REVISTA eixo revistaeixo.ifb.edu.br

ISSN 2238-5630
Brasília-DF, v. 8, n. 1, janeiro-junho de 2019

Autores | Authors

José Gonçalo dos Santos*
jose.santos@ifb.edu.br
Wilton Lucas Pires Gomes**
willtonlucas@outlook.com

ESTUDO COMPARATIVO ENTRE OS ALGORITMOS TRADICIONAIS E UM ALGORITMO BASEADO EM LÓGICA DIFUSA PARA TRATAMENTO DA IMPRECISÃO CONTIDA NA MATRIZ DE CO-OCORRÊNCIA USADA NO CÁLCULO DE ATRAÇÃO/REPULSÃO ENTRE PRODUTOS (BASE PARA*MARKET BASKET ANALYSIS*)

COMPARATIVE STUDY BETWEEN THE TRADITIONAL ALGORITHMS AND AN ALGORITHM BASED ON FUZZY LOGIC TO TREAT THE IMPRECISION CONTAINED IN THE MATRIX OF CO-OCCURRENCE USED IN THE CALCULATION OF ATTRACTION / REPULSION BETWEEN PRODUCTS (BASE FOR MARKET BASKET ANALYSIS)

Resumo: Este artigo mostra um estudo comparativo entre um algoritmo baseado em lógica difusa e os tradicionais, usados para determinar relações entre produtos no processo de Market Basket Analysis (MBA). Para tanto, inicialmente, fez-se um estudo dos algoritmos utilizados e foram selecionados alguns deles para os testes comparativos. A comparação levou em consideração os resultados obtidos de cada algoritmo. O algoritmo difuso mostrou ótimos resultados e mostrou-se viável no cálculo de atração/repulsão entre produtos. Com isso, esta pesquisa revela a potencialidade do uso da lógica difusa no processo de MBA.

Palavras-chave: Market basket analysis, Lógica difusa, Algoritmos.

Abstract: *This article shows a comparative study between a fuzzy logic-based algorithm and the traditional ones used to determine relationships between products in the Market Basket Analysis (MBA) process. To do so, initially, a study of the algorithms used was done and some of them were selected for the comparative tests. The comparison took into account the results obtained from each algorithm. The diffuse algorithm showed excellent results and proved to be feasible in the calculation of attraction / repulsion between products. With this, this research reveals the potentiality of using fuzzy logic in the MBA process.*

Keywords: *Market basket analysis, Fuzzy logic, Algorithms.*

Introdução

Os avanços tecnológicos e científicos proporcionam formas eficientes para capturar, organizar e armazenar imensas quantidades de dados. Essa facilidade de armazenamento em grande escala bem como a extração de informações valiosas a partir destes, tem contribuído grandemente para a área de pesquisa denominada Descoberta de conhecimento em bases de dados (DCBD), cujo principal objetivo é explorar grandes bases de dados a fim de extrair modelos de conhecimento (padrões) de mais alto nível, servindo de apoio à tomada de decisão, e também, obter melhor entendimento do fenômeno gerador dos dados (GOLDSCHMIDT, 2015).

A DCBD faz uso de diversas técnicas, métodos e algoritmos para que os modelos de conhecimento (padrões) sejam alcançados. Dentre as

Aceito em: 28/01/2019
Recebido em: 11/10/2017

técnicas utilizadas, está a de associação entre produtos, conhecida popularmente como Market Basket Analysis (MBA). O MBA tem grande importância no setor comercial, e visa identificar afinidades entre os produtos comprados pelos clientes, por exemplo, quem compra um produto A também, na maioria dos casos, compra B. Essa técnica possui grande potencial para melhorar as vendas e aumentar os lucros.

Devido a grande importância do MBA para os comércios, muitos modelos e algoritmos vêm sendo estudados a fim de alcançar melhores resultados. Um dos mais utilizados hoje é o Apriori e muitas variações dele também se encontram na literatura. Outros algoritmos com abordagens diferentes também são estudados. Dentre eles, se está o algoritmo baseado nos conceitos da lógica difusa proposto por Santos (2004).

Santos (2004), propôs e testou um algoritmo baseado no uso de conjuntos difusos e lógica difusa para o tratamento de imprecisão contida na matriz de coocorrência, que é de suma importância nos cálculos durante o processo de MBA. Em sua pesquisa o autor procurou verificar a adequação da abordagem difusa para modelar a imprecisão contida na matriz que é utilizada no cálculo da medida atração/repulsão entre itens. No desenvolvimento do seu trabalho, foram avaliadas várias combinações de funções de pertinência em conjunto com os principais modelos de regras, usando várias amostras de associações entre produtos oriundas de base de dados de três segmentos comerciais. Com base nisso, Santos (2004) propôs um método que mapeia entradas numéricas de frequências para termos linguísticos e que possibilita como saída a classificação de associação, podendo ser de atração ou repulsão, com os seguintes graus qualitativos: baixa, moderada ou alta. O método mostrou bons resultados e pode ser aplicado na área comercial para análise de dados históricos de vendas. Além disso, pode ser usado nos pontos de vendas para auxiliar o atendente a oferecer um novo produto a determinados clientes baseado na sua compra atual, porque a resposta do sistema pode ser dada em linguagem natural, o que torna acessível a qualquer usuário do sistema. Pode-se também usar o método para fazer consultas usando linguagem natural.

O trabalho de Santos (2004) mostrou que é perfeitamente possível usar a abordagem difusa para tratamento de imprecisão contida na matriz de coocorrência. Mas, o que se notou na pesquisa, é que não houve um efetivo estudo comparativo entre o algoritmo desenvolvido e os tradicionais, pois foi feita a comparação com apenas um deles. Neste sentido cabe levantar a seguinte hipótese que norteia este trabalho: “O algoritmo desenvolvido por Santos (2004) é viável se comparado com os demais algoritmos para Market Basket Analysis?”.

