REVISTA eixo
revistaeixo.ifb.edu.br
ISSN 2238-5630
Brasília-DF, v. 8, n. 2, junho-dezembro de 2019

Autores | Authors

* José Gonçalo dos Santos
[jose.santos@ifb.edu.br]
**Wilton Lucas Pires Gomes

USO DE FUZZY SET E FUZZY LOGIC PARA TRATAR IMPRECISÃO NO CÁLCULO DE ATRAÇÃO E REPULSÃO: UM ESTUDO DE CASO EM MARKET BASKET ANALYSIS

USE OF FUZZY SET AND FUZZY LOGIC TO TREAT INACCURACY IN CALCULATION OF ATTRACTION AND REPULSION: A CASE STUDY IN MARKET BASKET ANALYSIS

Resumo: O estudo de afinidades entre produtos, conhecido popularmente como MBA (Market Basket Analysis), utiliza uma matriz binária para representar as transações de compras de clientes e os diferentes produtos comprados. Por ser binária, despreza a intensidade da associação dos eventos e a quantidade de produtos comprados tornando-a imprecisa. Dessa forma, esta pesquisa faz um estudo de viabilidade da utilização dos conceitos e abordagens da lógica difusa para tratar a imprecisão encontrada. Assim, foi possível fazer um algoritmo difuso capaz de realizar cálculos de afinidade entre produtos com mais precisão. Foram feitos vários testes e análises com diferentes combinações de funções de pertinência a fim de obter melhores resultados. A validação do método difuso deu-se pela comparação entre o cálculo do LIFT e o algoritmo Apriori. O método difuso mostrou ser eficiente em mais de 80% dos casos quando comparado com o LIFT e 100% com o Apriori. Com isso, esta pesquisa modelou a imprecisão contida na matriz e mostrou que é viável usar lógica difusa no processo de MBA.

Palavras chave: lógica fuzzy, market basket analysis, algoritmo.

***Abstract**: The interrelationship between products, popularly known as MBA (Market Basket Analysis), uses a binary matrix to represent customer purchasing transactions and different products purchased. Because it is binary, it despises the intensity of the association of events and the quantity of products purchased rendering it inaccurate. Thus, this research makes a feasibility study of the use of fuzzy logic concepts and approaches to address the imprecision found. Thus, it was possible to make a diffuse algorithm capable of performing affinity calculations between products with more precision. Several tests and analyzes were done with different combinations of membership functions to obtain better results. The validation of the diffuse method was done by comparing with the LIFT calculation and the Apriori algorithm. The diffuse method showed to be efficient in more than 80% of the cases when compared to the LIFT and 100% with the Apriori. With this, this research modeled the imprecision contained in the matrix and showed that it is feasible to use fuzzy logic in the MBA process.*

Keywords: fuzzy logic, market basket analysis, algorithm.

Recebido em: 11/10/2017
Aceito em: 05/02/2019

INTRODUÇÃO

Atualmente as organizações têm se mostrado extremamente eficientes em capturar, organizar e armazenar imensas quantidades de dados obtidos de suas operações diárias, porém a maioria ainda não usa adequadamente essa gigantesca porção de dados para transformá-la em conhecimentos que possam ser utilizados em suas próprias atividades, sejam elas comercias ou científicas (CÔRTES et al., 2002).

A necessidade de transformar grandes porções de dados em conhecimentos realmente úteis tem contribuído grandemente para o crescimento da área de pesquisa denominada DCBD (Descoberta de Conhecimento em Bases de Dados), que une conceitos de estatística, de inteligência artificial e de banco de dados (HAN & KAMBER, 2001 apud SANTOS, 2004). A descoberta de conhecimento é um processo prolongado constituído de várias etapas, e considera-se a etapa de mineração de dados como a principal desse processo. Isso se deve ao fato de ser exatamente nessa etapa, que as informações ocultas dentro das bases de dados passam a ser vistas como informações valorosas, chamadas de padrões.

Há muitos modelos e técnicas já introduzidos no meio científico que auxiliam no processo de extração de padrões. Na mineração de dados, por exemplo, são usadas técnicas como associação entre itens, classificação, agrupamento (clustering), padrões sequenciais (predição) e regressão que fazem uso de métodos como: árvores de decisões, teorema de Bayes, regras de decisão, redes neurais artificiais, algoritmos genéticos, lógica fuzzy, além de outros métodos usados.

A tecnologia de mineração de dados que tem tido grande destaque no mundo dos negócios, especialmente na área de vendas, é conhecida popularmente como Market Basket Analysis (MBA) ou, em português, análise da cesta de compras. Essa técnica envolve associação entre itens e trabalha com geração de regras de associação para identificar afinidades entre os produtos comprados pelos clientes. O MBA tem grande importância no setor comercial, pois pode auxiliar significativamente para melhorar as vendas e aumentar os lucros.

A entrada para o processo de MBA é um lista de transações de compras que são montadas em uma tabela, chamada de matriz de co-ocorrência, onde nas linhas são representadas as diferentes transações e nas colunas os produtos comprados. Um problema nesse caso é o desprezo da intensidade com que os produtos são comprados, pois a representação descrita na matriz de co-ocorrência é por meio de 0s e 1s, onde 0 (zero) representa a ausência do produto na compra e 1 (um) a presença do produto.

Muitos métodos e algoritmos têm sido estudados e desenvolvidos por diferentes pesquisadores a fim de tornar o processo de MBA mais eficiente. No entanto, dentre os métodos tradicionais não há um que modela a imprecisão contida na matriz de co-ocorrência, em relação ao desprezo da quantidade de produtos comprados. Os primeiros estudos utilizando a abordagem difusa para modelar essa imprecisão foram realizados por Santos (2004), que fez em sua pesquisa a transformação dos valores numéricos da matriz de co-ocorrência em descritores difusos (qualitativos) para poderem ser analisados através de regras.

O uso da abordagem difusa para alcançar maior precisão nos resultados é amplo na literatura, além disso, abrange áreas como avicultura (PONCIANO et al., 2011), mecânica (BILOBROVEC et al., 2004), meteorologia (VIEIRA et al., 2014), medicina (BRANDAO et al., 2013), engenharia (CHERRI et al., 2011). Em todas essas pesquisas os resultados foram satisfatórios.

