Processamento Eficiente de Regras de Associação que Utilizam as Preferências dos Usuários: Uma Proposta

Rêuder N. Cerqueira Costa e João B. Rocha-Junior

[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 methods is the mining for association rules. However, it is difficult to select the best rules using this method, becouse 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 preference of the users. Unfortunately, these methods are memory-consuming to process. In this paper, we present novel algorithms for processing preference-based mining rule efficiently.*

1. Introdução

Os dados estão sendo gerados a todo o momento em intervalos de tempo cada vez menores dando origem a grandes volumes. Esse crescimento teve a sua origem com o aumento demográfico e reflexo da evolução das ferramentas tecnológicas de coleta de dados, redução do valor por giga byte (GB) armazenado e o aumento da capacidade dos dispositivos de armazemanto [Boja et al. 2012].

O volume de dados gerados, sendo parte deles oriundos de visitas de clientes a estabelicimentos comerciais (cestas de compras), também acompanham esse crescimento, surgindo a necessidade de sistematização e classificação do dados. Todos esses aspectos relacionados geram novas oportunidades para extração de informações em grandes bases de dados [Villars et al. 2011].

Um dos principais métodos utilizados para descobrir padrões ocultos em bases de dados transacionais é a mineração de regras de associação. Este método avalia transações em uma base de dados e algortimos são utilizados para extrair informações, sendo um deles algoritmo apriori [Agrawal et al. 1994]. Assim o mesmo é aplicado para descobrir padrões e associações entre os itens pertecentes as bases de dados transacionais [Abaya 2012].

As principais técnicas utilizadas para identificar as associações entre os itens verificam a co-ocorrência entre eles em uma mesma transação. Essas associações são geradas a partir dos bancos de dados que podem armazenar centenas de milhares de itens como os *datasets* presentes em grandes redes de supermercados. Uma das mais importantes aplicações das regras de associação é a análise de cestas de compras, nesta analise o objetivo é avaliar o comportamento das compras realizadas por consumidores, buscando informações estatísticas de quais produtos costumam ser os escolhidos em uma mesma transação (Compra), chamada de venda cruzada.

O cálculo das associações em base de dados transacionais é um dos problemas mais estudados no campo da mineração de dados. Estas abordagens investigam o relacionamento dos itens nas transações pertecentes ao *dataset*, e tem aplicação nas mais diversas áreas como, lojas de varejo, serviços e industria [Abaya 2012, Sahoo et al. 2015].

Assim dentre os métodos usados as regras de associação são muito exploradas [Agrawal et al. 1994]. O processo comumente utilizado para selecionar as regras de associação é através das métricas, *suporte* e *confiança*. O suporte captura a frequência de um item ou um conjunto de itens específicos presentes nas transações do *dataset* explorado, enquanto a confiança captura o grau de relacionamento entre os itens que compõem uma regra.

Assim, as associações devem atender a parâmetros pré-estabelecidos de suporte e confiança para serem selecionadas. Esses parâmetros especificados pelos usuários são relações do tipo X $\Rightarrow$ Y. Onde X é o antecedente da regra e Y é o consequente da associação gerada, descrita na literatura como uma regra de associação.

A Tabela 1 contém um conjunto de dados que reflete uma compra hipotética de produtos em um supermecado. 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, $Conf(X \Rightarrow Y)$ = $\frac{sup(X \bigcup Y)}{sup(X)}$. Então a confiança da regra arroz$\Rightarrow$pão é representada da seguinte forma, $Conf(arroz \Rightarrow pao)$=$\frac{sup(arroz \bigcup pao)}{sup(arroz)}$=($\frac{0.5}{0.87} = 0,57$). 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 de suporte e confiança para selecionar as melhores regras. Essa abordagem apresenta dois problemas principais: i) dificuldade para definir bons valores para o suporte e confiança que facilite identificação das regras desejadas; ii) a grande ou pouca quantidade de regras geradas a partir dos

valores de suporte e confiança definidos pelo usuário, que dificulta a seleção das regras que atendam as preferências do usuário.

Usando como exemplo os dados da Tabela 1, caso o usuário utilize 90% para suporte e 80% para confiança, nenhuma regra é retornada. Se o usuário utiliza os parâmetros 20% para suporte e 20% para confiança, várias regras são retornadas, dificultando a seleção das regras de interesse.

Para resolver esse problema vários trabalhos de pesquisa estão sendo propostos. Alguns destes [Bouker et al. 2012, Bouker et al. 2013, Mohammed et al. 2015, Tran et al. 2017, Tran et al. 2017] utilizam as preferências dos usuários e a redução do número de regras geradas ou redundantes para selecionar as melhores regras de associação. Neste trabalho duas abordagens serão estudadas: 1) uma abordagem que utiliza a consulta *skyline* para filtrar as regras mais relevantes e 2) uma abordagem que utiliza a consulta *top-k*.