Assim, o objetivo desta pesquisa é comparar a abordagem difusa proposta por Santos (2004) com os métodos tradicionais usados no cálculo de atração/repulsão entre produtos (base para Market Basket Análises). Para tanto, fez- se necessário um estudo aprofundado do método difuso apresentado por Santos (2004) e dos métodos tradicionais existentes para o mesmo fim. Depois, foram realizados diversos testes e comparados os resultados obtidos de cada algoritmo, e assim, foi possível saber a viabilidade do mesmo no processo de MBA.

O restante do artigo está organizado como segue. A seção 2 apresenta uma breve introdução à mineração de dados e ao processo de MBA. Na seção 3 é apresentado o método difuso proposto por Santos (2004) e os algoritmos tradicionais usados para o estudo comparativo. A seção 4 exibe a metodologia utilizada. A seção 5 mostra os resultados e discussões e por fim, na seção 6, são apresentadas as considerações finais.

Breve introdução à Descoberta de Conhecimento em Base de Dados (DCBD)

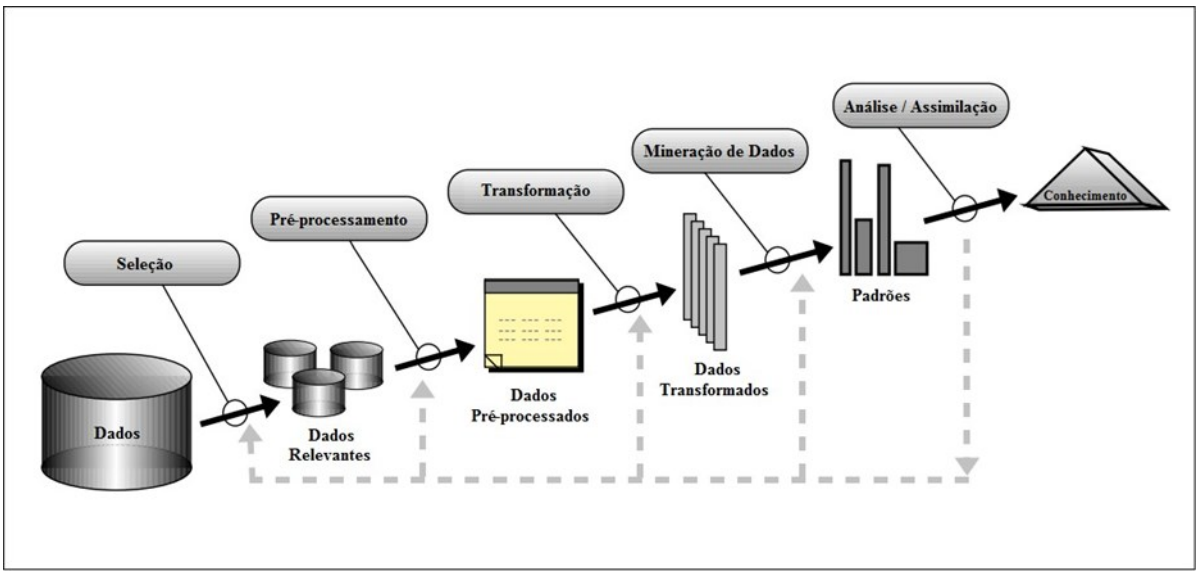
Atualmente produzimos enormes quantidades de dados que são armazenados em bases dos diferentes setores produtivos. Com o avanço da ciência, tornou-se cada vez mais fácil armazenar grandes quantidades de dados, porém, surge um novo cenário cujo problema é saber como melhor aproveitar estes dados. A partir disto, surgiu a necessidade de extrair modelos de conhecimento baseados nesses dados que servissem de apoio à tomada de decisão. Este cenário descrito levou ao surgimento de uma nova área de pesquisa, DCBD, que atende essa necessidade, mediante um processo constituído de várias etapas que faz uso de métodos, técnicas e ferramentas próprias.

As finalidades da aplicação de DCBD são diversas segundo o ambiente trabalhado. Goldschmidt et.al. (2015), por exemplo, cita algumas dessas finalidades, como otimizar procedimentos em uma empresa, aumentar os lucros, ajudar na compreensão dos resultados de um experimento científico, auxiliar médicos a interpretar os efeitos de um tratamento dentre outros propósitos. Dessa forma, para Goldschmidt et.al. (2015), questões do tipo “O que fazer com todos os dados armazenados?”, “Como utilizar o patrimônio digital em

1 Daqui em diante, usaremos a sigla LD para identificar a obra analisada, fazendo-se referência somente da página em que a citação se encontra.

benefício das instituições?", "Como analisar e utilizar todo o volume de dados disponível?", "De que maneira informações subjacentes aos dados armazenados podem ser úteis no contexto ao qual pertencem?", são frequentemente utilizadas e podem ser respondidas mediante métodos, técnicas e ferramentas disponíveis no processo de

Figura 1 - Visão geral das etapas de DCBD adaptado de Fayyad (1996).



DCBD.

