Uma Abordagem para Geração e Visualização de Regras de Associação de Acesso a Conteúdos de Portal de Notícias

Tayná Arruda Câmara da Silva
Programa de Pós-Graduação em Engenharia Elétrica e de Computação
Univ. Federal do RN
Natal, Brasil
arruda.camara.113@ufrn.edu.br

Prof. Dr. Luiz Affonso Guedes
Dep. de Eng. de Computação e Automação
Univ. Federal do RN
Natal, Brasil
affonso@dca.ufrn.br

***Abstract*—This paper aims to develop and visually association rules through a data mining project approach, using a data set of history navigation contents of an online news portal of a Brazilian magazine. The rules found allow us to understand users and subscribers access trends sequences through frequent items by applying the Apriori algorithm. To achieve this objective, the concepts of exploratory data analysis - EDA, scientific data visualization, knowledge discovery in databses - KDD and association rules were used. The result is a dataset and and visualization using parallel bar diagrams for the association rules found which bring together the main features of publications to frequently read together in sequence.**

***Index Terms*—Data Mining, Association Rules, Apriori, Visual data mining, history content navigation**

I. Introdução

Aproximadamente três décadas após o início da transposição do conteúdo dos jornais impressos para o formato digital, atingiu-se um fluxo de navegação de consumo dos conteúdos com característica puramente digital [1], onde o usuário tem a possibilidade de fluidez de navegação estruturada para esse formato. Dessa forma, existe para os jornais e revistas o desafio de fidelizar os usuários consumidores do seu conteúdo digital [2].

O relatório do Instituto *Reuters* para o ano de 2023 [2] mostra que 28% dos *publishers* já têm a inteligência artificial integrada em seus produtos como um meio de entregar uma experiência mais personalizada para os seus leitores e 39% vêm fazendo testes para introduzir essa área. Parte dos esforços para utilizar as tecnologias baseadas em dados está no sentido de fidelizar o público dos clubes de assinaturas de revistas e jornais, sendo citado por 80% dos entrevistados como uma das mais importantes prioridades [2].

Porém, dados gerados pelos diferentes tipos de serviços de mídias digitais *online* sobre os padrões de uso dos seus clientes demandam cada vez mais soluções eficientes para processá-los de modo a gerar informação útil ao negócio dessas empresas de mídia. Para se descobrir relacionamentos entre esses dados, utlilizam-se de técnicas de mineração de dados [3], as quais são essencialmente algoritmos de aprendizagem não supervisionada [4] [5]. Entre os algoritmos de aprendizagem não supervisionada mais básicos e populares, podemos citar o K-*Means*, para agrupamento de dados e o *Apriori*, para geração de regras de associação. Além do emprego de técnicas de aprendizagem de máquinas não supervisionadas, há a necessidade de se utilizar técnicas de análise expliratória de dados (*Exploratory Data Analysis* - EDA) e técnicas de visualização de dados, para melhor compreensão dos resultados obtidos [7].

Diante da relevância do tema, este trabalho tem como objetivo apresentar uma abordagem de descoberta de padrões de acesso de clientes de um portal de notícia *online*, de modo a melhor compreender seus hábitos e preferências de uso e consequente proposição de um sistema eficiente de recomendação de notícias visando a fidelização dos seus clientes. Essa abordagem possui quartro etapas: análise exploratória dos dados (EDA), pré-processamento dos dados, geração de regras de associação e visualização das regras. Na etapa de EDA são analisados os dados de modo a identificar as informações mais relevantes; na etapa de pré-processamento, os dados mais relevantes são preparados para permitir a geração das regras de associação via a utilização do algoritmo Apriori e na quarta etapa são utilizados diagramas de barras paralelas para melhor visualização dos resultados obtidos.

Este trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES). O trabalho está dividido em mais quatro seções. Na Seção II está o conceito de *Knowledge Discovery in Databases* - KDD e os conceitos de regras de associação, algoritmo Apriori e visualização científica de dados. Na Seção III está detalhada a abordagem proposta. Na seção IV são exibidos os resultados obtidos utilizando-se dados reais de acesso de clientes ao portal de notícias. Na Seção V estão as considerações finais.

II. *Knowledge Discovery in Databases* - KDD

Data Mining ou Mineração de Dados é uma área da computação que utiliza um conjunto de técnicas e algoritmos para descobrir relacionamentos entre dados e consequente suporte à tomada de decisão [10]. Aplica-se em diversas áreas, como: vendas, conteúdos digitais e saúde [11].

Por sua vez, *Knowledge Discovery in Databases* - (KDD) é o processo pelo qual técnicas de tratamento e mineração de dados são utilizadas a fim de se descobrir os relacionamentos relevantes entre os dados [3]. Esse processo se constitui basicamente em três grandes etapas: pré-processamento, mineração de dados e pós-processamento. A mineração de dados é uma etapa do processo de KDD que busca efetivamente por conhecimentos novos e úteis a partir dos dados [3].

Os padrões extraídos no processo de *KDD* podem ser preditivos ou descritivos. Os descritivos têm o objetivo de apresentar informações de interesse do especialista de negócio [3], os dados serão explorados sem hipótese pré-definida à procura de padrões de ocorrências frequentes, tendências e generalizações sobre os dados [9].

A. Regras de Associação

A aprendizagem não supervisionada é uma área da aprendizagem de máquinas que utiliza algoritmos para encontrar padrões ou estruturas em conjuntos de dados não rotulados. Técnicas de aprendizagem não supervisionada são empregadas na etapa de mineração de dados, especialmente com o objetivo de encontrar padrões descritivos [9], uma vez que nessa abordagem não se tem conhecimento *a priori* de uma saída esperada, ou seja, é realizada uma análise nos dados buscando-se relações entre eles sem que haja um alvo específico. Dentre as técnicas e algoritmos, destacam-se os voltados a agrupamento de dados e geração de regras de associação.

