Processamento eficiente de regras de associação preferenciais

Rêuder N. Cerqueira Costa[1], João B. Rocha-Junior[1]

[1]Programa de Pós-graduação em Computação Aplicada (PGCA)
Universidade Estadual de Feira de Santana

reudercerqueira@hotmail.com, joao@uefs.br

Abstract. *The discovery of patterns in transactional databases is a well-explored subject and many methods have been proposed to solve this problem. One of the most well known method is the mining for association rules. However, it is difficult to select the best rules using this method, because it requires a good number for support and confidence, which is not easy to set. In order to overcome this problem, preference-based mining rules methods have been proposed. They help in the process of finding the best rules, taking into account the preferences of the users. Unfortunately, these methods are costly to process. In this research, we present novel algorithms for processing preference-based mining rules efficiently. The algorithm proposed are evoluted using real datasets.*

Keywords: *Association Rule. Data Mining. Preferential Queries. Database. Transaction*

Resumo*. A descoberta de padrões em bancos de dados transacionais é um assunto bem explorado e muitos métodos têm sido propostos para resolver estes problemas. Um dos métodos mais conhecidos é a mineração de regras de associação. No entanto, é difícil selecionar as melhores regras usando esse método, porque ele requer um bom número para o suporte e para confiança, o que não é fácil de configurar. Para superar esse problema, métodos que utilizam as preferências do usuário na obtenção de regras de associação foram propostos. Estes métodos auxiliam no processo de encontrar as regras mais interessantes, levando em consideração as preferências dos usuários. Infelizmente, estes métodos são caros para serem processados. Neste artigo, apresentamos novos algoritmos para obter regras de associação, levando em consideração as preferencias dos usuários, de forma eficiente. O algoritmo proposto é evoluído em bases de dados reais.*

Palavras-chave: *Regras de Associação. Mineração de Dados. Consultas Preferenciais. Bando de Dados. Transação.*

1. INTRODUÇÃO

As corporações que operam em diversos setores como varejo, indústria e comércio armazenam uma enorme quantidade de dados durante suas operações comerciais. Por exemplo, grandes redes de hipermercados realizam grandes quantidades de transações de vendas dos seus produtos que são lançados e armazenados em bancos de dados distribuídos por diversos pontos de vendas localizados em suas filiais. Consequentemente o número de informações relacionadas as transações geradas vão se

acumulando com o decorrer do tempo. Todos esses aspectos relacionados ao crescimento destes dados criam novas oportunidades para extrair informações úteis que podem ser exploradas [Villars et al. 2011].

Uma estratégia para buscar informações em grandes bases de dados é o processo de seleção das regras de associação que tem como principal objetivo extrair informações que são úteis em bases de dados transacionais, auxiliando os usuários na tomada de decisões estratégicas de marketing e vendas. Por exemplo, ao analisar uma base de dados através da extração da regras de associação, o usuário pode descobrir que a compra de um determinado produto X implica na compra de um produto Y nas vendas registradas. Em posse deste conhecimento, o gerente de vendas poderá colocar os dois produtos na mesma prateleira aumentando as possibilidades de venda dos produtos X e Y. Sendo assim, é possível avaliar e extrair informações importantes de uma base de dados que estão relacionadas com as compras de clientes em estabelecimentos comerciais [Abaya 2012].

Agrawal et al. (1993) demonstraram que a busca por padrões em banco de dados transacionais tiveram a sua origem na análise de cestas de compras, sendo representada por transações de uma base de dados, formadas por um identificador e por um ou mais itens presentes, que são os produtos comprados por clientes em visita à estabelecimentos comerciais. Então, a avaliação dos dados busca encontrar padrões no comportamento das compras que são realizadas. O processo de prospecção de possíveis padrões realizado com o uso das regras de associação são muito importantes para identificar relacionamentos úteis, que na maiorias dos casos estão ocultos e devem ser avaliados por especialistas [Agrawal et al. 1994].

Assim, dentre os métodos usados no processo de mineração de dados, as regras de associação são muito exploradas [Agrawal et. al. 1994]. As regras de associação apresentam co-ocorrências de produtos nas vendas registradas em base de dados, isto significa que a ocorrência de um produtos X em uma transação de venda acarreta como consequência a presença do produto Y na mesma transação com uma certa probabilidade.

As regras de associação fazem o uso de duas medidas para selecionar as regras mais interessantes em base de dados transacionais, sendo elas o suporte e a confiança. O suporte é a razão da frequência de um item X, ou um conjunto de itens, dentre as transações pertencentes a uma base de dados, sendo o mesmo dividido pela quantidade de transações |D| que pertencem a base de dados D, o valor é obtido através da propriedade, suporte= $\frac{Freq(x)}{|D|}$. Para conceituar confiança é necessário entender como é especificada uma regra de associação, que é a relação entre itens pertencentes ao um banco de dados do tipo X $\rightarrow$ Y, a regra deve ser lida, se X então Y. Onde X é o antecedente da regra e Y é o consequente da regra de associação gerada. Sendo assim, a confiança captura o grau de relacionamento entre os itens que compõem uma regra onde é calculado a frequência da regra de associação dividido pela frequência do item

antecedente da regra, a propriedade é definida como, $Confian_{\text{ç}}a(X \Rightarrow Y) = \frac{sup(X \cup Y)}{sup(X)}$.

A seleção das regras de associação em base de dados transacionais é um dos problemas mais estudados em mineração de dados. Esta abordagem investiga o relacionamento de itens em transações, e tem aplicação nas mais diversas áreas como, lojas de varejo, serviços e indústria [Sahoo et al. 2015].

A Tabela 1 contém um conjunto de transações que reflete uma compra hipotética de produtos em um supermercado. Cada compra corresponde a uma transação. Cada transação é composta por quatro itens (produtos) comprados. Assim, a transação 100 é composta pelos produtos {arroz, feijão, pão, cerveja}, cada produto corresponde a um item da transação.