As finalidades da aplicação de DCBD são diversas segundo o ambiente trabalhado. Goldschmidt et al. (2015), por exemplo, cita algumas dessas finalidades, como otimizar procedimentos em uma empresa, aumentar os lucros, ajudar na compreensão dos resultados de um experimento científico, auxiliar médicos a interpretar os efeitos de um tratamento dentre outros propósitos. Dessa forma, para Goldschmidt et al. (2015), questões do tipo "O que fazer com todos os dados armazenados?", "Como utilizar o patrimônio digital em benefício das instituições?", "Como analisar e utilizar todo o volume de dados disponível?", "De que maneira informações subjacentes aos dados armazenados podem ser úteis no contexto ao qual pertencem?", são frequentemente utilizadas e podem ser respondidas mediante métodos, técnicas e ferramentas disponíveis no processo de DCBD.

Data mining e regras de associação

O termo em inglês, Data Mining, ou mineração de dados em português, refere-se a uma das etapas do processo de DCBD. Devido ser exatamente nessa etapa que as informações ocultas dentro das bases de dados passarem a ser vistas como informações valorosas, o termo ganhou maior popularidade do que o próprio processo de DCBD em geral. Fayyad et. al. (1996) traz uma definição concisa do termo, dizendo que data mining é um dos passos de DCBD que consiste em aplicar análise de dados e algoritmos de descoberta que são capazes de produzir modelos sobre os dados (padrões).

Na etapa de mineração de dados são utilizadas várias técnicas de análise. Dentre elas está a análise de regras de associação. Segundo Librelotto & Mozzaquatro (2014), Buttar & Kaur (2013) e Tudor (2008), a mineração de regras de associação pode ser vista como uma das mais importantes tarefas de mineração de dados. Para Berry & Linoff apud Côrtes et. al. (2002), a técnica consiste em descobrir as regras de associação condicionadas a valores de atributos que ocorrem juntos em um conjunto de dados, ou seja, determinar quais itens estão relacionados com outros itens. Amo (2004) define uma regra de associação como sendo um padrão da forma X → Y, onde X e Y são conjuntos de valores.

A importância de se utilizar a técnica de associação entre itens está fundada nos benefícios que podem ser adquiridos através de sua aplicação. Uma dessas vantagens, segundo Amo (2004), é na parte comercial, pois sabendo-se a relação dos produtos comprados pelos clientes em um determinado comércio, pode-se identificar aqueles que ocorrem de forma conjunta e assim, dispô-los na prateleira de forma a facilitar (ou dificultar) as compras do usuário, induzindo-o a comprar mais e então, consequentemente, aumentar as vendas e os lucros. Este é um cenário comumente encontrado na área de Market e será melhor discutido no tópico seguinte.

ISSN 2238-5630
Brasília-DF, v. 8, n. 1, janeiro-junho de 2019

Market Basket Analysis (MBA)

O MBA é o processo que utiliza a técnica de regras de associação para analisar hábitos de compra de clientes e encontrar associações entre os diferentes itens que os clientes colocam em sua "cesta de compra". É uma técnica matemática, frequentemente usada por profissionais de marketing, para revelar afinidades entre produtos individuais ou grupo de produtos (HAN & KAMBER 2001). A entrada de dados para o processo é uma lista de transações contendo o conjunto de produtos comprados pelos clientes. Tal lista é representada por uma matriz binária, também conhecida como matriz de co-ocorrência. Nela os dados são dispostos de forma que a primeira coluna é referente às compras dos clientes (transação) e as demais colunas são os produtos comprados, conforme a figura

Figura 2 - Matriz de co-ocorrência adaptado de (SANTOS, 2004)

Transação	arroz	carne	leite	ovo	feijão
001	1	1	0	0	1
002	0	0	1	0	1
003	1	0	1	1	1
004	1	1	1	0	1
005	1	0	1	1	1
006	1	1	1	1	1
...	...	...	...	...	...

seguinte.

A presença de um produto na compra é representada na matriz pelo valor lógico 1 (um) e, 0 (zero) para identificar ausência do produto naquela compra. Percebe-se então, que a matriz de co-ocorrência utilizada nesse processo, em sua forma binária, é incapaz de representar as transações de compras de forma precisa, pois, supondo que um cliente comprou 5 unidades de um determinado produto, a representação precisa na matriz seria 5 e não 1 como é representado, ou seja, a matriz de co-ocorrência despreza a intensidade com que os produtos foram comprados. Esse importante fator foi estudado por Santos (2004) e resultou na criação de um método baseado em lógica difusa que trata desta imprecisão. Em seguida são apresentados os algoritmos utilizados na técnica de associação de itens e o algoritmo difuso proposto por Santos (2004).

Algoritmos para Mineração de Regras de Associação

Um dos mais tradicionais algoritmos de mineração utilizando a estratégia de itens frequentes é o Apriori. Diversas variações deste algoritmo, envolvendo o uso de técnicas de hash, redução de transações, particionamento e segmentação podem ser encontradas na literatura (CAMILO & SILVA, 2009). Devido a grande similaridade dos algoritmos que tratam deste assunto, nesta pesquisa foi destacado dois deles que são bastante considerados e possuem abordagens diferentes.

Apriori

O algoritmo Apriori segundo Romão (1999), De Vasconcelos & De Carvalho (2004), Semaan et. al. (2006) e outros, é um dos algoritmos mais utilizados para descobrir regras de associação. O algoritmo Apriori foi proposto por Agrawal et. al. (1993). Na pesquisa, os autores propuseram um algoritmo eficiente, tanto em termos de memória quanto de processamento, que descobre regras úteis contendo as relações de itens mais significativas.

Fracalanza (2009) traz uma definição direta do

processo realizado pelo algoritmo Apriori e seus derivados, dizendo que o método consiste em encontrar todas as regras de associação que possuam suporte e confiança maiores ou iguais a um suporte mínimo (SupMin) e uma confiança mínima (ConfMin), especificados pelo usuário. O problema de descobrir todas as regras de associação, tal como formulado por Agrawal et al. (1993), pode ser decomposto em duas etapas: i) encontrar os conjunto de itens frequentes (aqueles com suporte ≥ SupMin) e ii) gerar as regras a partir dos conjuntos de itens frequentes (com confiança ≥ ConfMin) (ROMÃO et. al., 1999).