Os algoritmos de geração de regras de associação visam buscar relacionamentos relevantes entre itens, tal que a ocorrência de um item implica na ocorrência do outro. A análise irá determinar itens que ocorrem frequentemente juntos e os dados são representados em forma de transações [10] . O exemplo mais comum na literatura para o emprego de técnicas de geração de regras de associação é referente ao extrato de compra de supermercado, que consegue identificar itens que são comprados juntos nas transações e auxilia na distribuição deles nas prateleiras [4]. O objetivo ao aplicar um algoritmo de geração de regras de associação é explicitar as associações entre os dados, além de indicar as relevâncias dessas associações. Regras de associação estendem medidas estatísticas como correlação uma vez que podem mensurar níveis de relacionamentos entre duas ou mais variáveis.

As regras de associação são formadas por um ou mais antecessores (P) e um sucessor (Q), indicada por:

$$P \rightarrow Q$$

Se os atributos em P são verdadeiros, então o atributo em Q tende a ser verdadeiro também [11]. Os algoritmos de geração de regras de associação buscarão na base de dados relações entre os antecessores e os sucessores mais frequentes. Assim, quando determinados padrões de comportamento indicam associação entre itens com frequência, é entendido como uma regra de associação [9].

1) Heurísticas das Regras de Associação: Para avaliar a qualidade das associações que o algoritmo gerou devem ser utilizadas métricas de avaliação, elas irão orientar quão relevantes são as regras que se deseja gerar e serão os hiperparâmetros do algoritmo. Quanto mais restritas as regras, a tendência é que o subconjunto analisado seja menor. As três métricas principais para a definição das regras úteis, são: suporte, confiança e *lift*.

- Suporte: O suporte calcula a proporção da ocorrência de um ou mais itens na transação sobre o total de transações da base de dados. Consequentemente, quanto mais baixo o valor suporte, indica que um ou mais itens analisados aparecem poucas vezes no conjunto de dados. Assim, teremos os itens mais frequentes ordenando os resultados dos cálculos do suporte do maior para o menor. Escolher um valor de suporte mínimo muito baixo irá fazer com que o algoritmo teste um grande número de itens [10].
  Suporte é denotado formalmente como a divisão entre o número total de transações em que o item da regra está presente pelo número total de transações existentes na base de dados [4].
- Confiança: A confiança é composta por um item antecessor e um item sucessor e avalia a proporção da ocorrência dos itens sucessores e a proporção da premissa sobre o total proporcional de itens que contêm ambos sucessores e antecessores com ocorrência no antecessor. O valor de confiança será maior quando os itens estão mais associados, ou seja, aparecem com mais frequência juntos nas transações.
  Assim, para a regra:

  $$A \rightarrow B$$

  Confiança é definida formalmente [4] pela equação 1:

  $$Confidence = \left(\frac{support A \cup B}{support\,(A)} \right) \quad (1)$$

  No caso, o numerador é o suporte calculado quando A e B estão presentes na transação e o denominador se refere ao suporte de A.
- *Lift*: O *Lift* irá usar a informação da confiança e do suporte, de modo que fará a divisão da confiança do antecessor e sucessor dividido pelo suporte do sucessor. Assim, para a regra:

  $$A \rightarrow B$$

  Lift é dado por [4]:

  $$lift\,(A \rightarrow B) = \frac{support\,(A \cup B)}{support\,(A) * support\,(B)} \quad (2)$$

  Observa-se que o *lift* considera no numerador quando ambos A e B estão presentes na transação e para o denominador faz a multiplicação entre o suporte de A e o suporte de B.
  O *Lift* pode assumir os seguintes valores:
  - Igual a 1: quando o antecessor e o sucessor são independentes um do outro;
  - Maior que um: quando o antecessor e o sucessor são dependentes um do outro;

– Menor que um: quando o antecessor e o sucessor são substitutos um do outro

B. Algoritmo Apriori

O algoritmo Apriori é considerado um dos algoritmos mais clássicos [10] de geração de regras de associação e foi proposto por Agrawal e Shrikant em 1994 [4]. Esse algoritmo utiliza uma heurística de busca do tipo *bottom-up*, uma vez que busca os relacionamentos dos itens mais frequentes para os menos frequentes, sendo os níveis mínimos pré-definidos de suporte, confiança e *lift* os critérios de para do algoritmo.

O algoritmo pode trabalhar com um número grande de atributos e gerar diversas regras combinatórias entre eles [7]. Porém, como ele precisa analisar a base completa diversas vezes, seu desempenho computacional fica comprometido para base de dados muito grande [4]. Além disto, os valores mínimos de suporte, confiança e *lift* admitidos têm impacto direto no número de regras geradas e no tempo de busca na base de dados [9].

Neste trabalho, para geração das regras de associação utizamos o pacote *Apyori*, que é uma implementação do algoritmo *Apriori* em *Python*[1] [8].

C. Visualização Cientfica de Dados

As informações descobertas em um processo de aprendizagem precisam ser visualizadas de modo que tornem a sua compreensão e hipóteses possíveis e acessíveis. Para expor resultados de forma objetiva e clara, há necessidade de se usar recursos visuais adequados [13].

Os modelos de aprendizagem podem ser descritos de diferentes formas, porém, para o objetivo de gerar conhecimento precisamos considerar os fatores cognitivos dos conhecedores do negócio [16] de modo que não se crie algo que fique obsoleto. Na mineração de dados a visualização aumenta potencialmente a compreensibilidade e permite ações posteriores ao que foi encontrado com os modelos de mineração de dados [17].