A razão de usar os conceitos da lógica difusa nesse processo deu-se pela capacidade dela poder converter valores numéricos em descritores difusos conforme apresentado em (SANTOS, 2004) e modelar a imprecisão dos resultados, pois os resultados oferecidos pelos métodos tradicionais não consideram 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 têm forte atração, mas também os que se repelem.

Assim, o objetivo desta pesquisa é verificar a viabilidade do uso da lógica difusa no processo de afinidades entre produtos. Isso leva a responder a seguinte hipótese que norteia este trabalho: "No processo de descoberta de conhecimento em bases de dados é adequado utilizar a abordagem difusa para modelar a imprecisão contida na matriz de co-ocorrência?". Para responder essa questão, foi necessária uma pesquisa junto à literatura especializada, referentes à DCBD, à lógica difusa e à MBA. Dessa forma, foi possível criar um algoritmo de associação e classificação de produtos quanto à atração ou repulsão presente entre eles, utilizando a abordagem difusa. Com ele foram feitos testes e validações em cima de um conjunto de dados que

ISSN 2238-5630
Brasília-DF, v. 8, n. 2, junho-dezembro 2019

simulam um histórico de vendas de um comércio, e assim, foi possível saber a viabilidade do mesmo no processo de MBA.

Este artigo está estruturado da seguinte forma: nesta seção é apresentada a introdução contendo a descrição do problema e os objetivos da pesquisa. A seção 2 apresenta o referencial teórico dos principais temas trabalhados. A seção 3 exibe a metodologia utilizada. A seção 4 descreve o algoritmo difuso proposto. A seção 5 apresenta os testes realizados e os resultados obtidos e por fim, na seção 6, são apresentadas as considerações finais.

REFERENCIAL TEÓRICO

O pensamento humano é capaz modelar e lidar com as imprecisões de uma forma rápida e experiente (PACHECO et al., 1991). No entanto, tornar uma máquina capaz de modelar imprecisões não se trata de uma tarefa simples.

O tratamento de imprecisão é a forma de corrigir os problemas de situações incertas ou vagas, modelando as variáveis trabalhadas de forma a atender mais precisamente a necessidade do problema.

A literatura propõe muitos modelos e técnicas matemáticas que podem ser utilizadas para tratar as imprecisões. Nos últimos anos, o que se tem visto com frequência, é o uso dos conceitos da lógica difusa na solução de situações imprecisas, pois ela é um ferramental de excelência para esses fins (RIGNEL, 2011).

LÓGICA FUZZY

A ideia da lógica difusa ou lógica fuzzy está completamente baseada na teoria dos conjuntos fuzzy, que objetiva representar e manipular dados aproximados e vagos (RODRIGUES; DIMURO, 2010). Com essa potencialidade, é possível fazer com que o processamento computacional chegue mais próximo do pensamento humano e consequentemente a modelagem das situações imprecisas por máquinas torna-se mais eficiente.

A teoria dos conjuntos difusos ficou mundialmente conhecida a partir dos estudos de Lofti Asker Zadeh[1], mais precisamente em 1965. A partir de então, a sua aplicabilidade se estendeu por todas as áreas da ciência e tende crescer ainda mais devido à eficiência dos resultados encontrados em diversas pesquisas.

Zadeh desenvolveu a teoria de conjuntos difusos que permite trabalhar de acordo com o raciocínio humano, que é intrinsecamente impreciso e vago. A lógica fuzzy é uma técnica muito útil na solução de problema com extensa aplicabilidade. É usada atualmente nas áreas de negócios, sistemas de controle, eletrônica e engenharia de tráfego, meteorologia, entre outras aplicações (VIEIRA et al. , 2014).

Uma das diferenças entre a lógica clássica (booleana) e a lógica fuzzy é que a representação dos conjuntos na lógica fuzzy não possui limites bem definidos, conforme apontado por Santos (2004). Isso torna possível um elemento pertencer parcialmente a outro conjunto ou pertencer a dois conjuntos ao mesmo tempo, uma vez que a lógica fuzzy não considera apenas dois valores lógicos para representar a pertinência em um conjunto, como na lógica booleana, mas sim, um grau contido no intervalo [0,1]. Assim, com essa possibilidade, segundo Gomide e Gudwin (1994), pode-se obter um desempenho mais estável e preciso quando considerado processos mais complexos.

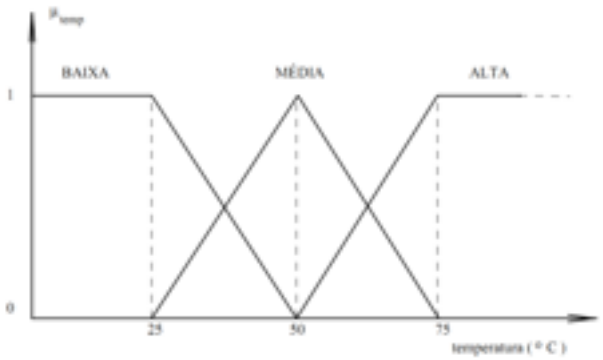
Para expressar os conceitos na lógica fuzzy é muito comum o uso de semânticas categóricas que servem para representar resultados qualitativos, como, "mais ou menos", "talvez", "muito", "pouco" "alto", "baixo", "médio". Estes conceitos são capturados pelas variáveis linguísticas. Gomide e Gudwin (1994) definem as variáveis linguísticas como um conjunto fuzzy definido sobre um universo de discurso comum. No mesmo sentido, Tanscheit et al. (1995) complementa a definição de Gomide dizendo que as variáveis linguísticas possuem "valores" dos quais esses constituem nomes de conjuntos fuzzy. Tanscheit et al. (1995) também cita que "a principal função das variáveis linguísticas é fornecer uma maneira sistemática para uma caracterização aproximada de fenômenos complexos ou mal definidos".

A Figura 1 traz a noção concreta da definição de variável linguística e função de pertinência.

A variável 'Temperatura' assume os valores: baixa, média e alta. No cálculo de atração e repulsão entre itens, por exemplo, uma variável linguística 'atração' ou 'repulsão', poderia assumir os estados: fraca, média e forte, de forma a identificar qualitativamente as relações presentes entre os itens.