Segundo Semaan et. al. (2006), o Apriori foi o primeiro algoritmo a tratar efetivamente o problema de extração de regras de associação. Goldschmidt et.al.

(2015) e Semaan et. al. (2006), citam diversos outros algoritmos que foram inspirados no Apriori, tais como, GSP, DHP, Partition, DIC, Eclat, MaxEclat, Clique, MaxClique, dentre outros, (LIU & WANG, 2007), (AGRAWAL et.al., 1994), que de alguma forma apresentam aperfeiçoamentos em relação ao algoritmo Apriori. Apesar dos diferentes algoritmos fornecerem tempos de execução diferentes para uma mesma base de dados, as ideias principais do algoritmo Apriori ainda são utilizadas e alcançam as mesmas regras. (SEMAAN et. al., 2006)

FP-Growth

Segundo Han et. al. (2000), um dos pontos que enfraquece o potencial do algoritmo Apriori, está na geração e teste do conjunto de candidatos. Esse problema foi tratado ao introduzir uma estrutura de dados nova e compacta, chamada árvore de padrões frequentes, ou FP-Tree. Baseado nessa nova estrutura foi proposto por Han et. al. (2000), o algoritmo FP-growth, que utiliza uma abordagem diferente do Apriori. Segundo testes realizados por Han et. al. (2000), dentre outros pesquisadores, tais como, Borgelt (2005), Gyorodi et. al. (2004) e Nandi et. al. (2015), o algoritmo FP-growth obteve maior eficiência, em termos de memória e de tempo de execução, do que o algoritmo Apriori.

Nandi et al. (2015) explica que os algoritmos tradicionais de associação adotam uma abordagem igual ou semelhante a do algoritmo Apriori, que se baseia na seguinte regra: se qualquer padrão de comprimento k não é frequente na base de dados, seu comprimento (k + 1) não será frequente. A ideia é, por meio de um processo iterativo, gerar o conjunto de padrões de candidatos de comprimento (k +1) a partir do conjunto de padrões de frequência de comprimento k (para k ≥ 1), e verificar suas frequências de ocorrência na base de dados. No entanto, a geração de conjunto candidato é ainda dispendiosa especialmente quando há um grande número de padrões e/ou estes são longos, ou seja, padrões formados por um número expressivo de itens. Também é dispendioso percorrer repetidas vezes a base de dados para verificar e testar todos os conjuntos de candidatos e seus padrões correspondentes (HAN et al., 2001).

Dessa forma o FP-Growth não gera conjuntos de candidatos iterativamente. O algoritmo codifica o conjunto de dados em uma estrutura de dados compacta em forma de árvore chamada FP-tree e extrai os conjuntos de itens frequentes diretamente desta estrutura (NANDI et. al, 2015).

Algoritmo baseado em lógica difusa de Santos (2004)

Proposto por Santos (2004), o algoritmo difuso consiste na utilização dos conceitos e abordagens da lógica difusa para tratar a imprecisão encontrada na matriz de co-ocorrência, que é de suma importância no processo de Market Basket Analysis (MBA). Segundo Santos, a razão de usar os conceitos da lógica difusa nesse processo deu-se pela capacidade dela poder converter valores numéricos em descritores difusos e modelar a imprecisão dos resultados, pois os resultados oferecidos pelos métodos tradicionais não considera a força de atração/repulsão entre os produtos, e com aplicação da lógica difusa é possível determinar essas relações em graus definidos por conjuntos difusos: baixo, médio e alto; sendo possível apontar não apenas os produtos que tem forte atração, mas também os que se repelem.

Em síntese, algoritmo difuso segue algumas etapas conforme a figura 3. O pré-processamento consiste em receber uma lista de transações e gerar a matriz de co-ocorrência a partir dela. Após isto, são geradas as associações

possíveis e calculadas as frequências individuais dos produtos antecedentes e consequentes (FR), além das frequências esperadas (FRE) e obtidas (FRO) de cada associação. Os valores calculados dessas frequências são as variáveis de entrada para o algoritmo. Estas, por sua vez, passam por um processo de fuzzificação onde recebem um valor difuso para cada um dos estados: baixo, médio e alto. Santos (2004) em seu trabalho definiu e testou várias combinações de funções de pertinência. A figura a seguir mostra uma das combinações de funções utilizadas para este processo.

Após a etapa de fuzzificação, é feito um processamento com base num conjunto de regras que foram e obtidas por Santos (2004) através de heurística, confor-

ISSN 2238-5630
Brasília-DF, v. 8, n. 1, janeiro-junho de 2019

Figura 3 - Etapas do Algoritmo Difuso proposto por Santos (2004)

Figura 4 - Combinação das funções L, Pi e Gama, usadas para a fuzzificação

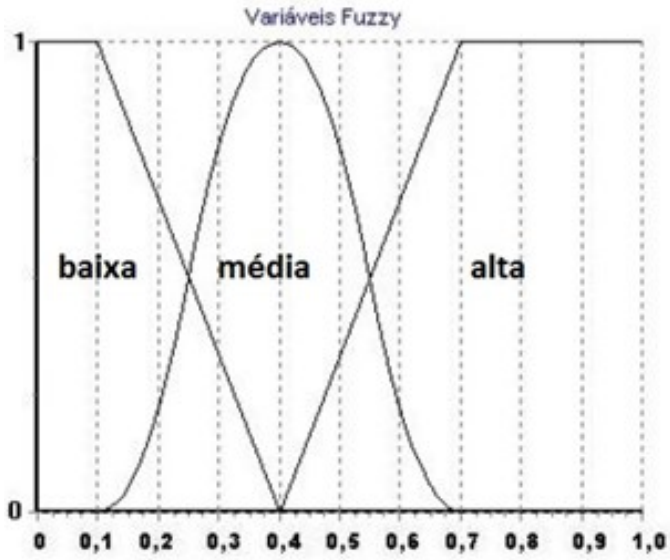


Tabela 1 – Regras de inferência do algoritmo difuso de Santos (2004).