A visualização de dados, entretanto, não auxilia apenas na compreensão dos resultados, uma vez que é um recurso importante em todas as etapas de diferentes metodologias de projetos de dados, desde o pré-processamento dos dados, análise exploratória, seleção de *features*, construção de modelos de aprendizagem, avaliação de modelos, entre outras [15].

Para os casos de aprendizagem não supervisionada e análises descritivas [15], a visualização possibilita a análise simplificada dos dados quando não se sabe exatamente quais questões precisam ser criadas no próximo passo de análise [14].

Quando concentramos em análises de regras de associação, podemos ter ocorrência de muitas regras geradas que precisam ser analisadas pelos especialistas de modo a validá-las ou não. A aplicação de técnicas de visualização de dados para conjuntos de itens frequentes e associação regras possibilita que as pessoas envolvidas na análise dos resultados possam perceber com mais facilidade as associações encontradas [18]. Aqui neste trabalho, propomos a utilização de diagramas de barras paralelas (ou gráfico de coordenadas pararlelas) para visualização das regras de associação [19]. Gráficos de coordenadas paralelas foram concebidos para visualizar e analisar conjuntos de dados de alta dimensão, onde cada barra paralela corresponde a uma variável (item, *feature*) e linhas interligam os valores desses itens para cada registro na base analisada.

[1]https://www.Python.org/

III. Abordagem Proposta

Nesta seção será apresentada a abordagem proposta neste artigo para geração e visualização de regras de associação entre dados de uso de serviço de um portal de notícias *online*, a qual é composta de quatro etapas: análise exploratória de dados, pré-processamento, geração de regras de associação e visualização de resultados. Na Figura 1 é ilustrado o sequenciamento dessas quatro etapas.

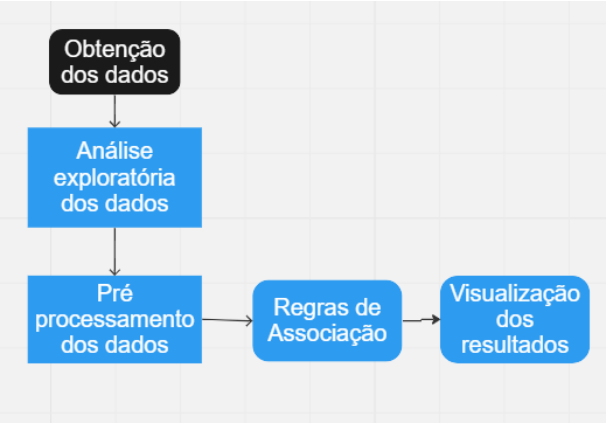

Fig. 1. Fluxograma das etapas da abordagem proposta

A etapa de análise exploratória de dados (*Exploratory data Analysis* - EDA) destina-se a entender o contexto da aplicação através do emprego de técnicas de visualização de características qualitativas e quantitativas dos dados, realizando-se para isto análise dos dados de maior interesse, dados necessários de codificação e indicação de geração de novas *features*, se necessário.

A etapa de pré-processamento dos dados está diretamente ligada ao resultado da EDA, onde a partir do conhecimento obtido sobre a base analisada é possível preparar adequadamente os dados para que se possa utilizar de forma eficiente os algoritmos de mineração de dados. Assim, nessa etapa ocorre o tratamento dos dados relevantes que foram identificados na etapa anterior.

Na etapa de geração de regras de associação utilizou-se o algoritmo Apriori, além de procedimentos de ajustes dos seus hiperparâmetros de modo a encontrar uma boa relação entre número de regras geradas e níveis de suas relevâncias. Na etapa de visualização de resultados, utilizamos os diagramas de barras paralelas, para melhor visualização e compreensão das regras geradas e consequente análise de suas pertinências pelos especialistas [12].

IV. Resultados

O conjunto de dados utilizado neste trabalho é referente ao histórico de acesso dos usuários de um portal de notícias. A base de dados analisada tem informações sobre o conteúdo e o momento do acesso e identificadores únicos para o leitor, ela pode ser visualizada em detalhes na tabela I. O conjunto completo contém 147.155 linhas (registros) entradas e 8 colunas (*features*).

O conjunto de dados foi fornecido em arquivos separado por períodos de 5 dias, compreendendo os meses de setembro, outubro e novembro de 2022, sendo este último período importante para análise pois engloba o período eleitoral de 2022. Para ser utilizado neste trabalho, todos os arquivos foram reunidos em um único conjunto de dados.

TABLE I
DESCRIÇÃO DAS COLUNAS DO CONJUNTO DE DADOS UTILIZADO

Colunas da base de dados		
Nome	***Descrição***	***Tipo do dado***
Título	Título da publicação	Objeto
Slug	URL da publicação	Objeto
reader_id	Identificador numérico único do leitor	float64
Product	Indica se o leitor é assinante ou não	Objeto
post_id	Identificador numérico único da publicação	float64
category	Categoria tema da publicação	Objeto
datetime	Data e hora do acesso registrado	datetime
post_id	Identificador numérico único da publicação	float64

A. Análise Exploratória dos Dados

A fase de análise exploratória dos dados analisou o histórico de acessos aos conteúdos da base obtida e objetivou compreender quais tipos de publicações os leitores acessaram com mais frequência em determinados períodos do dia ou da semana.

Como pode ser visto na Figura 2, para a coluna *category*, que corresponde ao tipo das notícias acessadas, observa-se que em torno de dois terços dos acessos estão concentrados em três tipos de notícias.