Tabela 1. Conjunto de Transações

ID	Transação
100	{**arroz**, **feijão**, pão, cerveja}
200	{pão, fralda, **feijão**, **arroz**}
300	{**feijão**, café, pão, fralda}
400	{**arroz**, **feijão**, café, pão}
500	{fralda, **arroz**, pão, café}
600	{**arroz**, fralda, cerveja, **feijão**}
700	{cerveja, **feijão**, **arroz**, café}
800	{fralda, café, **arroz**, cerveja}

Fonte:[Agrawal et al. 1994]

Considerando os dados apresentados na Tabela 1, o suporte do produto arroz é 87%, porque ele aparece em 7 transações de um total de 8, *sup*=$(\frac{7}{8}) = 87\%$. Já o suporte do {arroz, pão} é $(\frac{4}{8} = 50\%)$ porque o número de transações que contém {arroz, pão} é igual à 4, de um total de 8 pertencentes ao conjunto de transações. A propriedade da confiança é determinada como, $Conf(X \Rightarrow Y) = \frac{sup(X \cup Y)}{sup(X)}$. Então a confiança da regra arroz$\Rightarrow$pão é representada da seguinte forma, $Conf(arroz \Rightarrow pao) = \frac{sup(arroz \cup pao)}{sup(arroz)} = (\frac{0.5}{0.87} = 0{,}57)$. Ou seja, em 57% das transações que contém arroz, o pão também está contido nas mesmas.

Para grandes bases de dados, é difícil definir um bom valor para suporte e confiança que viabilize a seleção das regras mais interessantes. O principal problema dessa abordagem é: a dificuldade para definir bons valores para o suporte e confiança que facilite a seleção das regras de associação desejadas.

Para selecionas as regras de associação em grandes bases de dados, o usuário precisa estimar valores para suporte e confiança diversas vezes até alcançar o resultado esperado. Caso o usuário utilize um valor alto para estimar as medidas suporte e confiança, o resultado pode conter poucas ou nenhuma regra, se o usuário utiliza um valor baixo para as medidas suporte e confiança, muitas regras de associação serão retornadas, dificultando o processo de seleção das regras de interesse.

Além disso algoritmos que usam o modelo suporte/confiança para seleção das regras de associação geram dezenas de regras redundantes ou duplicadas [Zheng et al. 2001]. Assim, não só aumenta o custo de processamento, mas também torna difícil analisar e aplicar os resultados produzidos pois as regras geradas podem não traduzir as intenções de pesquisa dos usuários [Tran et al. 2017].

Assim, várias propostas são apresentadas para facilitar o processo de seleção das regras de associação entre itens pertencentes a base de dados transacionais [Bouker et al. 2012, Davis IV et al. 2009, Luna et al. 2014]. Algumas destas abordagens utilizam regras de associação preferenciais. Bouker et al. (2012), Bouker et al. (2013), Mohammed et al. (2015) e Tran et al. (2017) demonstram em seus trabalhos como utilizar as preferências dos usuários para selecionar as regras de associação. As regras de associação preferenciais apresentam como ponto positivo a simplificação do processo de seleção da regras de associação, mas para alcançar esse objetivo é necessário executar uma série de comparações entre as regras de associação geradas, sendo esse processo muito custoso. Ou seja, por uma perspectiva o processo de seleção da regras é simplificado e por outra o custo do processamento é comprometido.

Este trabalho visa desenvolver algoritmos para selecionar as regras com eficiência. Existem várias formas de selecionar as regras de associação preferenciais, nós optamos por duas consultas muito populares, a consulta top-k e a skyline. A consulta top-k determina um ranking das melhores regras de associação através de uma função de escore. A função de ranqueamento permite aplicar diversos pesos a atributos da regra, neste trabalho utilizaremos apenas o suporte e confiança.

A abordagem que utiliza a consulta skyline seleciona todas as regras de associação que não são dominadas por nenhuma outra em termos das medidas suporte e confiança. Assim como na abordagem top-k, optamos por utilizar apenas os atributos suporte e confiança na consulta skyline. Uma regra domina a outra quando o suporte e a confiança dela é melhor do que da outra, em nosso trabalho as maiores regras em termos de suporte e confiança serão selecionadas.

O objetivo principal deste trabalho é propor novos algoritmos para processar regras de associação em grandes bases de dados. Os algoritmos propostos serão avaliados e aplicados à bases de dados reais e sua eficiência analisada.

As principais contribuições deste trabalho são: 1) Especificar as consultas que serão utilizadas para minerar regras de associação preferenciais; 2) Desenvolver algoritmos para processar essas consultas preferenciais de forma eficiente; 3)Avaliar os algoritmos propostos aplicado os mesmos a base de dados reais.

O artigo está dividido da seguinte forma, na Seção 2, fundamentação teórica, Seção 3 trabalhos relacionados, Seção 4 metodologia, Seção 5, avaliação dos resultados.

2. FUNDAMENTAÇÃO TEÓRICA

Neste capítulo encontram-se os conceitos essenciais para o desenvolvimento deste trabalho. A Seção 2.1, contém o conceito de mineração de dados. Na Seção 2.2 falamos sobre regras de associação e na Seção 2.3 consultas preferenciais.

2.1 Mineração de Dados

Segundo Fayyad et al. (1996), a mineração de dados ou *Data Mining* é um processo de extração de informações de uma grande base de dados para auxiliar na tomada de decisões e pode ser aplicado em diversas áreas com objetivo de analisar os dados e prever comportamentos futuros. Assim "A Mineração de Dados é a análise dos conjuntos de dados observacionais, para encontrar relações insuspeitas e para resumir os dados de maneira compreensíveis e úteis para o proprietário dos mesmos" (Hand et al., 2001, p. 1).

O processo de mineração de dados, para extrair o conhecimento, utiliza diversas técnicas como, classificação, modelos de relacionamento entre variáveis, análise de agrupamento, sumarização, modelo de dependência e regras de associação. Os métodos de descoberta de conhecimento em grandes bases de dados (*KDD*) tem sido estudados nas últimas décadas, devido ao tamanho e diversidade das bases de dados. Novos modelos e ferramentas foram desenvolvidos para extrair conhecimentos dessas coleções permitindo a análise dos mesmos [Ribeiro et al. 2013].

As fases do processo de mineração de dados e descoberta de conhecimento são apresentados. Estas fases são, coleta de dados, seleção, pré-processamento, transformação, mineração de dados, interpretação e Resultado.