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, problema de pesquisa e contribuição, Seção 3, fundamentação teórica e trabalhos relacionados, Seção 4, desenvolvimento para a conclusão, Seção 5, avaliação dos resultados.

2. Problema de Pesquisa e Contribuição

Os métodos de seleção das regras de associação preferenciais apresentam comportamentos e características indesejáveis. Um dos fatores é o alto custo computacional para serem processadas, pois o usuário tem que estimar a priori valores para o suporte e a confiança que atenda com precisão o propósito de sua consulta, outras medidas também podem ser aplicadas aumentando ainda mais a sua complexidade.

Os algoritmos serão desenvolvidos com objetivo de reduzir o esforço dos usuários para conquistarem os resultados com eficiência, aplicando os conceitos de consultas preferenciais e mineração de regras de associação.

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 sua variações Algoritmo FP-Growth	Eficiência Popular	Excesso de intereções

[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
[Mohammed et al. 2015]	Algoritmo MDPref Algoritmo SkyRules Algoritmo ProfMiner	Analisa as preferências dos usuários	Limiar para as variáveis, SkyRules mantém regras entre aquelas que têm o mesmo medida
[Koh et al. 2014]	Algoritmo FastGrendy, Algoritmo SimpleGrendy	Reduz o espaço de busca de conjuntos candidatos Estima potenciais produtos candidatos	Excesso de Comparações
[Rocha-Junior 2013]	Algoritmo AGiDS	Aplica os conceitos de consultas preferenciais de forma eficiente e eficaz	As consultas preferenciais são aplicadas em contexto diferente da nossa proposta
[Bouker et al. 2013]	Algoritmo Regras de Associação Representativas	Compactção do número de regras, sendo submetidas a várias medidas de avaliação	Alto custo computacional na geração da regras

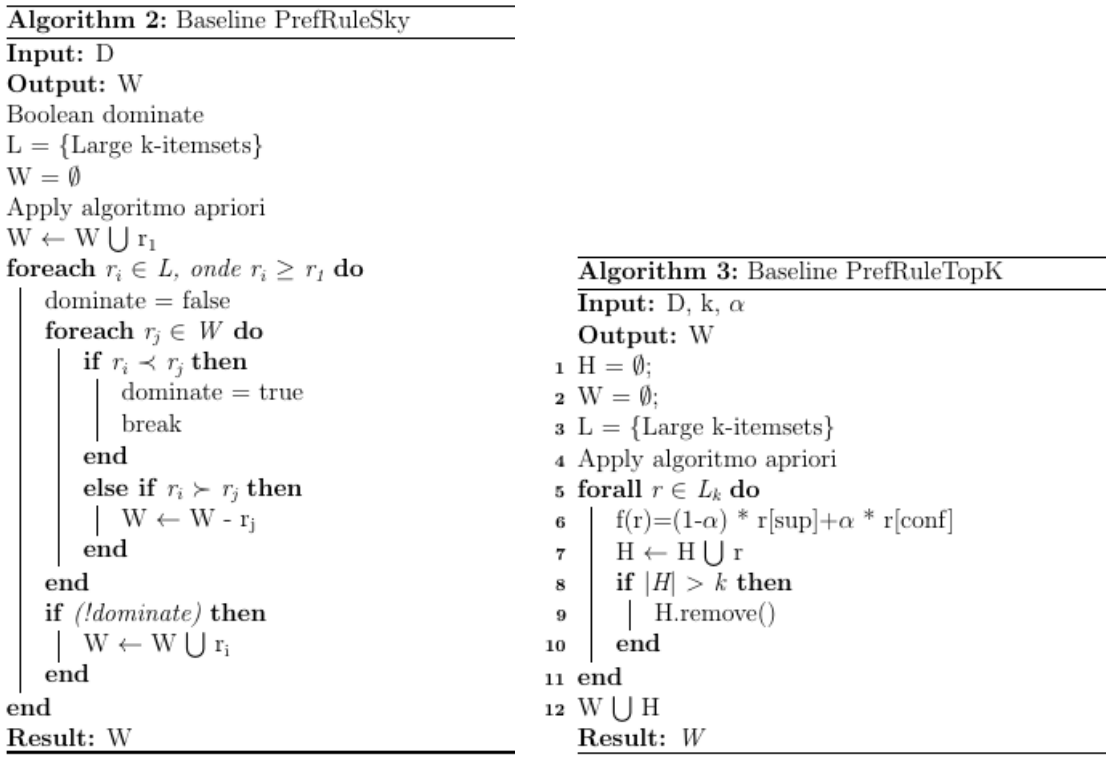
4. Desenvolvimento para a Conclusão

O trabalho está em desenvolvimento, encontra-se na fase de implementação dos algoritmos que estão sendo propostos como parte da metodologia para alcançar os resultados. O *baseline* foi desenvolvido, a base de dados com dados reais foi selecionada. Com aproximadamente 700 MB, 1 milhão de transações referente a itens comprados por clientes em visita a um estabelecimento comercial localizado na cidade de Feira de Santana, no estado da Bahia.