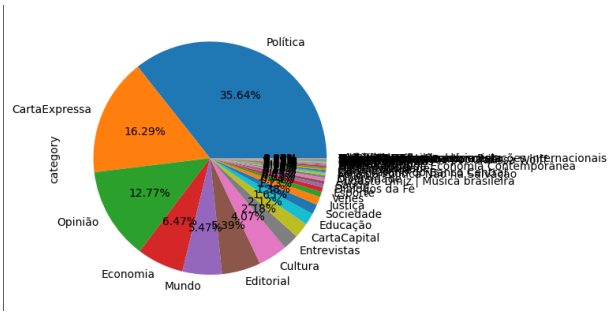

Fig. 2. Distribuição dos tipos de categorias das publicações

Ao se analisar a quantidade de acessos por hora do dia, Figura 3, verifica-se que os picos de acessos se concentram nos períodos entre 12h e 13h e entre 18h e 20h. O período entre 4h e 8h apresenta a menor quantidade de acessos. Também pode-se verificar que os períodos do dia (manhã, tarde e noite) têm padrões de acessos distintos.

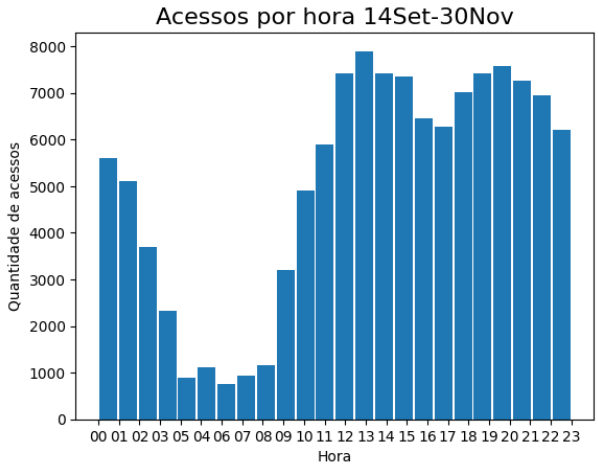

Fig. 3. Quantidades de acessos a notícias por hora

Para melhor compreender o comportamento dos usuários ao longo da semana, analisou-se a quantidade de acesso extratificado por dia da semana e categoria de notícias, Figura 4. Com isto, pode-se observar que a distribuição entre os tipos de notícias permanece praticamente inalterada, porém segunda-feira e sexta-feira foram os dias que apresentaram maiores quantidades de acessos.

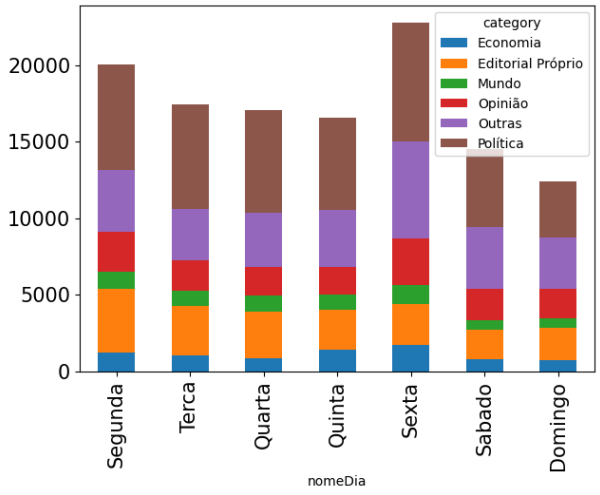

Fig. 4. Quantidade de acessos a notícias por dia e categoria

Observando as dez publicações mais acessadas no período de análise, Figura 5, verificou-se que o número de acessos da publicação mais lida é em torno de 1.400. Considerando que a base de dados completa e limpa tem o registro de 120.804, vemos que a notícia mais lida corresponde a pouco mais de 1% dos acessos.

Na Figura 6 são apresentados os dados de acesso referentes às notícias mais lidas em cada uma das quatro semanas do mês de novembro, visto que a revista publica edições semanais. Esse resultado indica que a notícia mais lida da semana tem os seus picos de acessos nos primeiros dias de sua publicação,

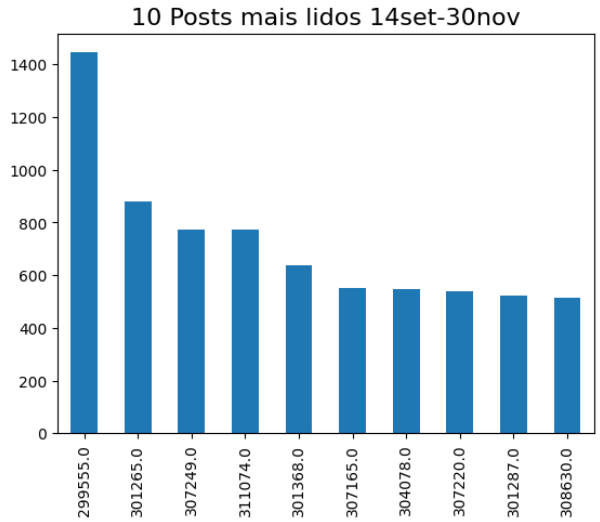

Fig. 5. Publicações mais lidas

caindo consideravelmente a quantidade de seus acessos ao longo dos dias seguintes. Com isto, pode-se observar que o tempo de interesse pelas notícias corresponde a alguns poucos dias e em alguns casos até um ou dois dias. Este resultado corrobora o fato das notícias mais lidas não corresponderem a grandes percentuais dos acessos.