- **Coleta dos Dados**, os dados são armazenados de forma estruturada. É onde será definida uma melhor compreensão da aplicação a ser tratada bem como identificar as características da base de dados, determinando o conhecimento que pode contribuir para atingir o objetivo [Chiara 2003].

- **Seleção de Dados** estabelece uma amostra do domínio que será executado no processo de descoberta. A seleção dos dados deve ser criteriosa pois, trata-se de uma das fases mais importantes do processo de descoberta de conhecimento [Fayyad et al. 1996].

- **Pré-Processamento** O objetivo é definir uma amostra que de fato represente todo o conjunto garantindo a qualidade dos dados, por exemplo, verificar a cardinalidade, dados inconsistentes e nulos [Claro et al. 2014];

- **Transformação**, nesta fase os dados são armazenados e as técnicas que preparam a base de dados para o processo de mineração são aplicadas [Silva 2004].

• **Mineração de Dados** é processo aplicado para extrair padrões ocultos em bases de dados com auxilio de algoritmos que avaliam estatisticamente a base de dados [Costa et al. 2013, Claro et al. 2014].

• **Interpretação** aqui é realizada análise dos resultados obtidos através da *mineração de dados*. Para Costa et al. (2013) esta análise deve ter coerência, caso contrário, será necessário repetir o processo até obter resultados satisfatórios.

• **Resultados/Conhecimento** é a descoberta do conhecimento o processo que possibilita que o usuário possa entender o comportamento dos seus dados e interpretar os resultados obtidos após a sua execução. Os padrões que são descobertos são analisados e um grau de certeza é estabelecido [Chiara 2003].

2.2 Regras de Associação

As regras de associação permitem realizar à buscar por padrões em grandes base de dados e muitos trabalhos foram desenvolvidos explorando e discutindo os problemas inerentes a esta abordagem como, alto custo computacional para gerar todas as regras de associação, grande quantidade de regras de associação geradas, [Agrawal et al. 1994, Sahoo et al. 2015, Luna et al. 2014].

Vários algoritmos foram propostos com objetivo de reduzir os problemas encontrados nessa metodologia, que consiste em gerar todas as associações entre os itens pertencentes há uma base de dados transacional [Abaya 2012]. As regras de associação tem como premissa básica a descoberta de elementos que implicam na presença de outro em uma mesma transação de uma base de dados transacional conceito introduzido por [Agrawal et al. 1993].

2.3 Consultas Preferenciais

As Consultas tradicionais aplicadas em bases de dados apresentam poucas alternativas para as selecionar características dos dados que serão recuperados [Lacroix and Lavency 1987]; obtendo como resposta um conjunto muito grande de respostas, ou um conjunto pequeno, de dados recuperados. Então, o tratamento das preferências do usuário tem se tornado uma ferramenta muito útil para Sistemas de Informação. Preferências estas que são usadas para reduzir o volume de dados apresentado ao usuário. As Consultas Preferenciais permitem que o usuário possa expressar as suas preferências de forma mais clara e precisa [Stefanidis et al. 2011].

A todo momento as pessoas fazem escolhas de acordo com seus desejos e intenções. Estudos buscam resolver problemas intrínsecos as pesquisas preferenciais em bancos de dados nas últimas décadas [Kie2002]. Obstáculos referentes as consultas preferenciais são discutidas e modelos estão sendo propostos para entregar resultados cada vez mais precisos, tais como: semântica intuitiva; base matemática concisa.

Segundo Rocha-Junior (2013) as consultas preferenciais são classificadas como quantitativas e qualitativas. As consultas qualitativas identifica a preferência entre os pares de objetos (tuplas) existentes na base de dados, para isto, é realizada um avaliação

subjetiva. Por outro lado, a consulta preferencial quantitativa define as preferências de forma indireta para cada objeto existente no conjunto de dados. Uma função de escore avalia os atributos de um objeto, e gera um valor numérico (escore) que mostra o qual importante é este objeto para as necessidades do usuário.

3. TRABALHOS RELACIONADOS

A Tabela apresenta em seu conteúdo um quadro comparativo de relatos descritos na literatura.

Publicação	Algoritmos	Principais Objetivos	Limitações
[Altaf et al. 2017]	Algoritmo Apriori e as suas variações e o Algoritmo FP-Growth	Eficiência Popular	Excesso de interações
[Dahbi et al. 2016]	Função Regras Dominadas - Ndom Função Regras Dominas e Melhores Medidas - Ndomb	Melhores regras de associação e as melhores medidas de avaliação (Lift, Perl, Ig, etc)	Gerar o score para cada regra e gera ranking para as médidas de avalição, alto custo
[Sahoo et al. 2015]	Algoritmo HUCI-Miner	Reduz o número de regras redundantes	Avaliação semântica
[Zaki 2000]	Algoritmo Eclat	Utiliza propriedades estruturais	combinação de diversas técnicas
[Koh et al. 2014]	Algoritmo FastGrendy, Algoritmo SimpleGrendy	Reduz o espaço de busca de conjuntos candidatos Estima potenciais produtos candidatos	Excesso de Comparações
[Bouker et al. 2013]	Algoritmo Regras de Associação Representativ as	Compactção do número de regras, sendo submetidas a várias medidas de avaliação	Alto custo computacional na geração da regras

4. METODOLOGIA

Este artigo tem a sua metodologia dividida em etapas: i) especificação das consultas ii) algoritmos propostos; iii) base de Dados ; iv) avaliação dos resultados. O processo metodológico praticado é conhecido como método experimental, que submete um estudo a alguma variação ou adição de fator para avaliar como o resultado pode alterar PRO-DANOV; FREITAS (2013).

4.1 Especificação das Consultas

A especificação tem como objetivo apresentar os modelos de consultas que serão desenvolvidos. A consulta *Prefrulesky* baseada no operador Skyline apresentado por Borzsony et al. (2001).

Outra abordagem que será desenvolvida é o algoritmo base PrefRuletopk baseado na consulta Top-k, definida como um modelo de consulta preferencial quantitativa [Chaudhuri and Gravano 1999].

4.2 Algoritmos Propostos

Durante a etapa de planejamento um algoritmo base é estabelecido para que o controle e a evolução do mesmo possa ser acompanhado pelo autor ou equipe que está participando do processo de desenvolvimento. O algoritmo base serve para nortear o que está sendo planejado, ou seja, é uma amostra que pode ser visualizada.