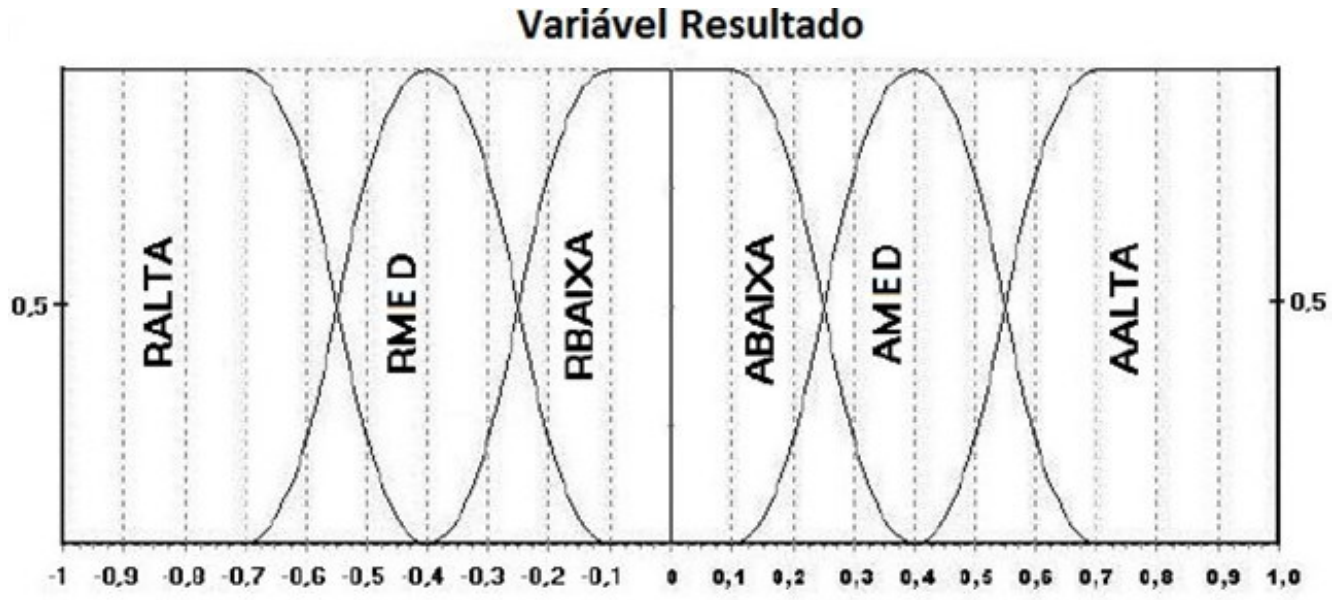
Regra	"Se FR A é"	"e FRE AB é"	"e FRO AB é"	"então RESULTADO é"
1	*Baixa*	*Baixa*	*Baixa*	*Atração média*
2	*Baixa*	*Média*	*Baixa*	*Repulsão alta*
3	*Baixa*	*Média*	*Média*	*Atração média*
4	*Média*	*Baixa*	*Baixa*	*Repulsão baixa*
5	*Média*	*Baixa*	*Média*	*Atração alta*
6	*Média*	*Média*	*Baixa*	*Repulsão alta*
7	*Média*	*Média*	*Média*	*Atração baixa*
8	*Média*	*Alta*	*Baixa*	*Repulsão alta*
9	*Média*	*Alta*	*Média*	*Repulsão alta*
10	*Alta*	*Baixa*	*Baixa*	*Atração baixa*
11	*Alta*	*Baixa*	*Média*	*Atração baixa*
12	*Alta*	*Média*	*Baixa*	*Repulsão média*
13	*Alta*	*Média*	*Média*	*Repulsão baixa*
14	*Alta*	*Média*	*Alta*	*Atração média*
15	*Alta*	*Alta*	*Baixa*	*Repulsão alta*
16	*Alta*	*Alta*	*Média*	*Repulsão média*
17	*Alta*	*Alta*	*Alta*	*Atração baixa*

ISSN 2238-5630
Brasília-DF, v. 8, n. 1, janeiro-junho de 2019

me Tabela 1.

O resultado das regras acima é dado com base num conjunto difuso com os seguintes termos linguísticos: AALTA (Atração Alta), ABAIXA (Atração Baixa), AMEDIA (Atração Média), RALTA (Repulsão Alta), RBAIXA (Repulsão Baixa) e RMEDIA (Repulsão Média) conforme mostrado na figura 5. Dessa forma o algoritmo é capaz de classificar as relações entre os produtos de uma forma mais precisa, pois identifica os produtos que tem forte relação (atração) como também os de fraca relação (repulsão), mostrando, em termos qualitativos, o quanto um item se

Figura 5 – Conjunto difuso para a variável Resultado



atrai ou se repele por outro.

Metodologia

A metodologia utilizada para o desenvolvimento desta pesquisa foi a pesquisa exploratória. Inicialmente, foi feito um estudo a respeito do algoritmo baseado em lógica difusa para Market Basket Analysis, em seguida foram levantados e estudados os métodos tradicionais para esse mesmo fim. De posse desses conhecimentos, foi feito um estudo comparativo entre o algoritmo baseado na lógica difusa e os tradicionais.