1 Zadeh, L.A. (1965). "Fuzzy Sets". Information and Control, V. 8: 338-353.

Figura 1- Variável Linguística Temperatura (GOMIDE & GUDWIN, 1994).



Estes estados ou valores são descritos por meio de conjuntos fuzzy, representados por funções de pertinência. As funções de pertinência definem os limites de um determinado conjunto na lógica fuzzy. Dentre as funções de pertinência mais utilizadas, pode-se destacar: Função Triangular, Função Trapezoidal, Função Gama, Função L, Função Gaussiana ou Pi, Função Z, Função Sino e Função Sigmoidal.

Muitas operações de conjuntos como união e intersecção, que são feitas na lógica booleana, também podem ser feitas na lógica fuzzy, além de serem válidas muitas propriedades da lógica clássica. A lógica fuzzy, como já mencionada, é aplicada em diferentes áreas e a forma de aplicação da mesma em sistemas e modelos que utilizam essa abordagem geralmente segue um processo sequencial conforme mostra a Figura 2.

De acordo com a Figura 2, uma aplicação da lógica fuzzy evolve três etapas fundamentais. A primeira delas se refere ao processo de fuzzyficação, onde as informações são convertidas em valores difusos para então ocorrer a formulação e execução de uma estratégia de controle. Na segunda etapa os dados fuzzyficados são passados por um sistema de regras, onde se efetua a inferência sobre o conjunto de regras obtendo os valores dos termos das variáveis de saída. Por último, uma vez obtidas as variáveis linguísticas de saída pode-se aplicar a defuzzificação, que consiste em converter os dados

Figura 2- Etapas do processo difuso

ISSN 2238-5630
Brasília-DF, v. 8, n. 2, junho-dezembro 2019

nebulosos para valores numéricos precisos (BILOBROVEC, 2004).

DESCOBERTA DE CONHECIMENTO EM BASES DE DADOS (DCBD)

Segundo Fayyad apud Carvalho (2010), a descoberta de conhecimento em bases de dados pode ser definida como um processo não trivial de identificar padrões novos, válidos, potencialmente úteis e, principalmente, percebíveis em meio às observações presentes em grandes bases de dados.

Para alguns autores como Han e Kamber apud Santos (2004), a DCBD é um campo interdisciplinar formado pela junção de três grandes áreas, a estatística, a inteligência artificial e banco de dados, que atuam diretamente para que os padrões sejam alcançados. Dessa forma, não se trata de um simples processo, mas sim, de um procedimento organizado e prolongado cujo tempo é variado segundo os objetivos que se deseja alcançar.

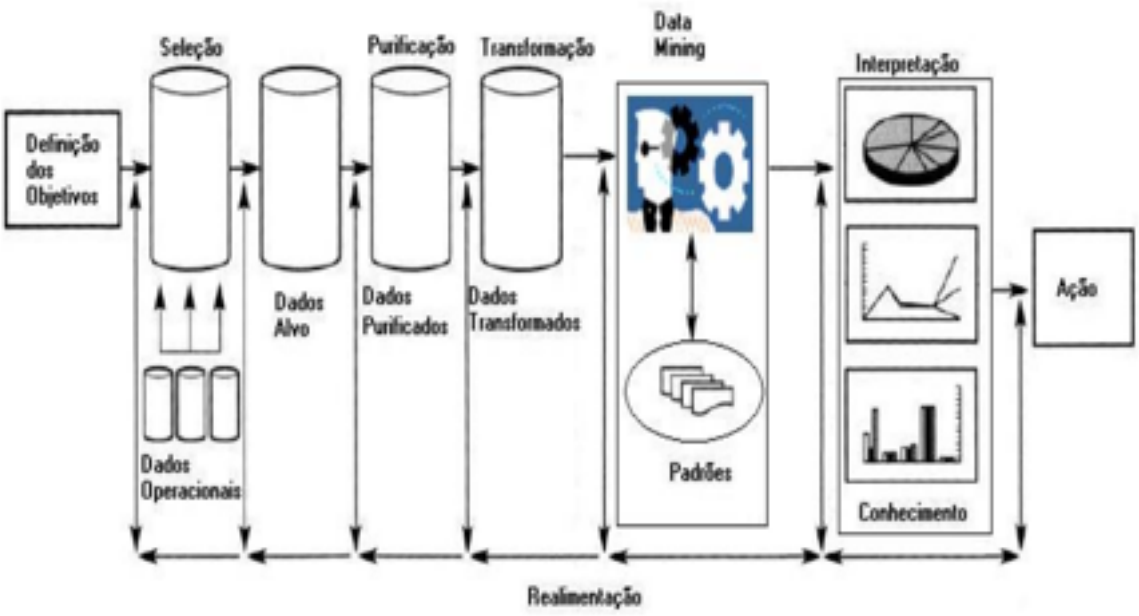
A Figura 3 mostra as fases que compõem o processo de DCBD, que foram extraídas de Fayyad et. al (1996 apud SANTOS, 2004).

Inicialmente é feito um processo de ETL (Extração, Transformação e Carga) dos dados deixando-os apropriados para a fase de mineração de dados, onde nessa fase, é feita realmente a descoberta dos conhecimentos ou padrões. Por fim, os dados são interpretados e validados para serem exibidos aos usuários (HAN & KAMBER 2001 apud SANTOS, 2004).

Considera-se a etapa de mineração de dados como a principal desse processo. É nessa etapa que são aplicadas técnicas e modelos a fim de extrair as informações ocultas e valorosas.

No processo de mineração de dados podem ser usadas várias técnicas que auxiliam na extração de padrões, tais como, classificação, estimação ou regressão, predição, agrupamento

Figura 3- Etapas do processo de DCBD adaptado de (FAYYAD et. al 1996 apud SANTOS, 2004)



(clustering), associação, onde em cada técnica podem ser aplicados diferentes métodos (CAMILO; SILVA, 2009).