O algoritmo base permite realizar uma comparação entre o previsto e o realizado e dá elementos para que o trabalho seja avaliado e o seu andamento comparado com outros semelhantes minimizando possíveis desvios. Então, primeiro será criado um algoritmo base usando técnicas do estado da arte com o objetivo de otimizar o método de seleção das regras entendendo como as consultas preferenciais podem auxiliar no processo de seleção das regras de associação em bases de dados transacionais.

4.3 Bases de Dados

A base de dados utilizada é uma base com dados reais extraída do estabelecimento comercial Grupo São Roque Ltda, rede de supermercados localizado na região de Feira de Santana, Bahia. Os dados representam todas as transações dos clientes em compras realizadas no período entre 01 de Janeiro de 2017 até 31 de Dezembro de 2017, a base de dados esta no formato CSV, e armazena 667 MB apresentando em seu conteúdo aproximadamente 633.899 transações e cada transação é composta pelas seguintes características, número das transações, código dos produtos e descrição dos produtos.

A base de dados foi subdividida em tamanhos distintos, o intuito da divisão realizada é analisar o comportamento dos algoritmos base durante processamento deste dados. O Objetivo é submeter os dados com características distintas ao processo de seleção da regras de associação para extrair os melhores resultados através dos algoritmos base implementados.

4.4 Avaliação dos Resultados

Os algoritmos propostos permitem processar as consultas especificadas PrefRulesky e PrefRulesTopK, que serão aplicadas na seleção de regras de associação e os resultados serão avaliados. Na avaliação, pretendemos medir o tempo de resposta dos algoritmos. Para que seja possível, vamos extrair as regras de associação sem estimar as variáveis suporte e confiança e também não será necessário a intervenção do usuário durante o processo de execução.

Para os usuários selecionarem regras de associação interessantes baseados nas métricas de avaliação suporte e confiança é um desafio, pois os mesmos devem realizar várias atualizações nos parâmetros até obter resultados satisfatórios. O motivo pelo qual estamos desenvolvendo os algoritmos, é reduzir o tempo de seleção das regras de associação garantindo um menor tempo de resposta para todo o processo, desde a extração dos conjuntos de itens, remoção dos conjunto infrequentes, definição das regras de associação e seleção das mais interessantes de acordo com o suporte e a confiança de cada regra.

5. ESPECIFICAÇÃO DAS CONSULTAS

Nessa pesquisa vamos focar em dois tipos de consultas preferenciais a primeira baseada na consulta *Skyline* chamada *PrefRuleSky* e a segunda baseada na consulta *Top-k* nomeada como *PrefRulesTopk*.

5.1 PrefRuleSky

O *PrefRulesSky* visa retornar um conjunto de regras preferenciais sem a necessidade do usuário especificar variáveis *targets*. A consulta avalia os objetos pertencentes a uma base de dados através de uma avaliação qualitativa, então cada objeto é selecionado e comparado com todos os outros [Borzsony et~al. 2001].

O *Prefrulesky* recebe como parâmetro uma base de dados transacional *D*, do qual é extraído um conjunto de regras que são acumuladas no conjunto **L**. O algoritmo retorna as melhores regras de associação de acordo com a Equação ??. Para avaliar as regras de associação o *PrefRuleSky* considera apenas as medidas suporte e a confiança de cada regra pertencente a L, geradas a partir de D, onde,L={r1[sup][conf],r2[sup][conf],r3[sup][conf]...rN[sup][conf]}. No presente estudo vamos nos concentrar em avaliar as regras de associação apenas observando as duas das médias supracitadas (suporte e confiança) [Dahbi et~al. 2016, Agrawal et~al. 1994, Silberschatz and Tuzhilin 1996].

O *PrefRuleSky* apresenta características da consulta Skyline, que tem com objetivo determinar um conjunto de regras que não são dominadas por qualquer outra regra pertencente ao conjunto de dados, e através de uma consulta binária as qualidades são avaliadas conforme Equação 1.

$$Q(r\,i\,, r\,j\,) = r\,i\,[\text{suporte}] \leq r\,j\,[\text{suporte}] \wedge r\,i\,[\text{confiança}] < r\,j\,[\text{confiança}]$$

$$\text{r i [suporte]} \geq \text{r j [suporte]} \wedge \text{r i [confiança]} \overset{V}{>} \text{r j [confiança]} \quad (1)$$

Definição: Consulta Skyline. Uma regra ri $\in$ *D, diz que domina outra regra rj* $\in$ *D, denotado como ri* $\prec$ *rj. (1) se em todas as dimensões di* $\in$ *D, ri* $\succeq$ *rj; e (2) em pelo menos uma dimensões dj* $\in$ *D, ri > rj. O skyline é um conjunto das regras Skyliner* $\subseteq$ *D que não são dominadas por qualquer outra regra r em D. As regras escolhidas são chamadas de conjunto de regras skyline.*

A consulta *Skyline* pode ser utilizada para minimizar ou maximizar os atributos. Na nossa abordagem estamos interessados em maximizar, visto que em geral os usuários estão interessados em regra com maior suporte e confiança, desta forma os maiores serão considerados como preferenciais, assim, as regras serão estabelecidas e acumuladas formando um conjunto de regras preferenciais, W $\subseteq$ L.

5.2 PrefRuleTopK

A consulta Top-k tem como objetivo gerar um *ranking* com as melhores regras de associação pertencentes a um conjunto de dados transacional baseada em uma função de ranqueamento. O algoritmo base *PrefRuleTopk* é desenvolvido aplicando os conceitos introduzido pela consulta Top-k. Para a avaliar as regras de associação existem várias medidas descritas por Bouket et al. (2013), mas vamos direcionar o nosso trabalho focando apenas em duas o suporte e a confiança.