A comparação entre o algoritmo difuso e os tradicionais utilizados no processo de Market Basket Analysis, se deu pela análise dos resultados obtidos de cada um deles. Devido a maior parte dos algoritmos tradicionais serem baseados no Apriori, os resultados apresentados por eles, são iguais. Dessa forma, fez-se a comparação, em especial, com os dois algoritmos mais utilizados: o Apriori e o FP- Growth. Outro motivo foi pelo fato de que ambos algoritmos apresentam formas bastante diferentes de geração de regras de associação. De forma indireta, foram envolvidos na comparação os seguintes algoritmos: AIS, SETM, GSP, DHP, Partition, DIC, Eclat, MaxEclat, Clique, MaxClique e AprioriTID.

Para a execução dos algoritmos foram utilizados os seguintes instrumentos que possibilitaram a investigação dos resultados: Software R; Software Weka e Netbeans IDE. A base de dados usada para os testes foi obtida através do pacote ARULES do Software R. Essa base consiste num conjunto de 9835 transações de compras que simulam um histórico de vendas de um comércio.

Dessa forma, levou-se em consideração o nível de precisão dos algoritmos, isto é, os resultados da classificação das regras de associação geradas pelo método difuso e as regras de associação dos métodos tradicionais em diferentes taxas de suporte e confiança. Assim, os resultados foram comparados e apresentados de maneira a esclarecer se é viável, ou não, usar a abordagem difusa para Market Basket Analysis. Em seguida é apresentado e discutido todo o procedimento aplicado e os resultados obtidos durante os testes.

Testes Realizados

Para a realização dos testes utilizou-se algoritmos já implementados em alguns softwares. O algoritmo Apriori, por exemplo, foi executado no software R Studio. O algoritmo FP-Growth, no software Weka. Já o algoritmo difuso de Santos (2004), foi desenvolvido em

ISSN 2238-5630
Brasília-DF, v. 8, n. 1, janeiro-junho de 2019

linguagem Java através da IDE NetBeans 8.2. Assim foi possível obter os resultados para serem comparados. Em seguida serão apresentados alguns pontos relevantes da base de dados trabalhada e os resultados obtidos durante os testes.

Explorando a base de dados

Para a realização dos testes foi utilizada uma base de dados bastante conhecida no software R, chamada groceries. A base está disponível no pacote ARULES, que é uma infraestrutura de manipulação de grande quantidade de dados criado por Hahsler et. al. (2005). A base de dados consiste num conjunto de transações que representam diferentes compras de clientes. Em números, a base contém 9.835 transações e 169 produtos (itens) distintos.

A quantidade de células preenchidas, ou seja, a quantidade de 1s (uns) na matriz de co-ocorrência gerada, é dita densidade da base. Na base de dados groceries, o valor dessa densidade é de 0.0261. Esse valor obtido, aparentemente baixo, é devido muitas transações possuírem um conjunto pequeno de produtos. Também, percebe-se que o maior valor da frequência relativa dentro da base é de 25%. Outra medida bastante importante é o maior valor de suporte válido na base. Neste caso foi de 0.074, ou 7,4%. Este parâmetro diz respeito ao máximo valor que pode ser fornecido ao algoritmo a fim de serem apresentadas as primeiras regras, ou seja, SupMin acima de 7,4%, nesta base, não serão obtidas regras de associação. O conhecimento deste valor permitiu estabelecer intervalos de suporte entre 0,1% à 7,4% para os testes deste trabalho.

O algoritmo difuso permite ajustar o intervalo das funções de pertinência utilizadas na fuzzificação. Nos primeiros testes, o valor do intervalo era calculado a partir do valor de maior frequência encontrada na base de dados. Dessa forma

chegou-se a bons resultados porém em alguns testes cuja confiança era muito baixa, houve uma discrepância da taxa de acerto do algoritmo difuso em relação aos demais algoritmos. Por meio desta análise, foi estabelecido um novo intervalo, [0,1], chegando assim a ótimos resultados.

Resultados Obtidos

A tabela 2 mostra os testes realizados com suas respectivas medidas de suporte e confiança e a quantidade

Tabela 2 – Testes realizados com seus respectivos parâmetros

Teste	*Sup Min*	*Sup Max*	*Conf Min*	*Conf Max*	Quantidade de regras: Apriori	Quantidade de regras: FP-Growth	Quantidade de regras: Eclat e outros
T01	0.030	0.074	0.20	0.45	25	25	**25**
T02	0.010	0.056	0.45	0.58	31	31	**31**
T03	0.006	0.009	0.60	0.64	8	8	**8**
T04	0.003	0.004	0.75	0.89	6	6	**6**
T05	**0.001**	**0.002**	**0.90**	**1.00**	**129**	**130**	**129**