A técnica de associação, segundo Berry e Linoff apud Côrtes *et al*. (2002), consiste em descobrir as regras de associação condicionadas a valores de atributos que ocorrem juntos em um conjunto de dados, ou seja, determina quais itens estão relacionados com outros itens. Essa técnica é bastante utilizada na área de Market. A técnica de classificação definida por Côrtes et al. (2002), "Consiste em examinar uma certa característica nos dados e atribuir uma classe previamente definida". Ambas as técnicas destacadas acima, são utilizadas no processo de MBA, que será discutido no tópico seguinte.

MARKET BASKET ANALYSIS (MBA)

O MBA é o processo que analisa hábitos de compra de clientes para 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 apud SANTOS, 2004)

De acordo com Santos (2004), o MBA é usado para determinar quais produtos são vendidos juntos, e a entrada é geralmente uma lista de transações de vendas.

As representações das cestas de compras dos clientes são feitas através da 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 (transações) e as demais colunas são os produtos comprados, conforme a Figura 4.

A matriz de co-ocorrência utilizada para o cálculo das medidas de associação em sua forma binária despreza a força de atração ou repulsão entre eventos. Esta força poderia ser forte, moderada ou fraca, caracterizando-se como uma variável linguística sob a presença da imprecisão quanto à ocorrência ou não ocorrência conjunta de eventos (SANTOS, 2004).

METODOLOGIA

Este trabalho consiste numa pesquisa aplicada na área de marketing e utiliza a forma de abordagem quantitativa para expressar os resultados. Refere-se também a uma pesquisa exploratória com o objetivo de averiguar os conceitos relevantes a essa área e entender os procedimentos utilizados em MBA (Market Basket Analysis). Para a realização deste trabalho considerou artigos publicados em periódicos e anais de conferência que envolvia a definição, conceitos e aplicabilidade da lógica fuzzy para modelar situações de imprecisão. Os artigos provinham de diferentes bases, dos quais foram buscados através da plataforma Google Acadêmico e SciELO - Scientific Electronic Library Online. Em números, foram 15 artigos publicados nas 3 últimas décadas.

Os artefatos utilizados possibilitaram: i) a identificação dos métodos mais usados em MBA; ii) os modelos de regras usados na lógica difusa; iii) a compreensão de como tratar as variáveis de entrada; iv) noções de adequação de limites dos

Figura 4- 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
...	...	...	...	...	...

ISSN 2238-5630
Brasília-DF, v. 8, n. 2, junho-dezembro 2019

intervalos das funções de pertinência; e v) a criação de um algoritmo difuso para determinar atração/repulsão entre itens.

Em síntese, o algoritmo difuso proposto se resume em 4 etapas: ETL (Extração, Transformação e Carga), Fuzzificação, Análise por meio de regras de inferência e Classificação. Os procedimentos ETL visam deixar os dados adequados para a etapa de mineração de dados. É nesta etapa que é gerada a matriz de co-ocorrência no qual os dados passam de um formato conforme a Tabela 1 para a forma matriz binária mostrada na Tabela 2. Dada a matriz de co-ocorrência, calcula-se as frequências: FRA (frequência relativa de A), FREAB (frequência relativa esperada de A e B), FROAB (frequência relativa obtida de A e B), onde A e B são os produtos antecedentes e consequentes das regras. Assim, inicia-se a etapa de fuzzificação onde os valores difusos de cada frequência (FRA, FREAB, FROAB) são obtidos das funções de pertinência nos graus baixo, médio e alto. Em seguida os resultados são analisados através de um conjunto de regras que foram obtidas através de heurística e são classificados conforme a regra de inferência melhor aplicada.

O procedimento foi implementado na linguagem JAVA através do Netbeans IDE 8.0. A escolha foi devido a facilidade dos autores em trabalhar com essa linguagem, contudo, a linguagem não interfere nos objetivos da pesquisa. A implementação resultou em um sistema que foi chamado de CARI (Cálculo de Atração e Repulsão entre Itens) que serviu para executar todas as etapas do algoritmo difuso e determinar o grau de afinidade entre os itens. Além disso, com o CARI, foi feita a validação do algoritmo difuso comparando com o valor do LIFT de Groth apud Santos (2004), calculado conforme a Figura 8 e o algoritmo Apriori que é um dos mais utilizados hoje no processo de MBA. O LIFT resulta num valor que está entre -1 e 1. Dessa forma, quanto mais próximo for o valor de -1 ocorre uma repulsão, caso contrário, mais próximo de 1, uma atração. Se o valor resultante for exatamente 0, significa que os itens são independentes. Esse fenômeno explica a escolha do LIFT para a validação do método difuso. Assim, foi obtida uma taxa de equidade entre a classificação do LIFT e do método difuso. Para o Apriori, comparou as regras geradas por ele com a classificação de no mínimo "Atração Baixa" dada pelo método difuso. Isto possibilitou saber uma relação de similaridade entre os dois métodos. Para a comparação com o algoritmo Apriori foram realizados vários testes com diferentes taxas de SupMin (suporte mínimo) e ConfMin (Confiança Mínima) afim de explorar melhor os dados usados.

O algoritmo Apriori foi executado pelo Software R Studio e o difuso pelo Netbeans IDE 8.0.

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). Esta consiste num conjunto de transações que representam diferentes compras de clientes contendo 9.835 transações e 169 produtos (itens) distintos.

Foram feitos vários testes e análises com dezesseis combinações de funções de pertinência a fim de encontrar aquela com melhores resultados. Diante destes procedimentos foi possível saber a viabilidade do uso da lógica difusa no processo de MBA. Em seguida é detalhado o algoritmo difuso e os resultados obtidos durante os testes.

ALGORITMO DIFUSO

Neste tópico serão mostrados os passos do algoritmo difuso proposto para cálculo de atração/repulsão entre itens. O algoritmo segue as etapas fundamentais do processo de DCBD e de aplicação de lógica difusa conforme a Figura 3 e Figura 2 respectivamente. Logo, as etapas serão tratadas e discutidas em subtópicos seguintes.

SELEÇÃO, PURIFICAÇÃO E TRANSFORMAÇÃO DOS DADOS

A primeira etapa do algoritmo é fazer um tratamento do conjunto de dados usados, ou seja, é a realização dos procedimentos ETL (Extração, Transformação e Carga) dos dados. Essa transformação visa deixar os dados adequados para a etapa de mineração de dados. Nessa fase também é gerada a matriz de co-ocorrência, que será usada para fazer todos os cálculos de frequências dos itens.