Para execução do algoritmo base *PrefRulesTopk*; Além da base de dados transacionais D, o mesmo recebe como parâmetro k e α, onde k é o número de resultados que serão retornados, com o parâmetro alfa será possível balancear os valores entre as medidas suporte e confiança, como descrito na Equação 2.

$$f(r) = (1 - \alpha) * ri[sup] + \alpha * rj[conf] \quad (2)$$

Definição: Dado um conjunto dados transacional D, que contém L, L={r1[sup][conf],r2[sup][conf],r3[sup][conf]...rN[sup][conf]}, onde r representa cada regra, a função f(r), define um peso para cada regra ri,rj $\in$ *D, ri é melhor do que rj se o score de f(ri) > f(rj). Desta forma as regras são avaliadas e um ranking entre as regras é definido. Como resultado da avaliação as k regras com os maiores escores serão retornadas.*

6. ALGORITMOS

Nesta seção vamos descrever o baseline dos algoritmo que serão desenvolvidos visando alcançar os objetivos esperados pelo presente trabalho de pesquisa. A seção está dividida da seguinte forma, Seção 6.1 baseline, Seção 6.2 o baseline do algoritmo *PrefRuleSky* e na Seção 6.3, baseline do algoritmo *PrefRulesTopK*.

6.1 Baseline

O baseline é um protótipo, um guia do que foi planejado já com tudo ou a maioria do que foi estabelecido, ou seja, é a amostra visual de que o projeto está pronto para ser iniciado ou continuado.

6.2 Baseline PrefRuleSky

Figura 1. Baseline PrefRuleSky

Algorithm 1: Baseline PrefRuleSky

```
   Input: D
   Output: W
 1 FreqMin = 1
 2 Boolean dominate
 3 L = {Large k-itemsets}
 4 W = ∅
 5 for (k=2; L_{k-1} ≠ ϕ; k++) do
 6 |   C_k = generation(L_{k-1})
 7 |   if candidates c ∈ C_k then
 8 |   |   c.count ++
 9 |   end
10 |   L_k={C_k| C.count > FreqMin}
11 end
12 L = generationAssociationRules(L_k)
13 foreach r_i ∈ L do
14 |   dominate = false
15 |   foreach r_j ∈ W do
16 |   |   if r_i ≺ r_j then
17 |   |   |   dominate = true
18 |   |   |   break
19 |   |   end
20 |   |   else if r_i ≻ r_j then
21 |   |   |   W ← W - r_j
22 |   |   end
23 |   end
24 |   if (!dominate) then
25 |   |   W ← W ⋃ r_i
26 |   end
27 end
   Result: W
```

Fonte: Própria Autória

Descrição: O algoritmo apresentado recebe como parâmetro um banco de dados **D** e retorna uma window. A execução do algoritmo da linha 4 até 11 incorpora o mesmo princípio do algoritmo Apriori para gerar o grande conjunto Lk. Nesta execução não eliminamos o processo de poda, apenas retiramos os itens sem representação significativa, ou seja, maior que 1. Aqui vamos analisar todos *itemsets* gerados independente da sua frequência no banco de dados D. Na linha 12 são geradas todas as regras de associação.

A partir da linha 13 as regras de associação geradas serão acumuladas conjunto Lk, pois as mesmas são provenientes de um processo de associação entre os itens pertencentes aos conjuntos de dados analisados. Na linha 13 inicia um *loop*, onde cada regra, ri $\in$ Lk será analisada, se existir, então é definido que ri $\succeq$ r1, ou seja, não podem ser iguais.

Da linha 14 até à 26 as seguintes tarefas são executadas, à variável booleana *dominate* é definida como falsa, na linha 17 um novo *loop* é definido com o objetivo de verificar se rj está presente no conjunto W, se existe, verifica, ri é dominada por rj, se verdade, dominate é atualizada recebendo o valor verdadeiro, ou caso contrário, se ri domina rj, neste caso rj é retirada do conjunto W. Caso a condição do *loop* não for válida, será avaliado se a regra em questão não é dominada por nenhuma outra regra pertencente ao conjunto, então a mesma é adicionada no conjunto W.

Por fim, depois de todas as regras pertencentes ao grande conjunto Lk analisadas e todos os loops chegarem ao seu fim, o conjunto W será composto por uma ou mais regras que não são dominadas por qualquer outra pertencente ao conjunto de dados. É válido ressaltar que todas as regras avaliadas são compostas por um *itemset*, as medidas *suporte* e *confiança*, valores que serão usados para definir qual regra não é dominada dentro do conjunto de dados D.

6.3 Baseline PrefRuleTopK

Figura 2. Baseline PrefRulesTopK

Algorithm 2: Baseline PrefRuleTopK

```
   Input: D, k, α
   Output: W
 1 FreqMin = 1
 2 minHeap = ∅;
 3 W = ∅;
 4 L = {Large k-itemsets}
 5 for (k=2; L_{k-1} ≠ ϕ; k++) do
 6 |   C_k = generation(L_{k-1})
 7 |   if candidates c ∈ C_k then
 8 |   |   c.count ++
 9 |   end
10 |   L_k={C_k| C.count > FreqMin}
11 end
12 L = generationAssociationRules(L_k)
13 forall r ∈ L do
14 |   f(r)=(1-α) * r[sup]+α * r[conf]
15 |   minHeap ← minHeap ⋃ r
16 |   if |minHeap| > k then
17 |   |   minHeap.remove()
18 |   end
19 end
20 W ⋃ minHeap
   Result: W
```

Fonte: Própria Autória.

Descrição: O algoritmo apresentado recebe como parâmetro uma base de dados D, um valor **k** e outro α e retorna um conjunto *W* (*window*). Na linha 1 e na de número 2, são inicializados dois conjuntos minHeap (uma estrutura de dados heap) e W (*uma window*) *status*, vazio. A execução do algoritmo da linha 4 até 10 incorpora o mesmo princípio do algoritmo descrito Apriori para gerar o grande conjunto Lk. Será eliminado apenas os itens sem representação significativa, ou seja, com valor menor ou igual à 1.

Assim, vamos analisar todos *itemsets* gerados no conjunto de dados D.

A partir da linha 10 é gerada o grande conjunto Lk que são compostos pelas de regras de associação, pois, os mesmos são provenientes de um processo de associação entre os itens pertencentes aos conjuntos de dados analisados. Cada regra de associação é composta de valores de suporte e a confiança e a descrição do *itemset*. E na linha 12 são geradas todas as regras de associação calculando as medidas suporte e confiança de cada regra.