ISSN 2238-5630
Brasília-DF, v. 8, n. 1, janeiro-junho de 2019

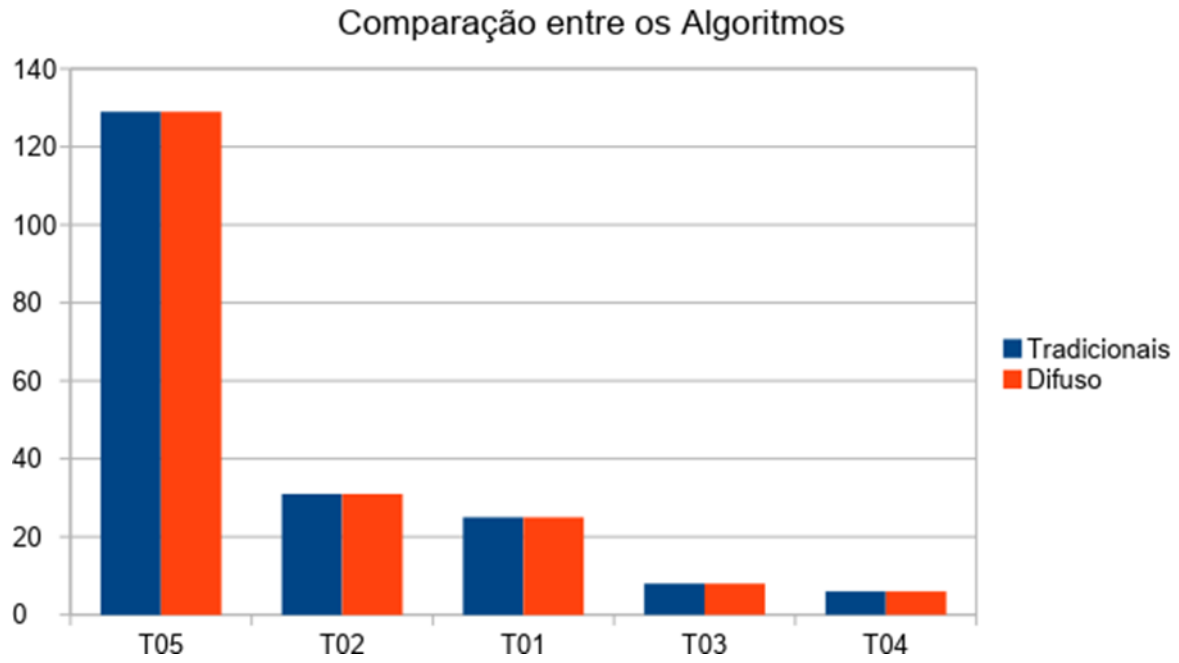
de regras obtidas pelos algoritmos com estes parâmetros.

Foram realizados vários testes com diferentes valores de suporte e confiança. Dessa forma, através da análise dos resultados, fez-se uma determinação dos limites dos intervalos de confiança, que resultou em 5 (cinco) testes conforme a tabela 2, de forma que as regras encontradas, em todos os cinco testes, são distintas. Assim sendo, foram avaliadas 199 regras diferentes.

Mediante os testes realizados, como podem ser vistos na tabela 2, percebe- se em todos eles a mesma quantidade de regras geradas pelos algoritmos, que era já esperado. As regras obtidas foram comparadas com a classificação resultante do algoritmo difuso. Isto possibilitou saber uma relação de acertos entre os algoritmos, ou seja, descobrir em porcentagem a quantidade de regras, dentre as regras geradas pelos algoritmos, que foram classificadas como "atração" pelo método difuso. Dessa forma, foi possível determinar a eficiência do método difuso na classificação das melhores regras geradas pelos algoritmos testados.

O gráfico a seguir mostra a quantidade de regras obtidas pelos algoritmos tradicionais e quantas delas foram classificadas como sendo do tipo "atração" pelo algoritmo difuso.

Gráfico 1 – Relação de acertos



Com base no gráfico, percebe-se que todas as regras analisadas foram classificadas pelo algoritmo difuso como sendo, no mínimo, atração baixa. Isso leva a uma taxa de 100% de acerto do método difuso, provando assim, a eficiência do algoritmo na classificação das regras.

Diante dos resultados apresentados, o método difuso mostrou-se bastante eficiente. Além de o método determinar regras que possuem atração e repulsão, o mesmo também é capaz de dar um fator de precisão maior em relação às regras, pois não só informa que há, ou não, atração entre os itens, como também informa o quanto se atrai ou repele um conjunto de item pelo outro, nos seguintes graus qualitativos: baixo, médio e alto.

Conclusão

O desenvolvimento deste trabalho possibilitou a realização de um estudo comparativo entre um algoritmo difuso proposto por Santos (2004) e alguns dos principais métodos e algoritmos tradicionais utilizados no cálculo de atração/repulsão entre itens, que é base para Market Basket Analysis. Além disso, também foi possível uma análise dos resultados oferecidos por cada algoritmo tradicional em relação ao algoritmo difuso.

De modo geral, em sua grande maioria, os algoritmos utilizados para cálculo de relação entre itens são baseados no algoritmo Apriori e o que está repleto na literatura são otimizações em relação a esse. Outro algoritmo utilizado nesta pesquisa foi o FP-Growth, que tomou grande importância devido possuir abordagem distinta do apriori e ser eficiente nesse processo.

Ao fazer a comparação entre os algoritmos, analisou-se as respostas por eles geradas. Isso permitiu

fazer uma relação entre cada algoritmo tradicional e o baseado em lógica difusa, calculando uma taxa de acerto das regras geradas pelos algoritmos tradicionais e a classificação obtida pelo método difuso. Com isso os resultados obtidos foram analisados e apresentados conforme o Gráfico 1, e assim, possibilitou saber a viabilidade do algoritmo proposto por Santos (2004) em relação aos tradicionais.

O método difuso proposto por Santos (2004) mostrou ótimos resultados quando comparado aos demais algoritmos, além disso, mostrou ser viável no cálculo de atração e repulsão entre itens, e traz ao meio científico uma nova abordagem de determinar relações entre itens com base nos conceitos da lógica difusa.

Referências

AGRAWAL, Rakesh; IMIELIŃSKI, Tomasz; SWAMI, Arun. Mining association rules between sets of items in large databases. In: **Acm sigmod record**. ACM, 1993. p. 207-216.