Ao analisar esses resultados, possivelmente ao se aplicar um algoritmo de geração de regras de associação, como o Apriori, o valor do índice de suporte, que mede a proporção da ocorrência de um item sobre o total de registros da base de dados, tende a ser baixo. Além disto, os acessos se concentram em poucos tipos de notícias.

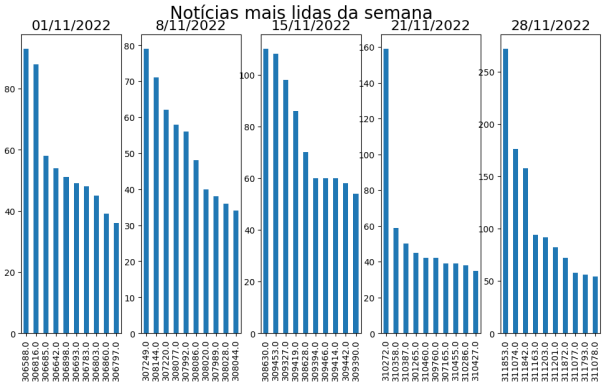

Fig. 6. Publicações mais lidas por semana (Nov/2022)

B. *Pré-processamento*

Baseando-se nas conclusões obtidas na etapa de EDA, a informação foi organizada de modo a maximizar o desempenho do algoritmo de geração de regras de associação. Para isto, as datas de acessos foram divididas em duas colunas: *nomeDia* e *hora* correspondendo, respectivamente aos dias da semana e aos períodos do dia, que possui três valores: manhã (das 4h às 11h), tarde (das 12h às 17h) e noite (das 18h às 3h). Foram considerados ainda apenas seis tipos de notícias, sendo os cinco tipos mais acessados e o tipo **Outras**, que reúne as demais categorias existentes. Na Figura 7 é apresentada a distribuição do quantitativo de acesso a notícias por tipo de assunto após o procedimento de redução de cardinalidade.

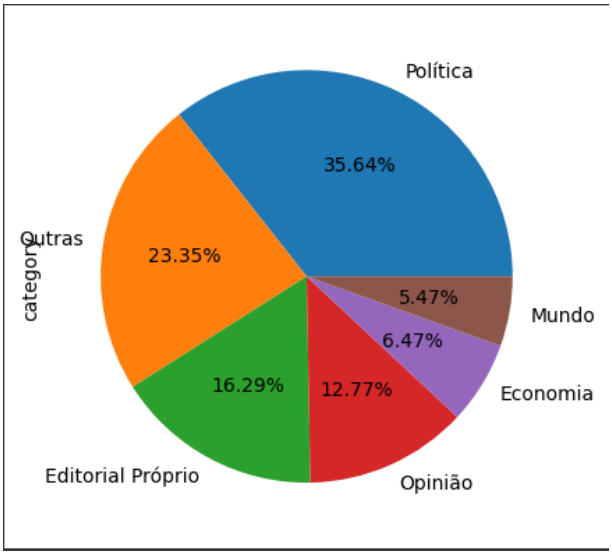

Fig. 7. Quantidade por tipo de publicação depois da redução de cardinalidade

Após o pré-processamento foram consideradas apenas 4 colunas (*category*, *hora*, *product* e *nomeDia*) para a geração das regras de associação. Para iniciar a análise foi gerado o diagrama de barras paralelas que leva em consideração a categoria, período do dia e tipo do leitor (assinante - CASD ou CADASTRADO) obtendo um indicativo inicial das associações na base completa, Figura 8.

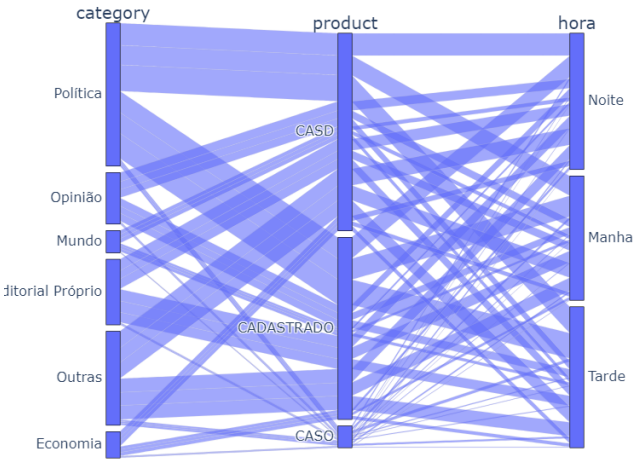

Fig. 8. Diagrama de barras paralelas da base tratada

O diagrama de barras paralelas mostra que a categoria **Política** tem uma alta frequência junto com leitores assinantes

que leem a tarde. Para a categoria **Editorial Próprio** a maior frequência é de leitores CADASTRADOS lendo à noite, como também para a categoria **Outras**, porém para leitores assinantes digitais (CASD).

Finalmente, para se analisar as associações entre as publicações e identificar postagens lidas em sequência, foi realizado o agrupamento dos dados pela coluna *reader_id*. Assim, construiu-se uma base onde cada linha corresponde a um determinado leitor e as colunas correspondem às notícias acessadas por ele com respectivos horários e dias, correspondendo às transações que o algoritmo necessita para executar. Assim, foi possível ter o histórico de acessos realizados por cada leitor.

C. Geração e Visualização de Regras de Associação

Nesta subseção são apresentados os resultados de geração de regras de associação e visualizações via gráficos de barras paralelas, que correspondem às terceira e quarta etapas da abordagem proposta neste trabalho.

