Department of Computer Science and Engineering, Easwari Engineering College, Chennai, India
	Sentiment analysis	Das, A., National Institute of Technology Calicut, India
big data	Privacy-preserving representation learning for big data	Zhu, X., Xiaofeng Zhu is with Guangxi Normal University, Guilin, China; Shang, S., Shuo Shang is with University of Electronic Science and Technology of China, Chengdu, China; Kim, M., Minjeong Kim is with the University of North Carolina at GreensboroNC, United States
	Big data analytics for seismic fracture identification using amplitude-based statistics	Udegbe, E., Department of Energy and Mineral Engineering, The Pennsylvania State University, 110 Hosler Building, University Park, State College, PA 16802-5000, United States; Morgan, E., Department of Energy and Mineral Engineering, The Pennsylvania State University, 110 Hosler Building, University Park, State College, PA 16802-5000, United States; Srinivasan, S., Department of Energy and Mineral Engineering, The Pennsylvania State University, 110 Hosler Building, University Park
decision tree	A bottom-up oblique decision tree induction algorithm	Barros, R.C., Department of Computer Science, ICMC, University of São Paulo (USP), São Carlos - SP, Brazil; Cerri, R., Department of Computer Science, ICMC, University of São Paulo (USP), São Carlos - SP, Brazil; Jaskowiak, P.A., Department of Computer Science, ICMC, University of São Paulo (USP), São Carlos - SP, Brazil; De Carvalho, A.C.P.L.F., Department of Computer Science, ICMC, University of São Paulo (USP), São Carlos - SP, Brazil
	Quantum decision tree classifier	Lu, S., School of Computer Science and Technology, Huazhong University of Science and Technology, Wuhan 430074, China; Braunstein, S.L., Department of Computer Science, University of York, York YO10 5GH, United Kingdom
feature selection	Predictability score-based feature selection	Crăciun, M.V., Department of Computer and Information Technology, Dunarea de Jos University of Galati, 800323, Romania; Cocu, A., Department of Computer and Information Technology, Dunarea de Jos University of Galati, 800323, Romania; Dumitriu, L., Department of Computer and Information Technology, Dunarea de Jos University of Galati, 800323, Romania
	Robust ensemble feature selection for high dimensional data sets	Ben Brahim, A., LARODEC, ISGT, University of Tunis, Tunisia; Limam, M., LARODEC, ISGT, University of Tunis, Tunisia; Dhofar University, Oman

Para o mesmo *Dataset 1*, o algoritmo foi executado, dessa vez, considerando os artigos produzidos por pesquisadores de universidades brasileiras, como consta na Tabela 2. Similarmente, a palavra-chave “*Sentiment Analysis*” retorna artigos relacionados ao tema em questão, como pode-se conferir nos trabalhos: “*Anonymous real-time analytics monitoring solution for decision making supported by sentiment analysis*” e “*Anonymous real-time analytics monitoring solution for decision making supported by sentiment analysis*”. Pode-se observar também que todos os primeiros autores dos artigos são de alguma instituição brasileira.

Table 2. Amostra de Artigos brasileiros recomendados para cada Palavra-chave

Palavra-chave	Artigo Recomendado	Primeiro Autor e Afiliação
sentiment analysis	Anonymous real-time analytics monitoring solution for decision making supported by sentiment analysis	de Oliveira, G.A., Jr., Cyber Security INCT Unit 6, Laboratory for Decision-Making Technologies (LATITUDE), Department of Electrical Engineering (ENE), Faculty of Technology, University of Brasília (UnB), Brasília, DF 70910-900, Brazil
	Opinion-meter: A framework for aspect-based sentiment analysis	Farias, D.S., Inst. de Ciências Matemáticas e de Computação, USP, Universidade Federal do Mato Grosso do Sul, Brazil
machine learning	Ranking machine learning classifiers using multicriteria approach	De Moura Rezende Dos Santos, F., Computer Science, Universidade de Brasília, Brazil
	Cryptographic Algorithm Identification Using Machine Learning and Massive Processing	De Mello, F.L., Polytechnic School, Federal University of Rio de Janeiro (UFRJ), Rio de Janeiro, Brazil
text mining	Neutron activation analysis and data mining techniques to discriminate between beef cattle diets	Tejeda Mazola, Y., Nuclear Energy Center for Agriculture, University of São Paulo
	Graph Pattern Mining and Learning through User-Defined Relations	Teixeira, C.H.C., Department of Computer Science, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil;
decision tree	A bottom-up oblique decision tree induction algorithm	Barros, R.C., Department of Computer Science, ICMC, University of São Paulo (USP), São Carlos - SP, Brazil;
	A statistical decision tree algorithm for data stream classification	Cazzolato, M.T., Computer Science Department, Federal University of São Carlos, São Carlos, Brazil;
classification	Classification, Association and Clustering of Water Body Data: Application to Water Quality Monitoring	Bertholdo, L., Federal Institute of Education, Science and Technology of São Paulo, São Paulo, Brazil, Faculty of Technology, University of Campinas, Limeira, Brazil;
	Making data stream classification tree-based ensembles lighter	Costa, V.G.T.D., Computer Science Department, Londrina State University, Londrina, Brazil
clustering	Pattern clustering using ants colony, Ward Method and Kohonen Maps	Villwock, R., University of West of Paraná, Universitária Street, 2069, Cascavel, Brazil
	Graph-based Clustering of miRNA Sequences	Kasahara, V.A., Computer Science Department, Federal University of São Carlos, S. Carlos, SP, Brazil
association rules	Identification of freight patterns via association rules: The case of agricultural grains	Moreira, C.E.S., University of Campinas (UNICAMP), School of Electrical and Computer Engineering, Campinas, SP 13083- 852, Brazil
	Comparative study of algorithms for mining association rules: Traditional approach versus multi-relational approach	Valêncio, C.R., Depto. de Ciências de Computação e Estatística, Universidade Estadual Paulista - Unesp, São José do Rio Preto, Brazil
opinion mining	Neutron activation analysis and data mining techniques to discriminate between beef cattle diets	Tejeda Mazola, Y., Nuclear Energy Center for Agriculture, University of São Paulo, Avenida Centenário 303, Piracicaba, SP 13416-000, Brazil
	Graph Pattern Mining and Learning through User-Defined Relations	Teixeira, C.H.C., Department of Computer Science, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil

Ressalta-se que a execução do algoritmo implementado retorna os 10 primeiros artigos recomendados, número que pode ser facilmente modificado no código implementado caso seja necessário retornar mais artigos para cada palavra-chave em questão.

5.4. Análise de Comunidades de Autores

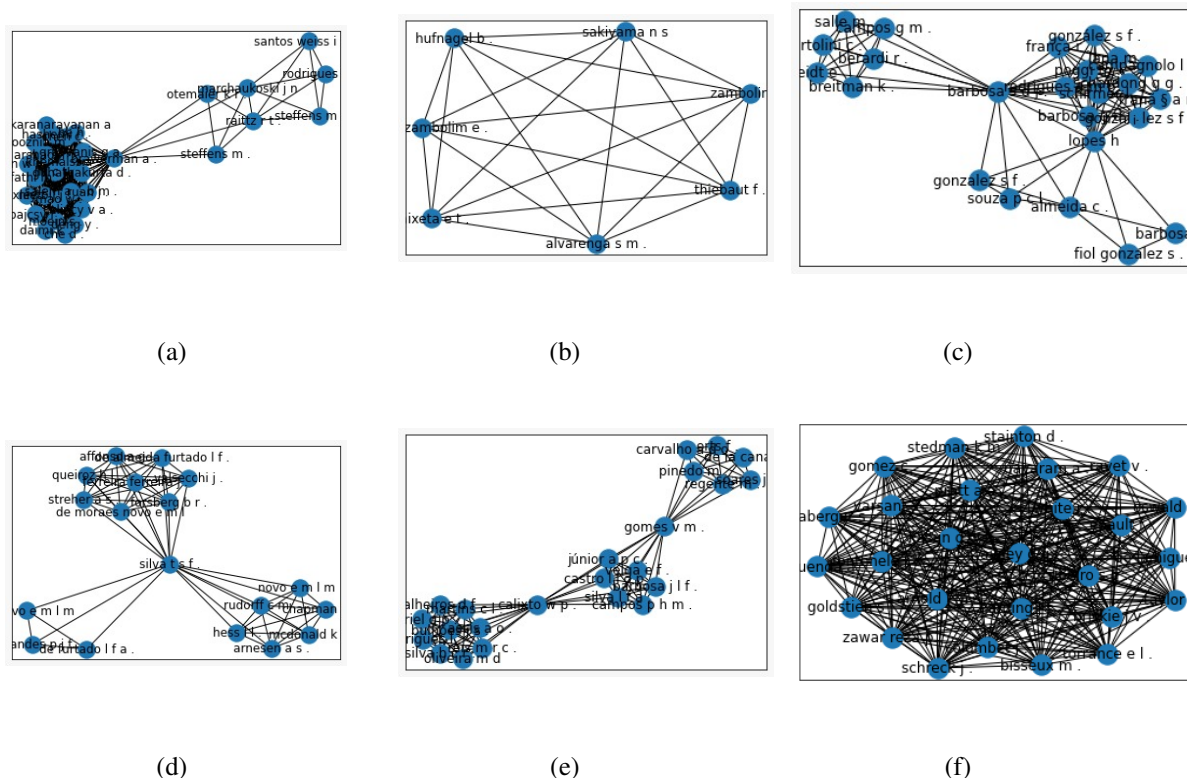
A partir do PLN e dos datasets auxiliares, mencionados na seção 4.3.5, foi possível montar um grafo para cada conjunto de dados inicial, no caso do Dataset 1 o grafo resultante tinha 46212 vértices(autores) e 205211 arestas(relações) e o grafo do dataset 2 tinha 10670 vértices e 46789 arestas. Para identificar as comunidades, o algoritmo de Girvan–Newman foi aplicado em cada grafo, resultando em 6795 grupos, para o dataset 1 e 1101, para o dataset 2. Devido o tamanho dos grafos originais optamos por cortar e extrair o sub-grafo de cada grupo e fazer uma análise mais pontual desses grupos. Ao analisar as comunidades de cada grafo é possível observar comportamentos similares nos dois conjuntos. Em ambos os grafos, uma porção grande dos vértices, 40.04% para o dataset 1 e 39.12% para o dataset 2, fazem parte de uma comunidade só. Os outros vértices foram divididos em grupos menores.