Na linha 12 um *loop* é definido onde todas as regras pertencentes ao grande conjunto Lk serão analisadas. Na linha 14 uma função de ranqueamento é definida, o valor de α (alfa) definido pelo analista e o suporte e a confiança de cada regra irá indicar um peso para cada regra avaliada. Assim, conforme a linha 15 as regras serão adicionadas a *heap*, conjunto minHeap, na linha 16 definindo um limite para o conjunto minHeap, onde apenas as k's melhores regras serão inseridas no conjunto W, conforme linha 20. E como resultado as k melhores regras de acordo com o peso definido pela função de ranqueamento fará parte do conjunto W.

7. AVALIAÇÃO DOS RESULTADOS

Neste capítulo apresentamos as descrições e principais conceitos utilizados, bem como a metodologia aplicada. Será feita a caraterização da base de dados que será utilizada/analisada e que servirá de base ao trabalho desenvolvido. Este capítulo tem também como finalidade a avaliação dos resultados parciais obtidos

8. PRÉ-PROCESSAMENTO DOS DADOS

A base de dados transacional, no seu estado inicial, continha várias características dos dados, então a base foi exportada para um arquivo com extensão CSV, sendo necessário uma limpeza, pois possuía muitos registros nulos e repetidos que poderia enviesar o processo extração e seleção das regras de associação. Então, na fase inicial foi necessário realizar filtragem e pré-seleção de dados. Esta tarefa foi efetuada através do software TextPad na versão 7.6.4, 32 bits, também foi desenvolvido um algoritmo na linguagem de programação Java para auxiliar no processo. Esta fase de limpeza e pré-seleção de dados está prevista nas fases inicias do processo de descoberta do conhecimento Sachin e Vijay (2012), sendo de extrema relevância uma vez que a partir deste processo os dados serão limpos, selecionados e disponibilizados para o processo de análise

Na fase de pré-seleção dos dados e limpeza foram realizadas as seguintes atividades:

• Remoção das linhas sem qualquer informação útil, incluindo data, hora, valor e quantidade de cada item;

• Remoção das informações pessoais dos clientes;

• Eliminação de itens repetidos quando os mesmos apareciam mais do que uma vez na mesma transação;

• Foi através de uma média ponderada foi definida o limite de 5 itens por transação;

• Foi considerado para o experimento nesta etapa uma amostra de 1000 transações;

A base de dados para o experimento foi reparticionada em tamanhos distintos e algumas alterações em suas estruturas também foram realizadas. Primeiro dividimos de acordo com o número de transações. A divisão foi estabelecida em 5 partes com as seguintes características, 200 transações, 400 transações, 600 transações, 800 transações e 1000 transações.

A base de dados foi submetida a uma nova divisão. Dessa vez, foi definida 5 bases de dados com o número de transações preestabelecido em 500 transações. Após, realizada essa etapa, variamos a largura (quantidade de itens por transação) em cada base, a primeira com 2 itens por transação, a segunda 3 itens por transação, a terceira 4 itens e a quinta com 5 itens por transação na base de dados.

Ao final desta etapa foi obtida várias partes distintas da mesma bases de dados. Cada transação da base de dados tem um número indicativo que corresponde a compra de um ou mais itens efetuada por um cliente em um determinado momento. Cada linha representa uma transação que é identificada pelo código da transação e pode ter um ou mais itens associados a transação em questão. As compras com mais de um item transacionado serão apresentadas na base de dados na mesma linha, conforme descrito na Figura 4.

Na Tabela 2 é possível visualizar como os dados ficaram dispostos na base de dados. Na transação de compra número 3448930 os itens comprados foram os 130273 e o 167174.

Tabela 2. Amostra da Base de Dados

Cod. Transação	Cod. Itens
3448930	130273 167174
3145122	88620 135429 135483 135495 135582 135603
3448931	163423 163540 163628 164242 165633 165666
3145123	135495 135603 135668 136594 159690
3448932	150232 152074 156229 169542 169913
3145124	154238 156226 156642 157643 157946 158891
3448933	164282
3145125	163626 166137 166538 168066 168195 129011
3448934	132797 130750 162234 130811 130814 130812
3145126	129547 130432 142393 146719 166182
3448935	129546 129824 130050 130271 130446 131180

Fonte: Autoria Própria

9. CONFIGURAÇÃO

Um ambiente para realizar os testes foi montado com o objetivo de rodar os experimentos, que é constituído por um computador com a seguinte configuração: processador Intel Core i5-5200U CPU 2.20 GHz x 4; HD 500 GB; Memória 8GB; sistema operacional Linux Ubuntu versão 16.04. Para os resultados apresentados neste trabalho foram considerados o tempo para realizar as consultas de regras de associação preferenciais.

Cada consulta foi executada com 50 repetições, com o intuito de evitar problemas de performance devido a execução de algum processo inesperado em um determinado momento pelo sistema operacional.

A Tabela 3 apresenta os parâmetros que serão estudados nos experimentos submetidos ao algoritmo base *PrefRuleSky*.

Tabela 3. Parâmetros Aplicados ao Algoritmo Base PrefRuleSky

Parâmetros	Valores
Base de Dados 1	200 transações
Base de Dados 2	400 transações
Base de Dados 3	600 transações
Base de Dados 4	800 transações
Base de Dados 5	1000 transações
Base de Dados 6	500 transações e 2 itens/transação
Base de Dados 7	500 transações e 3 itens/transação
Base de Dados 8	500 transações e 4 itens/transação
Base de Dados 9	500 transações e 5 itens/transação

Na Tabela 4 apresenta os parâmetros que serão estudados nos experimentos submetidos ao algoritmo base *PrefRuleTopK*.