A transformação dos dados significou passar de um formato conforme a Tabela 1 para a forma da matriz de co-ocorrência, mostrada na Tabela 2 abaixo.

MINERAÇÃO DE DADOS

Na etapa de mineração de dados foram aplicados os conceitos da lógica difusa. Os primeiros cálculos realizados com

a matriz de co-ocorrência foram de probabilidades e frequências, pois assim, pode-se saber como estão distribuídos os produtos na matriz em relação as suas quantidades. Então, os produtos foram combinados em pares, gerando as possíveis associações de primeiro nível, isso é, associação do tipo A→B, onde A é o produto antecedente e B o produto consequente.

Para cada uma das associações geradas, foram calculadas as frequências individuais dos produtos antecedentes e consequentes, além das frequências esperadas e obtidas de cada associação conforme as fórmulas de cálculo abaixo.

MINERAÇÃO DE DADOS

Na etapa de mineração de dados foram aplicados os conceitos da lógica difusa. Os primeiros cálculos realizados com a matriz de co-ocorrência foram de probabilidades e frequências, pois assim, pode-se saber como estão distribuídos os produtos na matriz em relação as suas quantidades. Então, os produtos foram combinados em pares, gerando as possíveis associações de primeiro nível, isso é, associação do tipo A→B, onde A é o produto antecedente e B o produto consequente.

Para cada uma das associações geradas, foram calculadas as frequências individuais dos produtos antecedentes e consequentes, além das frequências esperadas e obtidas de cada associação conforme as fórmulas de cálculo abaixo.

Os valores das frequências calculadas são passadas pelo processo de fuzzificação onde serão transformados em descritores difusos.

Tabela 1- Transações de compras de clientes

Transações	Produtos			
001	Leite	Maçã	Banana	Carne
002	Arroz	Batata	Feijão	Leite
003	Chocolate	Ovo	Batata	Carne
004	Cerveja	Arroz	Carne	Batata
005	Carne	Maçã	Arroz	Feijão
006	Leite	Ovo	Chocolate	Cerveja
...				...

Tabela 2- Matriz de co-ocorrência gerada a partir da tabela 1

Transação	Leite	Maçã	Banana	Carne	Arroz	Batata	Feijão	Chocolate	Ovo	Cerveja
001	1	1	1	1	0	0	0	0	0	0
002	1	0	0	0	1	1	1	0	0	0
003	0	0	0	1	0	1	0	1	1	0
004	0	0	0	1	1	1	0	0	0	1
005	0	1	0	1	1	0	1	0	0	0
006	1	0	0	0	0	0	0	1	1	1
...										...

ISSN 2238-5630
Brasília-DF, v. 8, n. 2, junho-dezembro 2019

Figura 5- Fórmula para cálculo das frequências.

$$FRA = freq\_rel(A) = \frac{\sum_{1}^{n} oA}{n}, \qquad FRB = freq\_rel(B) = \frac{\sum_{1}^{n} oB}{n},$$

$$FREAB = FRAxFRB, \qquad FROAB = freq\_relO(A^B) = \frac{\sum_{1}^{n} (oA^{\wedge} oB)}{n},$$

Onde n é o número de transações, FRA é a frequência relativa do produto A, FRB é a frequência relativa do produto B, FREAB é a frequência relativa esperada de A e B, FROAB é a frequência relativa obtida de A e B, o A é a ocorrência de A na tabela, o B é a ocorrência de B na tabela, (oA ^oB) é a ocorrência de A e B simultaneamente.

FUZZIFICAÇÃO

A *fuzzificação* se refere ao processo no qual os valores das frequências FRA, FREAB e FROAB de cada associação são passadas pelas funções de pertinência.

Foram utilizadas 8 funções de pertinência distintas e combinadas de 3 em 3, que resultou em 16 combinações possíveis. O motivo de serem combinadas de 3 em 3 é devido a quantidade de conjuntos difusos: baixo, médio, e alto. Dessa forma, as frequências FRA, FREAB e FROAB são fuzzificadas e recebem um valor difuso para cada um dos estados, baixo, médio e alto.

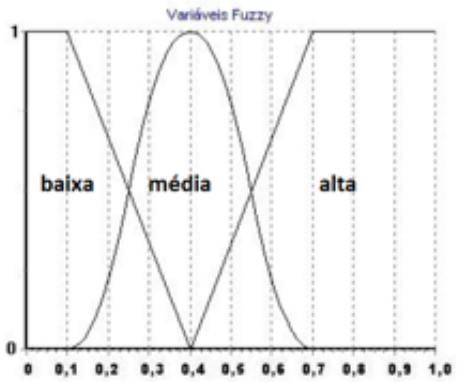
A figura a seguir mostra uma das combinações de funções utilizadas para este processo.

Para exemplificar, considere FRA = 0,5. Para esse valor seria resultado os seguintes valores difusos: FRAbaixa = 0,00 ; FRAmedia= 0,75 ; FRAalta = 0,30; e assim segue o mesmo critério para as demais frequências de cada associação.

REGRAS DE INFERÊNCIA

Os valores difusos passam por um conjunto de regras que foram obtidas através de heurística. Para a criação dessas regras, foram analisadas as diferentes combinações dos resultados das frequências, ou seja, uma frequência FRA, por exemplo, geraria 3 valores difusos. Como é necessário fazer a fuzzificação de 3 frequências para cada associação e cada frequência gera 3 valores difusos, o total de possibilidades é calculado por 33 , totalizando 27 possíveis resultados. Assim, seriam necessárias vinte e sete regras, porém, fazendo algumas análises, foram desprezadas aquelas cuja possibilidade de ocorrência é mínima ou nula em um am-

Figura 6- Combinação das funções de pertinencia L Pi e Gama utilizadas para fuzificação.



biente real. Desse modo, tomaram dezessete regras que classificam as associações conforme a Tabela 3.

CLASSIFICAÇÃO

O resultado apresentado nas 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 7.

Para classificar as associações utilizou-se o método "Max-Produto". O método consiste num produto entre as frequências difusas de cada regra individual.