O objetivo principal dessa etapa é ter uma quantidade de regras geradas que nos permitem verificar a sequência em que os conteúdos estão sendo associados, mostrando assim a frequência que esses itens são acessados juntos, de modo que ao serem fornecidas aos especialistas de negócio, as regras podem indicar o melhor fluxo de leitura do portal de notícias para contribuir com o objetivo de engajar os leitores com boas indicações a se ler na sequência.

Foram avaliados três cenários: o primeiro sendo com agrupamento pela coluna *reader_id* representando as transações de um mesmo leitor, o segundo e terceiro compreendendo as regras que consideram os parâmetros da categoria da notícia, tipo do leitor e período do dia, sendo geradas regras com *lift* maior que um e menor que um, onde poderemos avaliar também o cenário com regras de itens que podem ter impacto negativo. Para determinar os hiperparâmetros foram levadas em consideração as análises realizadas na etapa de EDA, como indicativo principal um suporte baixo, visto que, de acordo com a Figura 5, uma publicação considerada mais lida tem em torno de 1.400 acessos de uma base de 147.155 acessos (entradas), sendo a definição por experimentação a partir desse indicativo.

Deste modo, no primeiro cenário configurou-se os hiperparâmetros do algoritmo Apriori para considerar transações que contêm os itens em 1,5% da base (min_support=0,015) e que têm no mínimo 20% de chances de ocorrerem (min_confidence=0,2) e com *lift* maior que 1,0, ou seja, antecessor e sucessor dependentes (min_lift=2). Com essa configuração paramétrica foram geradas 28 regras de associação entre as publicações (*posts*).

Para analisar as regras mais relevantes foi aplicado um filtro na saída das 28 regras geradas de modo que ficassem apenas regras que contivessem as 10 publicações mais acessadas, ou seja, as 10 publicações com mais ocorrências de acessos na base de dados. Na Tabela II está exibida a lista das 10 publicações mais lidas da base com suas respectivas categorias.

TABLE II
LISTA DAS 10 PUBLICAÇÕES MAIS LIDAS

Colunas	
ID	***Categoria***
299555	Outras
301265	Opinião
307249	Outras
311074	Política
301368	Outras
307165	Opinião
304078	Outras
307220	Outras
301287	Política
308630	Outras

As regras geradas para as transações de acesso de um mesmo leitor podem ser observadas na Tabela III, onde são indicadas as 16 associações mais relevantes dos acessos entre as notícias publicadas. Essas regras de associação podem ser visualizadas na Figura 9 via gráficos de barras paralelas.

Na Tabela III estão listados os identificadores das publicações que foram consideradas para as regras como antecessores e sucessores, sendo geradas pelo algoritmo *Apriori* onde na primeira parte seleciona-se os subconjuntos da base obedecendo o suporte mínimo e na sequência, utilizando os subconjuntos já encontrados agrega ao parâmetro da confiança mínima, logo, busca-se os subconjuntos mais frequentes e estende-se a partir deles para gerar os candidatos considerando a frequência que ocorrem juntos.

Considerando as publicações mais acessadas (Tabela II) e analisando-se as regras geradas, pode-se observar que os antecessores são das categorias **Outras**, **Opinião** e **Política**. Além disto, as regras mostram uma frequência de publicações da categoria **Outras** com a categoria **Política** e a associação que mais ocorre, sendo uma confiança de 34%, é a leitura de duas publicações de **Opinião**.

TABLE III
ASSOCIAÇÕES DOS ACESSOS ENTRE PUBLICAÇÕES

Associações dos acessos entre publicações				
Antecessor	***Sucessor***	***Suporte***	***Confiança***	***Lift***
299555	296627	0,022	0,560	4,220
299555	298054	0,018	0,580	4,380
299555	299373	0,023	0,570	4,300
299555	301287	0,020	0,370	2,810
301368	299555	0,028	0,210	3,350
299555	304078	0,021	0,420	3,180
307249	299555	0,020	0,310	2,360
301265	307165	0,028	0,340	6,310
301368	301287	0,020	0,490	6,250
301368	302616	0,018	0,280	5,930
301368	304078	0,016	0,260	5,150
307249	304078	0,019	0,380	5,950
307249	305547	0,015	0,410	6,460
307249	307220	0,017	0,380	6,090
307249	308630	0,018	0,290	6,600
311074	311203	0,015	0,230	6,270

No cenário dois, com os parâmetros que consideram transações que contém os itens em 0,5% da base

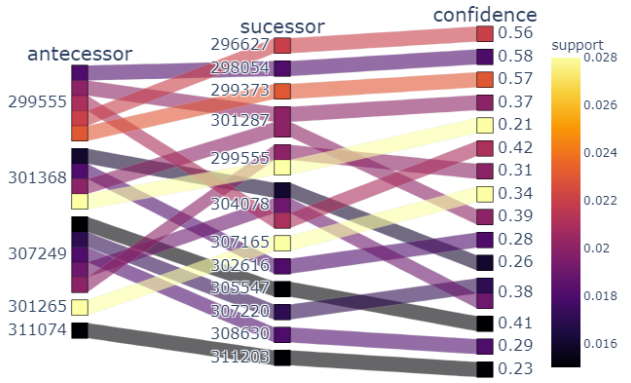

Fig. 9. Diagrama de barras paralelas para regras entre publicações

(min_support=0,005) e que tem no mínimo 15% de chances de ocorrerem (min_confidence=0,15) considerando antecessor e sucessor dependentes (min_lift=1,4), foram geradas 5 regras de associação entre as características dos acessos (*category, hora, product e nomeDia*), exibidas na Tabela IV.

As regras geradas contém dois sucessores, sendo assim, uma regra com três dimensões, considerada uma longa cadeia de informações, são bastante ricas pois consideram o período do dia, a categoria da publicação e o tipo do leitor.