AMO, Sandra De. Técnicas de mineração de dados. **Jornada de Atualização em Informática**, 2004.

BORGELT, Christian. An Implementation of the FP-growth Algorithm. In: **Proceedings of the 1st international workshop on open source data mining: frequent pattern mining implementations**. ACM, 2005. p. 1-5.

BUTTAR, Harveen; KAUR, Rajneet. Association Technique in Data Mining and Its Applications. **International Journal of Computer Trends and Technology (IJCTT)-volume4Issue4–April 2013**, 2013.

CAMILO, Cássio Oliveira; SILVA, João Carlos da. Mineração de dados: Conceitos, tarefas, métodos e ferramentas. **Universidade Federal de Goiás (UFC)**, p. 1-29, 2009.

CÔRTES, S. C.; PORCARO, R. M.; LIFSCHITZ, S. **Mineração de dados – Funcionalidades, técnicas e abordagens**. PUC-Rio Inf. MCC10/02, maio 2002.

DE VASCONCELOS, Lívia Maria Rocha; DE CARVALHO, Cedric Luiz. Aplicação de regras de associação para mineração de dados na web. **Instituto de Informática da Universidade Federal de Goiás**, 2004.

FAYYAD, Usama; PIATETSKY-SHAPIRO, Gregory; SMYTH, Padhraic. From data mining to knowledge discovery in databases. **AI magazine**, v. 17, n. 3, p. 37, 1996.

FRACALANZA, Lívia Fonseca. **Mineração de Dados voltada para Recomendação no Âmbito de Marketing de Relacionamento**. 2009. Dissertação de Mestrado. PUC-Rio.

GOLDSCHMIDT, Ronaldo; BEZERRA, Eduardo; PASSOS, E. Data mining: conceitos, técnicas, algoritmos, orientações e aplicações. **Rio de Janeiro-RJ: Elsevier**, p. 56-60, 2015.

GYORODI, Cornelia; GYORODI, Robert; HOLBAN, Stefan. A comparative study of association rules mining algorithms. In: **Hungarian Joint Symposium on Applied Computational Intelligence, Oradea**. 2004.

HAHSLER, Michael; GRUEN, Bettina; HORNIK, Kurt. **arules - A Computational Environment for Mining Association Rules and Frequent Item Sets**. Journal of Statistical Software 14/15. 2005.

HAN, Jiawei; PEI, Jian; YIN, Yiwen. Mining frequent patterns without candidate generation. In: **ACM sigmod record**. ACM, 2000. p. 1-12.

HAN, Jiawei & KAMBER, Micheline. **Mineração de dados: Concepts and Techniques**. USA: Morgan Kaufmann, 2001.

LIU, Hanbing; WANG, Baisheng. An association rule mining algorithm based on a Boolean matrix. **Data Science Journal**, v. 6, p. S559-S565, 2007.

LIBRELOTTO, Solange Rubert; MOZZAQUATRO, Patricia Mariotto. Análise dos algoritmos de mineração J48 e Apriori aplicados na detecção de indicadores da qualidade de vida e saúde. **Revista Interdisciplinar de Ensino, Pesquisa e Extensão**, v. 1, n. 1, 2014.

NANDI, Júlio César Borba; PEREIRA, Ruano Marques; FELIPPE, Gabriel. O Algoritmo de Associação Frequent Pattern-Growth na Shell Orion Data Mining Engine. **Anais SULCOMP**, v. 7, 2015.

ROMÃO, Wesley et al. Extração de regras de associação em C&T: O algoritmo Apriori. **XIX Encontro Nacional em Engenharia de Produção**, v. 34, p. 37-39, 1999.

SANTOS, J. G. dos. **Uso de conjuntos difusos e lógica difusa para cálculo de atração e repulsão: uma aplicação em Market Basket Analysis.** 2004. 113f.. Tese (Doutorado em Ciência da Computação) – Universidade Federal de Santa Catarina –UFSC, Florianópolis. Dez. 2004.

SEMAAN, Gustavo Silva; GRAÇA, A. A.; DIAS, Carlos Rodrigo. Extração de Associações em Bases de Dados de Varejo. **XXXVIII Simpósio Brasileiro de Pesquisa Operacional (SBPO), Goiânia**, 2006.

TUDOR, Irina. Association rule mining as a data mining technique. **Seria Matematică-Informatică–Fizică-Buletin**, v. 1, p. 49-56, 2008.

CURRÍCULOS

* possui graduação em Licenciatura Plena Em Matemática pela Universidade Federal de Mato Grosso (1994) e Tecnologia em Gestão de TI pela Universidade Católica de Brasília, especialização em matemática computacional pela Universidade Federal de Mato Grosso (1996) e em engenharia de requisitos e processo de negócio pela Universidade Federal do Rio Grande do Sul (2012), mestrado em Ciências da Computação (área de Inteligência Artificial) pela Universidade Federal de Santa Catarina (2001) e doutorado em Ciência da Computação (área de Inteligência Artificial) pela Universidade Federal de Santa Catarina (2004). Tem experiência na área de Ciência da Computação, com ênfase em Arquitetura de Sistemas de Computação, atuando principalmente nos seguintes temas: sis-

ISSN 2238-5630
Brasília-DF, v. 8, n. 1, janeiro-junho de 2019

temas especialistas, inteligência artificial, análise estatística, data mining, banco de dados, ITIL, Análise Orientada a Objetos, Programação Orientada a Objetos, Linguagens de Programação (JAVA, C/C++, DELPHI, PHP) , SQL e Desenvolvimento JAVA WEB.

** Atualmente cursa Ciência da Computação no Instituto Federal de Brasília.