Como mostrado na Figura 15, podemos observar a presença de grafos completos 16b, onde todos os autores estavam conectados diretamente, esse comportamento foi observado nos dois datasets e em diferentes tamanhos, em alguns casos os autores de um grupo de um sub-grafo completo eram todos de um mesmo artigo, porém em outros, como mostrado na figura 16f, o grupo representava mais de um artigo. No total, 85.81% dos grupos do dataset 1 eram grafos completos, e 81.6% para o dataset 2.

Porém, apesar da maioria dos grupos em ambos datasets serem sub-grafos completos, também tiveram ocorrências de outros formatos. Foi possível observar casos como o das figuras 16d e como no caso 16a onde um único autor se conecta de forma exclusiva

com dois ou mais grupo fechados, provavelmente esse tipo de grupo representa algum tipo de supervisor ou orientador que trabalha com vários times. Instancias como o caso 16e, também foram observadas nos dois conjuntos de dados. Por último, instancias difusas também foram encontradas, como no caso 16c, onde os autores apesar de serem conectados fortemente com um grupo, ainda possuem conexões com outros.

Figure 15. Comunidades de Autores



6. Considerações Finais

Realizou-se uma preliminar revisão bibliográfica semi-automática utilizando mineração de dados. O processo ocorre desde a busca dos artigos nas bases de origem como Scopus e IEEE até a sumarização das informações encontradas através das etapas de mineração de dados. Primeiramente, foi utilizada PLN para verificar os temas mais utilizados pela comunidade brasileira e internacional, e Regras de Associação utilizando as palavras-chaves dos artigos. Além disso, LSH foi utilizado para recomendação de artigos considerando as palavras-chaves que são utilizadas com mais frequência e, por fim, utilização do algoritmo de Girvan-Newman para a geração de comunidades de autores.

Como trabalhos futuros, pretende-se incluir mais bases que contenham as informações necessárias. Além disso, para abranger o maior número possível de artigos, planeja-se implementação de um *Crawler* para realização de uma busca automática nas bases científicas de origem. Por fim, o desenvolvimento de uma interface faz-se necessária para que o pesquisador possa introduzir buscas relacionadas à outros domínios e tópicos a fim de obter informações bibliométricas do estado da arte.

References

- Bezerra, C. A. and Guimarães, A. J. R. (2014). Mineração de texto aplicada às publicações científicas sobre gestão do conhecimento no período de 2003 a 2012. *Perspectivas em Ciência da Informação*, 19(2):131–146.
- Camilo, C. O. and Silva, J. C. d. (2009). Mineração de dados: Conceitos, tarefas, métodos e ferramentas. *Universidade Federal de Goiás (UFG)*, pages 1–29.
- de Amo, S. (2004). Técnicas de mineração de dados. *Jornada de Atualização em Informática*.
- de Araujo, R. M., Athayde Silveira, B., Yusuke Muramatsu, T., and Revoredo, K. (2015). Minerando publicações científicas para análise da colaboração em comunidades de pesquisa-o caso da comunidade de sistemas de informação. *Revista Eletrônica de Sistemas de Informação*, 14(1).
- Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI magazine*, 17(3):37–37.
- Goldschmidt, R. and Passos, E. (2005). *Data mining: um guia prático*. Gulf Professional Publishing.
- Han, J., Pei, J., and Kamber, M. (2011). *Data mining: concepts and techniques*. Elsevier.
- Hand, D. J. and Adams, N. M. (2014). Data mining. *Wiley StatsRef: Statistics Reference Online*, pages 1–7.
- Hirji, K. K. (1999). Discovering data mining: From concept to implementation. *ACM SIGKDD Explorations Newsletter*, 1(1):44–45.
- Martino, R., Oliveira, J., and Souza, J. (2009). Mineração de dados científicos para prospecção tecnológica e identificação de especialistas. In *V Workshop em Algoritmos e Aplicações de Mineração de Dados*. Ceará.
- Mourão, A., Pasquini, R., Villaça, R., and Camargos, L. (2016). Busca por similaridade no cassandradb.
- Real, R. and Vargas, J. M. (1996). The probabilistic basis of jaccard’s index of similarity. *Systematic biology*, 45(3):380–385.