Como podemos observar na Figura 10, 25% dos acessos são referentes às publicações de categoria **Editorial Próprio** pela manhã. Essa regra ganha relevância observando que os acessos no dia de **Segunda** da categoria **Editorial Próprio** possuem a regra que tem 24% de confiança.

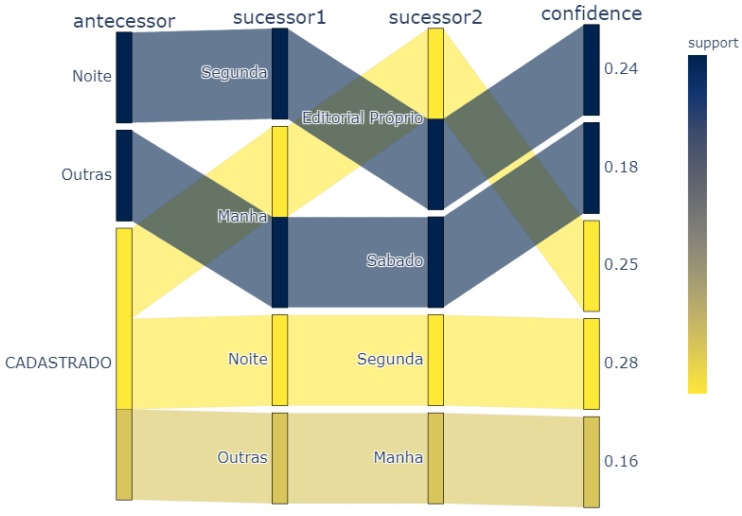

Fig. 10. Diagrama de barras paralelas para regras entre características dos acessos

No último cenário foram geradas regras com o *lift* menor que 1, ou seja, regras que indicam que um item antecessor pode substituir o item sucessor. Considerando um suporte mínimo de 10% e uma confiança de 15%, para o *lift* de 0,9, foram geradas 9 regras, como mostrado na Tabela V.

As regras geradas podem ser denominadas como *regras negativas* e as observações geradas nessas regras nos indicam perfis de leitores diferentes, a sua dimensão é menor do que quando comparado com as regras com *lift* maior que um que utilizam as mesmas dimensões para a sua geração *(category, hora, product e nomeDia)*, nesse caso, aparecem nas regras apenas a categoria e o período do dia, não aparecendo o tipo do leitor, ou seja, não possuem uma cadeia longa de informação.

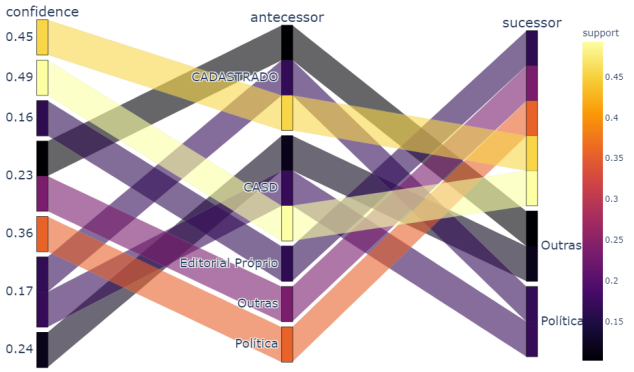

Fig. 11. Diagrama de barras paralelas para regras com *lift* menor que 1

As regras que podem ser visualizadas na Figura 11 são aquelas que encontram itens substitutos. A visualização via gráfico de barras paralelas evidencia que o período do dia (manhã, tarde e noite) e o tipo do leitor (CADASTRADO ou CASD) tem impacto na categoria da publicação que ele irá ler.

A Tabela VI mostra o resumo dos cenários e seus respectivos parâmetros e quantidade de regras geradas.

V. Conclusão

Este artigo teve como objetivo geral desenvolver regras de associação para acessos de publicações de um portal de notícias *online* que irá servir como base para um sistema de recomendação de conteúdos a ser desenvolvido em trabalho posterior, tomando como informações os dados relativos ao histórico de acesso das notícias publicadas de setembro a novembro de 2022.

O objetivo específico foi encontrar associações não evidentes para os especialistas no negócio, validando-as através das heurísticas estabelecidas para o algoritmo *Apriori* e visualizando através do diagrama de barras paralelas.

O uso da *Data Mining* com o algoritmo *Apriori* para a obtenção das regras de associação se baseou em dois subconjuntos da base de dados: Subconjunto de publicações agrupadas por leitor e Subconjunto de características principais

TABLE IV
CARACTERÍSTICAS DOS ACESSOS 1

Características dos acessos 1					
Antecessor	***Sucessor 1***	***Sucessor 2***	***Suporte***	***Confiança***	***Lift***
Editorial Próprio	Segunda	Noite	0,013	0,250	1,500
Outras	Sábado	Manhã	0,013	0,180	1,470
Editorial Próprio	Segunda	CADASTRADO	0,006	0,250	1,530
Editorial Próprio	Segunda	Noite	0,006	0,280	1,710
Sábado	Outras	CADASTRADO	0,007	0,160	1,530

TABLE V
CARACTERÍSTICAS DOS ACESSOS 2

Características dos acessos 2				
Antecessor	***Sucessor***	***Suporte***	***Confiança***	***Lift***
CADASTRADO	Nulo	0,454	0,450	1
CASD	Nulo	0,493	0,490	1
Editorial Próprio	Nulo	0,163	0,160	1
Outras	Nulo	0,234	0,230	1
Política	Nulo	0,356	0,360	1
Outras	CADASTRADO	0,102	0,230	0,960
Política	CADASTRADO	0,168	0,170	1
CASD	Outras	0,118	0,240	1,020
CASD	Política	0,170	0,170	1

TABLE VI
RESULTADOS PARA CADA CENÁRIO

Cenários para regras				
Cenário	***Suporte***	***Confiança***	***lift***	***Regras***
Publicações	0,0015	0,20000	2	28
Características dos acessos 1	0,0050	0,1500	1,4000	5
Características dos acessos 2	0,1000	0,15000	0,5000	9

de acesso. Partindo, assim, da hipótese de ser possível gerar regras de associação confiáveis a partir dos dados de acesso às publicações e utilizando a aprendizagem de dados não supervisionada a partir do histórico, onde serão extraídas informações das experiências passadas.