Tabela 4. Parâmetros Aplicados ao Algoritmo Base PrefRuleTopK

Parâmetros	Valores
k melhores resultados	7
Alfa (α)	0.5
Base de Dados 1	200 transações
Base de Dados 2	400 transações
Base de Dados 3	600 transações
Base de Dados 4	800 transações
Base de Dados 5	1000 transações
Base de Dados 6	500 transações e 2 itens/transação
Base de Dados 7	500 transações e 3 itens/transação
Base de Dados 8	500 transações e 4 itens/transação
Base de Dados 9	500 transações e 5 itens/transação

10. TEMPO DE RESPOSTA PARA CONSULTA VARIANDO O NÚMERO DE ITENS POR TRANSAÇÃO

Nesta seção vamos avaliar o tempo de resposta de uma consulta para os algoritmos base, *PrefRuleSky*, *PrefRuleTopk*. Para entender o comportamento dos algoritmos a base de dados foi subdividida em 5 partes, variando a quantidade de itens por transação, cada base de dados com 500 transações. Por exemplo: A primeira base de dados com 2 itens, a segunda com 3, a quarta com 4 e a quinta base de dados com 5 itens, e cada base possui um total 500 transações. O tempo de resposta deste experimento foi medido em segundos.

A Tabela 5 apresenta em seu conteúdo os resultados encontrados com a execução do algoritmo baseline *PrefRuleSky*. A Tabela 5 é composta dos seguinte dados, o tempo de execução em segundos e a quantidade de itens por transação da base de dados descrita.

Tabela 5. Resultados encontrados após execução do algoritmo baseline PrefRuleSky

Quantidade de Itens	Tempo de Execução (s)
2	4,046
3	11,801
4	24,136
5	45,481

Fonte: Própria Autoria

A Tabela 6 apresenta em seu conteúdo os resultados encontrado com a execução do algoritmo baseline *PrefRuleTopK*. A Tabela 6 é composta dos seguintes dados, o tempo de execução em segundos e a quantidade de itens por transação da base de dados descrita.

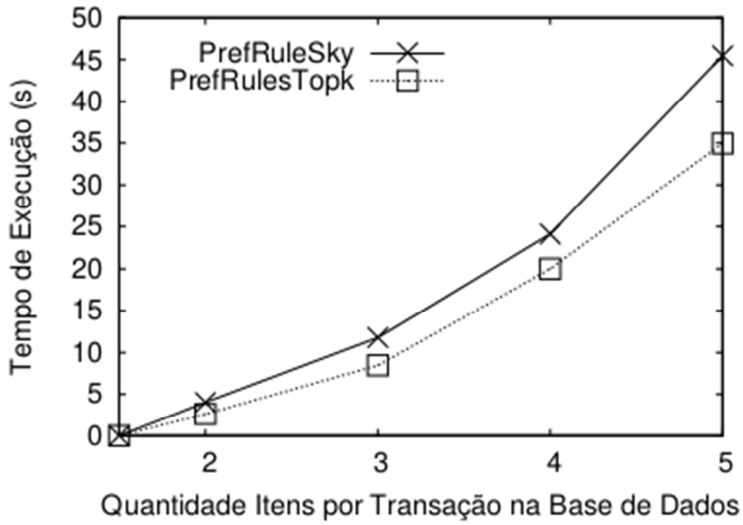
Tabela 6. Resultados encontrados após execução do algoritmo baseline PrefRuleTopK

Quantidade de Itens	Tempo de Execução (s)
2	2,630
3	8,469
4	20
5	34,957

Fonte: Própria Autoria

Na Figura 3 os dados descritos nas Tabelas 5 e 6 foram plotados, dados e estes que foram produzidos pela execução dos algoritmos baseline *PrefRuleSky* e o *PrefRuleTopk*.

Figura 3. Tempo de Resposta Para Execução dos Algoritmos



Fonte: Autoria Própria

Os resultados encontrados após a execução dos experimentos nos permitem realizar algumas considerações a respeito. Apesar dos resultados serem obtidos através dos algoritmos baseline, o primeiro aspecto a ser analisado é que, ao aumentar a quantidade de itens por transação temos como reflexo um aumento no tempo de execução como descrito na Tabela 5 e na Tabela 6, onde o algoritmo baseline PrefRuleSky demonstra um resultado inferior ao algoritmo base PrefRuleTopk. No que se refere ao tempo de execução, quanto menor esse tempo mais eficiente será o algoritmo.

11. TEMPO DE RESPOSTA PARA CONSULTA VARIANDO O TAMANHO DA BASE DE DADOS

A Tabela 7 apresenta em seu conteúdo os resultados encontrados com a execução do algoritmo baseline *PrefRuleSky*. A Tabela 7 é composta dos seguinte dados, tempo de execução em segundos e o tamanho da base de dados.

Tabela 7. Resultados encontrados após execução do algoritmo baseline PrefRuleSky

Tamanho da base	Tempo de Execução (s)
200	4,383
400	24,103
600	70,528
800	147,475
1000	271,898

Fonte: Própria Autoria

A Tabela 8 apresenta em seu conteúdo os resultados encontrados com a execução do algoritmo baseline *PrefRuleTopK*. A Tabela 8 é composta dos seguintes dados, o tempo de execução em segundos e o tamanho da base de dados.

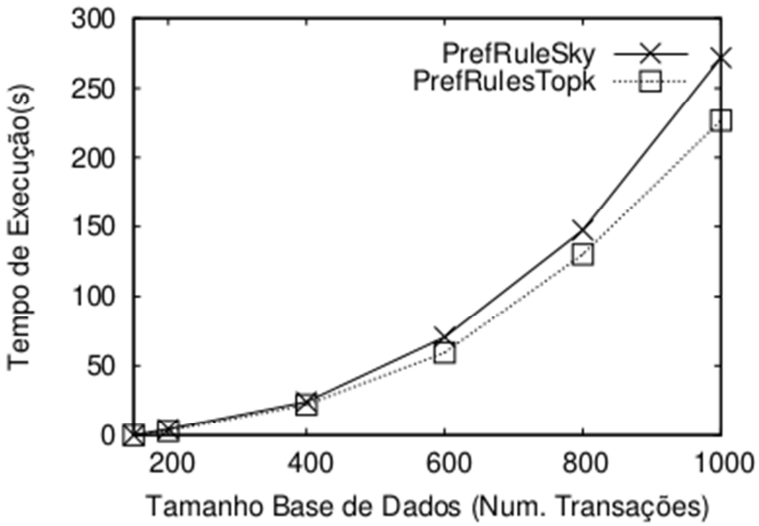
Tabela 8. Resultados encontrados após execução do algoritmo baseline PrefRuleTopK

Tamanho da base	Tempo de Execução (s)
200	3,022
400	22,019
600	59,395
800	130,176
1000	227,474

Fonte: Própria Autoria

Na Figura 4 o algoritmo baseline *PrefRuleSky* e o algoritmo baseline *PrefRuleTopk* foram submetidos a bases de dados com tamanhos distintos conforme descritos nas Tabela 9 e Tabela 10. O tempo de resposta deste experimento está sendo medido em segundos, conforme descrição gráfica.