Para exemplificar, considere a regra 1 (um) como a análise da vez. Para ela seriam multiplicados os valores difusos FRAbaixo X FREABbaixo X FROABbaixo. O resultado obtido seria comparado com os resultados das demais regras (2,3,4,...17) a fim de ser escolhido o maior deles. Assim, a regra que apresentar maior resultado será a regra vencedora.

Tabela 3- Regras usadas para classificação

Regra	"Se FRA é"	"e FREAB é"	"e FROAB 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. 2, junho-dezembro 2019

Dessa forma, a classificação será dada conforme o resultado da regra vencedora. No exemplo citado acima com a regra 1, caso ela atinja maior valor dentre as demais regras, seria atribuído uma classificação de atração média conforme a Tabela 3.

VALIDAÇÃO DO MÉTODO DIFUSO PROPOSTO

Para validar o método proposto fez-se uma comparação com o valor do LIFT de Groth apud Santos (2004). Essa comparação também foi implementada no sistema CARI e aparece como taxa de desempenho vista na Figura 9.

Além disso, fez-se uma comparação com um dos algoritmos mais usados hoje para determinar relação entre itens, a saber, o Apriori. Para este, foram usados diferentes valores de suporte mínimo e confiança mínima, uma vez que estes valores são definidos como parâmetros de entrada.

Dessa forma, a comparação tanto para o LIFT quanto para o Apriori resultou em um valor que mostra o quanto de desempenho o algoritmo difuso atingiu.

Figura 7- Representação dos conjuntos difusos para a variável resultado.

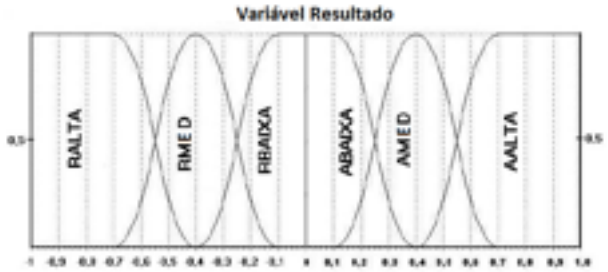


Figura 8- Fórmula para cálculo do lift.

$$lift(A \rightarrow B) = \frac{freq\_relo(A \wedge B) - freq\_rele(A \wedge B)}{freq\_rel(A)},$$

onde: $lift(A \rightarrow B)$ = Medida de atração entre os itens A e B,

$freq\_rele(A \wedge B)$ = Freqüência Relativa esperada de A e B,

$freq\_relo(A \wedge B)$ = Freqüência Relativa obtida de A e B,

$freq\_rel(A)$ = Freqüência Relativa de A.

SISTEMA CARI

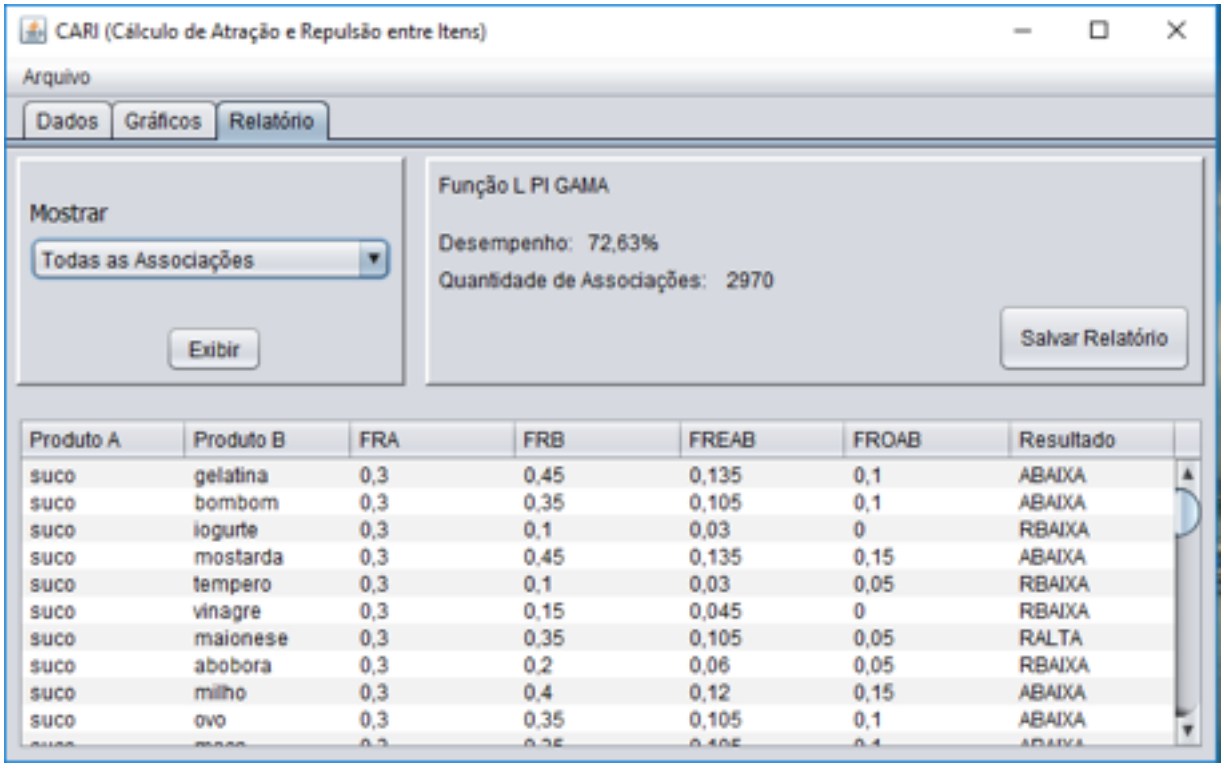
O algoritmo difuso foi implementado na linguagem Java e resultou em um sistema chamado CARI (Cálculo de Atração e Repulsão entre Itens). O CARI auxilia em todas as etapas do algoritmo desde a transformação dos dados de entrada até a classificação da força de atração/repulsão entre os itens.

A entrada de dados para o processamento do sistema é um arquivo no formato txt que atende os requisitos conforme a Tabela 1.