Foram analisadas associações entre notícias buscando encontrar sequência de leitura, levando em consideração a categoria da postagem, dia da semana e período do dia em que ocorreu o acesso.

As regras de associação geradas trarão um retorno sobre o perfil de navegação dos leitores da revista. Dessa maneira podemos constatar que o objetivo do trabalho foi atingido, tendo em vista o êxito obtido nos resultados demonstrados nas regras de associação entre publicações e associações entre as características do acesso e a sua visualização nos diagramas de barras paralelas.

REFERENCES

[1] AUGUSTO FREITAS LOHMANN; MONAT, A. Sistemas de recomendação de conteúdo em sites de notícias. Arcos Design, v. 9, n. 2, p. 65–76, 1 jan. 2016.

[2] NEWMAN, N. Journalism, Media, and Technology Trends and Predictions 2023. [s.l: s.n.]. Disponível em: ¡https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2023-01/Journalism_media_and_technology_trends_and_predictions_2023.pdf¿. Acesso em: 2 fev. 2023.

[3] GOLDSCHMIDT, R.; PASSOS, E. Data Mining: um guia Prático. [s.l.] Gulf Professional Publishing, 2005.

[4] Vaibhav Verdhan. Mastering Unlabeled Data. [s.l.], v06, 2020.

[5] DAYAN, P. Unsupervised Learning, Appeared in Wilson, RA & Keil, F, editors. The MIT Encyclopedia of the Cognitive Sciences.: MIT. Disponível em: ¡https://web.math.princeton.edu/šswang/developmental-diaschisis-references/dun99b.pdf¿.

[6] Parallel Categories Diagram in Python. Disponível em: ¡https://plotly.com/Python/parallel-categories-diagram/¿. Acesso em: 13 abr. 2023.

[7] Romão, W., Niederauer, C. A., Martins, A., Tcholakian, A., Pacheco, R. C., & Barcia, R. M. (1999). Extração de regras de associação em C&T: O algoritmo Apriori. XIX Encontro Nacional em Engenharia de Produção, 34, 37-39.

[8] Apyori 1.1.2. Disponível em: ¡https://pypi.org/project/apyori/¿ Acesso em: 17 mar. 2023.

[9] Silva, Glauco Carlos, Mineração de Regras de Associação Aplicada a Dados da Secretaria Municipal de Saúde de Londrina - PR / Glauco Carlos Silva - Porto Alegre: Programa de Pós-Graduação em Ciência Computação, 2005.

[10] LAYTON, R. Learning data mining with Python : harness the power of Python to analyze data and create insightful predictive models. Birmingham, Uk: Packt Publishing Ltd., July, 2015.

[11] Heikki Mannila, Hannu Toivonen and A. Inkeri Verkamo, Improved Methods for Finding Association Rules, University of Helsinki, Department of Computer Science, February, 1994.

[12] PENG, T. T., Trevor Campbell, and Melissa Lee Foreword by Roger. Data Science. A First Introduction. September, 2022.

[13] Visualização de dados com python — Ciência de dados - UFF. Disponível em: ¡https://cienciadedadosuff.github.io/cursos/notebooks/caderno-3.html¿. Acesso em: 24 maio. 2023.

[14] Tamara Munzner. Visualization Analysis & Design. Tamara Munzner. Department of Computer Science University of British Columbia. 2014.

[15] TAYO,B. Role of Data Visualization in Machine Learning. 2020. Disponível em: ¡https://medium.com/towards-artificial-intelligence/role-of-data-visualization-in-machinelearning-a6dd62ad1082¿.Acesso em: 21 maio. 2023.

[16] TURKAY,C.;LARAMEE,R.;HOLZINGER,A. On the challenges and opportunities in visualization for machine learning and knowledge extraction: A research agenda. In:.[S.l.:s.n.], 2017. p.191–198. Disponível em: ¡https://www.researchgate.net/publication/319248120_On_the_Challenges_and_Opportunities_in_Visualization_for_Machine_Learning_and_Knowledge_Extraction_A_Research_Agenda Acesso em: 18 Maio.2023.

[17] ALÍPIO MÁRIO JORGE; JOÃO POÇAS; AZEVEDO, P. J. A Methodology for Exploring Association Models. p. 46–59, 1 abr. 2008. Disponível em: https://link.springer.com/chapter/10.1007/978-3-540-71080-6_4 Acesso em 22 Maio. 2023.

[18] YANG, L. Visual Exploration of Frequent Itemsets and Association Rules. p. 60–75, 1 abr. 2008. Disponível em: ¡https://link.springer.com/chapter/10.1007/978-3-540-71080-6_4¿ Acesso em 22 Maio. 2023.

[19] Weitz, Darío. Parallel Sets & Alluvial Diagrams, Why & How. 2021. Disponível em: https://towardsdatascience.com/parallel-sets-alluvial-diagrams-adf40514546 Acesso em 22 Maio. 2023.