Figura 4. Tempo de Resposta Para Execução dos Algoritmos



Fonte: Autoria Própria.

No segundo experimento foram aplicadas aos algoritmos bases de dados com tamanhos e características distintas como descrito na Tabela 8 e na Tabela 9. A Figura 4 apresenta o comportamento dos algoritmos a serem submetidos aos testes. É possível visualizar que o algoritmo baseline PrefRulesTopk apresentou resultados mais interessantes em termos de tempo de execução, conforme descrito também na Tabela 10.

12. CONCLUSÃO

Os experimentos realizados neste trabalho evidenciam os obstáculos para elaborar novas estratégias para extrair e selecionar as regras de associação para que seja possível desenvolvermos algoritmos mais eficientes para o processamento de regras de associação preferencias baseados nas medidas de avaliação suporte e confiança. Apesar de apresentar uma conclusão lógica do problema a ser enfrentado, estes resultados nos direcionam para caminhos que viabilizam resultados mais interessantes no contexto apresentado.

O próximo passo deste trabalho é aprimorar os algoritmos base apresentados e avaliar-los em face dos resultados encontrados por algoritmos já existentes. Além disso, avaliar a possibilidade de implementar estratégias de poda inteligentes e eficientes, conforme descritas na Seção de trabalhos relacionados a essa obra com o objetivo de propor uma abordagem que seja capaz de processar as consultas, e após, realizaremos novos experimentos e vamos comparar com os resultados encontrados.

REFERÊNCIAS

[Agrawal et al. 1993] Agrawal, R., Imieliński, T., and Swami, A. (1993). Mining association rules between sets of items in large databases. In *Acm sigmod record*, volume 22, pages 207–216. ACM.

[Agrawal et al. 1994] Agrawal, R., Srikant, R., et al. (1994). Fast algorithms for mining association rules. In *Proc. 20th int. conf. very large data bases, VLDB*, volume 1215, pages 487–499.

[Borzsony et al. 2001] Borzsony, S., Kossmann, D., and Stocker, K. (2001). The skyline operator. In *Data Engineering, 2001. Proceedings. 17th International Conference on*, pages 421–430. IEEE.

[Bouker et al. 2012] Bouker, S., Saidi, R., Yahia, S. B., and Nguifo, E. M. (2012). Ranking and selecting association rules based on dominance relationship. In *Tools with Artificial Intelligence (ICTAI), 2012 IEEE 24th International Conference on*, volume 1, pages 658–665. IEEE.

[Chaudhuri and Gravano 1999] Chaudhuri, S. and Gravano, L. (1999). Evaluating top-k selection queries. In *VLDB*, volume 99, pages 397–410.

[Claro et al. 2014] Claro, D. B., Santos, M. S., Pereira, Q. L., Santana, L. C. d., Silva, M. A. d. S., Teles, A. R. T. F., Lopes, D. C. P., Ribeiro, S. S. C., Lima, V. M. C., and Santos, V. V. d. (2014). Análise da retenção do alunado da ufba via desempenho acadêmico.

[Costa et al. 2013] Costa, E., Baker, R. S., Amorim, L., Magalhães, J., and Marinho, T. (2013). Mineração de dados educacionais: conceitos, técnicas, ferramentas e aplicações. *Jornada de Atualização em Informática na Educação*, 1(1):1–29.

[Dahbi et al. 2016] Dahbi, A., Jabri, S., Balouki, Y., and Gadi, T. (2016). A new method for ranking association rules with multiple criteria based on dominance relation. In *Computer Systems and Applications (AICCSA), 2016 IEEE/ACS 13th International Conference of*, pages 1–7. IEEE.

[Davis IV et al. 2009] Davis IV, W. L., Schwarz, P., and Terzi, E. (2009). Finding representative association rules from large rule collections. In *Proceedings of the 2009 SIAM International Conference on Data Mining*, pages 521–532. SIAM.

[Fayyad et al. 1996] Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI magazine*, 17(3):37.

[Kie2002] Kießling, W. (2002). Foundations of preferences in database systems. In *Proceedings of the 28th international conference on Very Large Data Bases*, pages 311–322. VLDB Endowment.

[Luna et al. 2014] Luna, J. M., Romero, J. R., Romero, C., and Ventura, S. (2014). Reducing gaps in quantitative association rules: A genetic programming free-parameter algorithm. *Integrated Computer-Aided Engineering*, 21(4):321–337.

[Ribeiro et al. 2013] Ribeiro, V. G., Silveira, S. R., Silveira, A. L. M. d., Atkinson, R., and Zabadal, J. R. S. (2013). O emprego de técnicas de mineração de dados para definição de estratégias em processos de divulgação cientfica em periódicos de design. *Strategic design research journal [recurso eletrônico]. São Leopoldo, RS. Vol. 6, no. 2 (May-Aug 2013), p. 85-94.*

[Sahoo et al. 2015] Sahoo, J., Das, A. K., and Goswami, A. (2015). An efficient approach for mining association rules from high utility itemsets. *Expert Systems with Applications*, 42(13):5754–5778.

[Silberschatz and Tuzhilin 1996] Silberschatz, A. and Tuzhilin, A. (1996). What makes patterns interesting in knowledge discovery systems. *IEEE Transactions on Knowledge and data engineering*, 8(6):970–974.

[Stefanidis et al. 2011] Stefanidis, K., Pitoura, E., and Vassiliadis, P. (2011). Managing contextual preferences. *Information Systems*, 36(8):1158–1180.

[Tran et al. 2017] Tran, A., Truong, T., and Le, B. (2017). Efficiently mining association rules based on maximum single constraints. *Vietnam Journal of Computer Science*, pages 1–17.

[Zheng et al. 2001] Zheng, Z., Kohavi, R., and Mason, L. (2001). Real world performance of association rule algorithms. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 401–406. ACM.