A Figura 9 mostra uma janela do sistema que apresenta o relatório do processo. Como pode ser observado na tabela apresentada pelo sistema, têm-se as seguintes informações: produto A (produto antecedente da associação), produto B (produto consequente da associação), FRA (frequência relativa do produto A), FRB (frequência relativa do produto B), FREAB (frequência relativa esperada de A e B) e FROAB (frequência relativa obtida de A e B).

Na Tabela 4 é mostrado o resultado das regras de algumas associações feitas pelo sistema CARI. Como pode ser observado, para a saída 1 a regra que apresentou maior valor pelo Max-Produto foi a de número dez resultando em uma associação do tipo RBAIXA (Repulsão Baixa) e a saída 2 é classificada como RALTA (Repulsão Alta) pelos mesmos critérios.

Figura 9- Interface do Sistema CARI (Cálculo de Atração e Repulsão entre Itens)



Produto A	Produto B	FRA	FRB	FREAB	FROAB	Resultado
suco	gelatina	0,3	0,45	0,135	0,1	ABAIXA
suco	bombom	0,3	0,35	0,105	0,1	ABAIXA
suco	iogurte	0,3	0,1	0,03	0	RBAIXA
suco	mostarda	0,3	0,45	0,135	0,15	ABAIXA
suco	tempero	0,3	0,1	0,03	0,05	RBAIXA
suco	vinagre	0,3	0,15	0,045	0	RBAIXA
suco	maionese	0,3	0,35	0,105	0,05	RALTA
suco	abobora	0,3	0,2	0,06	0,05	RBAIXA
suco	milho	0,3	0,4	0,12	0,15	ABAIXA
suco	ovo	0,3	0,35	0,105	0,1	ABAIXA

ISSN 2238-5630
Brasília-DF, v. 8, n. 2, junho-dezembro 2019

Tabela 4- Valores de saída para cada regra.

Regra	Saída 1	Saída 2
1	0,000	0,000
2	0,000	0,000
3	0,000	0,000
4	0,120	0,264
5	0,060	0,132
6	0,090	0,396
7	0,050	0,198
8	0,000	0,000
9	0,000	0,000
10	0,230	0,000
11	0,115	0,000
12	0,200	0,000
13	0,100	0,000
14	0,000	0,000
15	0,000	0,000
16	0,000	0,000
17	0,000	0,000

TESTES REALIZADOS

Foram realizados testes com diferentes combinações de funções de pertinência. São elas: Função Triangular, Função Trapezoidal, Função Gama, Função L, Função Gaussiana ou Pi, Função Z e Função Sigmoidal. Os limites e intervalos de cada função levaram-se em consideração os seguintes critérios: função baixa está entre 0 e 0,4; função média está entre 0,1 e 0,7 e a função alta está entre 0,4 e 1,0. Vide exemplo na Figura 6.

Ao determinar o maior valor de suporte válido na base de dados groceries, a saber, 7,4%; foram feitos 5 testes com o algoritmo Apriori em diferentes valores de SupMin (Suporte Mínimo) e ConfMin (Confiança Mínima) mostrados na Tabela 5. Isso possibilitou a varredura completa da base de dados e a compreensão do comportamento do método difuso quando comparado a diferentes taxas de suporte e confiança.

RESULTADOS E DISCUSSÕES

A Tabela 6 mostra o resultado dos testes realizados com as diferentes combinações de funções de pertinência. Para cada

Tabela 5 – Testes realizados com o Apriori e seus respectivos parâmetros

Teste	*SupMin*	*SupMax*	*ConfMin*	*ConfMax*	Quantidade de regras geradas
T01	0.030	0.074	0.20	0.45	25
T02	0.010	0.056	0.45	0.58	31
T03	0.006	0.009	0.60	0.64	8
T04	0.003	0.004	0.75	0.89	6
T05	0.001	0.002	0.90	1.00	129

Tabela 6 – Resultado das funções de pertinência comparadas ao LIFT

Função de pertinência	Desempenho atingido
L Triangular Gama	77,03%
L Triangular Sigmoidal	77,03%
L Trapeizoidal Gama	76,47%
L Trapezoidal Sigmoidal	76,47%
L Pi Gama	83,87%
L Pi Sigmoidal	83,87%
L Sino Gama	76,47%
L Sino Sigmoidal	76,47%
Z Triangular Gama	77,44%
Z Triangular Sigmoidal	77,44%
Z Trapeizoidal Gama	77,44%
Z Trapezoidal Sigmoidal	77,44%
Z Pi Gama	83,87%
Z Pi Sigmoidal	83,87%
Z Sino Gama	77,03%
Z Sino Sigmoidal	77,03%

ISSN 2238-5630
Brasília-DF, v. 8, n. 2, junho-dezembro 2019

uma delas obteve-se uma taxa de desempenho em relação ao LIFT.

Os resultados do método difuso aplicando diferentes combinações de funções de pertinência foram satisfatórios. Em média, o desempenho atingido foi de 78,70%. As combinações de funções com melhores taxas foram as compostas pela função Pi. Um dos motivos é a característica da função Pi possuir bordas mais suaves em relação a outras funções usadas. Apesar da diferença entre as taxas de cada função, percebe-se que, quando comparado ao LIFT, o desempenho final não foi tão distante entre elas. O algoritmo difuso usando qualquer uma das combinações mostrou-se eficiente e adequado para o cálculo de afinidade entre itens.

A Tabela 7 mostra os resultados do algoritmo difuso comparado com o algoritmo Apriori.

O algoritmo Apriori resulta nas melhores regras de associação com valores de SupMin e ConfMin fornecidos como entrada. Assim, 199 regras foram comparadas uma a uma com a classificação dada pelo método difuso. O critério de desempenho, nesse caso, foi a quantidade de regras que foram classificadas como "atração" pelo método difuso, podendo ser baixa, média ou alta.

Em todos os testes, com exceção do T01, o algoritmo difuso correspondeu 100% com o Apriori. Já no primeiro teste o desempenho foi de 88%. Essa taxa é aceitável, pois, no teste T01, as regras geradas pelo Apriori possuem valores baixos de confiança. Nos testes, percebe-se mais uma vez o ótimo desempenho do algoritmo difuso quando comparado ao LIFT, e chega a atingir taxas de 100% na maioria deles.