5. Avaliação dos Resultados

Na Figura 1 o baseline dos algoritmos que serão desenvolvidos visando alcançar os objetivos esperados.

Figura 1. BaseLine



Algorithm 2: Baseline PrefRuleSky

Input: D
Output: W

```
1  Boolean dominate
2  L = {Large k-itemsets}
3  W = ∅
4  Apply algoritmo apriori
5  W ← W ⋃ r_1
6  foreach r_i ∈ L, onde r_i ≥ r_1 do
7      dominate = false
8      foreach r_j ∈ W do
9          if r_i ≺ r_j then
10             dominate = true
11             break
12         end
13         else if r_i ≻ r_j then
14             W ← W - r_j
15         end
16     end
17     if (!dominate) then
18         W ← W ⋃ r_i
19     end
20 end
```

Result: W

(a) PrefRuleSky

Algorithm 3: Baseline PrefRuleTopK

Input: D, k, α
Output: W

```
1  H = ∅;
2  W = ∅;
3  L = {Large k-itemsets}
4  Apply algoritmo apriori
5  forall r ∈ L_k do
6      f(r)=(1-α) * r[sup]+α * r[conf]
7      H ← H ⋃ r
8      if |H| > k then
9          H.remove()
10     end
11 end
12 W ⋃ H
```

Result: W

(b) PrefRuleTopK

A Figura 1 descreve o baseline (a) *prefrulesky* que recebe como parâmetro um *dataset* **D** e retorna uma ***window***. Na linha 4 o algoritmo apriori apresentado por Agrawal et. al (1994) é usado para gerar o grande conjunto L_k. Dá linha 6 até a 20 todos os itens do conjunto L_k que contém todas as regras de associação pertencentes a base de dados. Rocha-Junior (2013) aborda o conceito de consulta skyline para definir um ranking entre os elementos pertencentes ao um *dataset*, seguindo esse princípio realizaremos a análise e apenas os regras que não são dominados por qualquer outro pertecente ao conjunto serão inseridos no conjunto **W**, representando um ranking das melhores regras de associação de acordo com medidas suporte e confiança.

A Figura 1 apresenta o baseline (b) *prefruletopk* que recebe como parâmetro um *dataset* **D**, um valor **k**, parâmetro da consulta top-k e α variável que integra uma função de *score* e retorna o conjunto ***W*** uma (*window*). Na linha 1 e na de número 2, são inicializados dois conjuntos **H** (uma estrutura de dados heap) e W (*uma window* ou pilha). O algoritmo apriori é aplicado baseline na linha 4. Agrawal et. al (1994) usaram o apriori para gerar e extrair as regras de associação do dataset, vamos aplicar a mesma estrategia para gerar o conjunto L_k que contém todas as regras geradas. Rocha-Junior (2013) aplica o conceito de consulta top-k para definir um raking entre elementos pertencentes a um *dataset*, assim, da linha 5 à 12 um *ranking* é gerado de acordo a consulta preferêncial e as melhores regras com auxilio de uma *heap* serão selecionadas, onde as **k** melhores regras são inseridas no conjunto **W** e apresentada com resultado.

6. Conclusão

Os algoritmos base acima descritos ainda encontram-se em fase de desenvolvimento e não foram submetidos a testes que avaliam o seu desempenho. Dessa forma o resultado parcial apresentado nesta proposta foi o desenvolvimento do baseline Prefrulesky e Prefruletopk.

Referências

Abaya, S. A. (2012). Association rule mining based on apriori algorithm in minimizing candidate generation. *International Journal of Scientific & Engineering Research*, 3(7):1–4.

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.

Altaf, W., Shahbaz, M., and Guergachi, A. (2017). Applications of association rule mining in health informatics: a survey. *Artificial Intelligence Review*, 47(3):313–340.

Boja, C., Pocovnicu, A., and Batagan, L. (2012). Distributed parallel architecture for"big data". *Informatica Economica*, 16(2):116.

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.

Bouker, S., Saidi, R., Yahia, S. B., and Nguifo, E. M. (2013). Towards a semantic and statistical selection of association rules. *arXiv preprint arXiv:1305.5824*.

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.

Koh, J.-L., Lin, C.-Y., and Chen, A. L. (2014). Finding k most favorite products based on reverse top-t queries. *The VLDB Journal*, 23(4):541–564.

Mohammed, M., Taoufiq, G., Youssef, B., et al. (2015). A new way to select the valuable association rules. In *Knowledge and Smart Technology (KST), 2015 7th International Conference on*, pages 81–86. IEEE.

Rocha-Junior, J. B. (2013). *Efficient Processing of Preference Queries in Distributed and Spatial Databases.* PhD thesis, Norwegian University of Science and Technology, Trondheim, Norway.

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.

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.

Villars, R. L., Olofson, C. W., and Eastwood, M. (2011). Big data: What it is and why you should care. *White Paper, IDC*, 14.