CONSIDERAÇÕES FINAIS

O desenvolvimento deste trabalho possibilitou saber se é adequado utilizar a abordagem difusa para modelar a imprecisão contida na matriz de co-ocorrência. Assim, verificou-se a viabilidade de usar lógica difusa no processo de MBA (Market Basket Analysis).

Diante dos testes realizados, o algoritmo difuso proposto 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 nos seguintes graus qualitativos: baixo, médio e alto. Foram

Tabela 7 – Resultado dos testes comparativos

Testes	QTD de regras com "atração" (algoritmo difuso)	QTD de regras com "repulsão" (algoritmo difuso)	Desempenho em relação ao Apriori	Desempenho em relação ao LIFT
T01	22	3	88%	96%
T02	31	0	100%	100%
T03	8	0	100%	100%
T04	6	0	100%	100%
T05	129	0	100%	100%

feitos testes comparativos com algoritmo Apriori e o valor do LIFT.

O desempenho atingido pelo método difuso quanto ao LIFT foi em média de 78% e quanto ao Apriori de 98%. Houve testes que o desempenho foi de 100% conforme mostrado na Tabela 7. Os testes realizados com o algoritmo difuso e a validação feita quando comparado ao Apriori e LIFT, mostrados na Tabela 6 e na Tabela 7, permitiram que os objetivos propostos fossem alcançados.

Determinar relações entre itens que estão ocultas em grandes bases de dados é uma das principais tarefas de mineração de dados que tem ganhado foco nos últimos anos. Assim, otimizar este processo tornando-o mais preciso implica melhorias de muitos procedimentos em diferentes áreas.

O método difuso mostrou ótimos resultados quando comparado ao LIFT e Apriori, 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

BILOBROVEC, M.; MARÇAL, R. F. M.; KOVALESKI, J. L. **Implementação de um sistema de controle inteligente utilizando a lógica fuzzy**. XI SIMPEP, Bauru/Brasil, 2004.

BRANDAO, Euzeli da Silva et al . Proposta de reconhecimento de padrão de conforto em clientes com pênfigo vulgar utilizando a Lógica Fuzzy. **Rev. esc. enferm**. USP, São Paulo , v. 47, n. 4, p. 958-964, Aug. 2013.

CAMILO, C. O.; SILVA, J. C. **Mineração de dados**: Conceitos, tarefas, métodos e ferramentas. Universidade Federal de Goiás (UFC), p. 1-29, 2009.

CARVALHO, R. A. S. Data mining no contexto de customer relationship management em uma franquia coca-cola company. 2010. 152 f.. Dissertação (mestrado em ciências da computação) – Universidade Federal de Pernambuco, Recife. Out. 2010.

CHERRI, Adriana Cristina; ALEM JUNIOR, Douglas José; SILVA, Ivan Nunes da. Inferência fuzzy para o problema de corte de estoque com sobras aproveitáveis de material. **Pesquisa Operacional**, v. 31, n. 1, p. 173-195, 2011.

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

GOMIDE, F. A. C.; GUDWIN, R. R. Modelagem, controle, sistemas e lógica fuzzy. **SBA Controle e Automação**. Campinas –SP, v. 4 , n. 3, p. 97-115, set./out. 1994.

HAHSLER, Michael; GRUEN, Bettina; HORNIK, Kurt. arules - A Computational Environment for Mining Association Rules and Frequent Item Sets. J**ournal of Statistical Software** 14/15. 2005. URL: http://dx.doi.org/10.18637/jss.v014.i15.

OLIVEIRA, R. B. T. de. O processo de extração de conhecimento de base de dados apoiado por agentes de software. 2000. 115 f.. Dissertação (mestrado em ciências da computação e matemática computacional) – Universidade de São Paulo- USP, São Carlos. Out. 2000.

PACHECO, R. C. S. et al. **Tratamento de imprecisão em sistemas especialistas**. 1991. Disponível em: https://repositorio.ufsc.br/xmlui/handle/123456789/157703 Acesso em: 04 de maio de 2017

PONCIANO, P. F.; LOPES, M. A.; YANAGI JUNIOR, T.; FERRAZ, G. A. S. Análise do ambiente para frangos por meio da lógica fuzzy: uma revisão. **Archivos de Zootecnia**, Córdoba, v.60, n.1, p.1-13, 2011.

RIGNEL, D.G.S. ; CHENCI, G.P. ; LUCAS, C.A.. Uma introdução à lógica fuzzy. **Revista eletrônica de sistemas de informação e gestão tecnológica**. v 1, n. 12. p. 17-28, março 2011.

RODRIGUES, L. M.; DIMURO, G. P. **Utilizando Lógica Fuzzy para Avaliar a Qualidade de uma Compra Via Internet**. 2010.

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.

ISSN 2238-5630
Brasília-DF, v. 8, n. 2, junho-dezembro 2019

Dez. 2004.

SCHAEFFER, A. G. Data mining no varejo: estudo de caso para loja de materiais de construção. 2003. 86 f.. Tese (mestrado em ciência da computação) – Universidade Federal do Rio Grande do Sul, Porto Alegre. Abril 2003.

SPERONI, R. M. Obtenção de dados para mineração da utilização da web. 2003. 73 f.. Dissertação (mestrado em ciência da computação) – Universidade Federal de Santa Catarina, Florianópolis. Set. 2003.

TANSCHEIT, R; GOMIDE, F. A. C.; GUDWIN, R. R. Conceitos fundamentais da teoria de conjuntos fuzzy, lógica fuzzy e aplicações. In: Proc. 6 th IFSA Congress-Tutorials. 1995. p. 1-38.

VIEIRA, Samuel Oliveira et al . Aplicação do método Fuzzy na classificação da zona de convergência do Atlântico Sul no sul da Amazônia. Rev. bras. meteorol., São Paulo , v. 29, n. 4, p. 621-631, Dec. 2014 .

Agradecimentos: Ao Conselho Nacional de Pesquisa e Desenvolvimento Científico e Tecnológico (CNPq) por ter financiado esta pesquisa.

CURRÍCULOS

* http://lattes.cnpq.br/3262542868714570
** http://lattes.cnpq.br/5406666543919084