UNIVERSIDADE ESTADUAL DE CAMPINAS - FACULDADE DE TECNOLOGIA
BACHARELADO EM SISTEMAS DE INFORMAÇÃO
TÓPICOS EM COMPUTAÇÃO E INFORMÁTICA I

JULIANA PINHO MARCHI 177291
LETÍCIA SOUSA DE OLIVEIRA 201506
LUCIANO SOUZA GOMES DO NASCIMENTO 202215
THIAGO HENRIQUE VIOTTO 187583

Relatório do trabalho

Mineração de dados

LIMEIRA - SP
2018

SUMÁRIO

Este relatório refere-se à documentação do desenvolvimento do trabalho prático da disciplina *TT001A - Tópicos em Computação e Informática I*, o qual consistia no estudo e escolha de um cenário para aplicar a técnica de mineração de dados, de modo que possa-se obter informações para gerar conhecimento e sanar determinados problemas. Portanto, para a melhor compreensão do trabalho, é necessário esclarecer alguns pontos referentes à esta prática que tem sido aplicada cada vez mais.

1.1 O que é mineração de dados?

Com o crescimento exponencial da quantidade de dados atualmente, o estudo manual deles tornou-se praticamente inviável, além da dificuldade da distinção de dados úteis de dados não úteis. Tal cenário acaba deixando inoperável grandes quantidades de dados, dos quais poderiam ser obtidos uma grande gama de informações, gerando os “cemitérios” de dados.

Mineração de dados é uma ferramenta criada justamente para a análise destes dados, consistindo na descoberta e exploração de grandes bases de dados (*Knowledge Discovered in Databases*), a fim de extrair regras e padrões, de modo a gerar conhecimento.

Existem várias técnicas para sua aplicação, das quais se destacam: a classificação, o agrupamento e a associação. No desenvolvimento deste trabalho, o foco será a aplicação da técnica de associação.

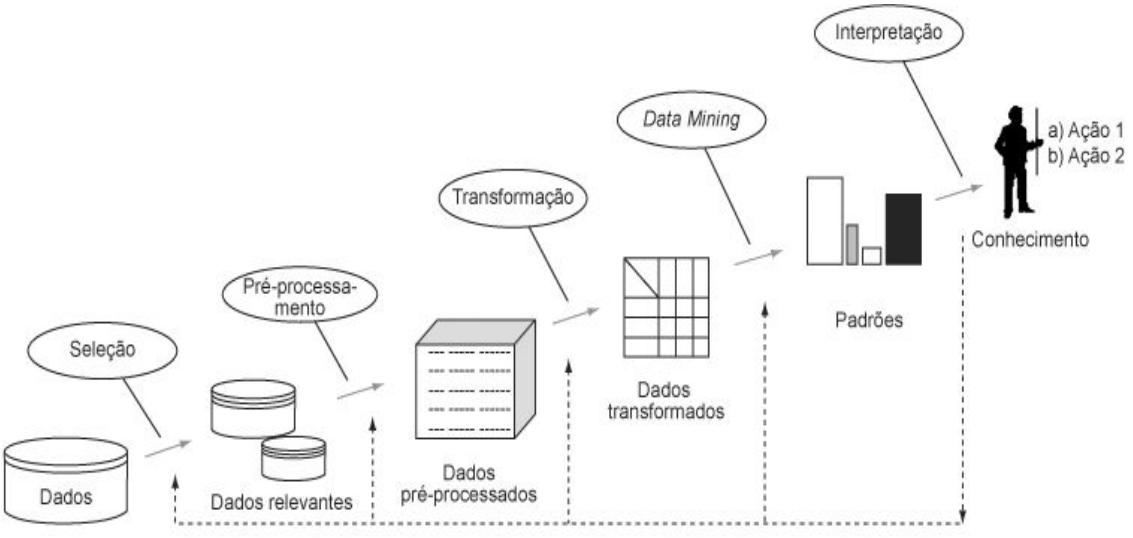

Figura 01: Etapas do processo KDD (Fayyad et al. (1996))

Na *Figura 01*, é possível observar as etapas do processo do *KDD*. Primeiramente, os dados são selecionados. Em seguida é feito o pré-processamento, a fim de identificar anomalias e inconsistências. Deste modo, os dados são transformados, sendo aplicado uma determinada técnica de mineração de dados, possibilitando a identificação de regras e padrões, possibilitando a interpretação dos dados finais.

1.2. Tarefas de Mineração de Dados

1.2.1 Classificação

A técnica de classificação consiste na identificação de classes (ou rótulos) de um registro específico, prevendo uma classe de um dado novo automaticamente. Uma característica do algoritmo de classificação é sua necessidade de treinamento, ou seja, é necessário que um conjunto de dados já rotulados sejam submetidos ao programa, de modo que ele gere regras de classificação. Se o conjunto de treinamento for satisfatoriamente bom, o modelo será capaz de predizer novos registros não rotulados.

Para sua avaliação de desempenho, observa-se a capacidade de generalização, a qual refere-se o quão bom o algoritmo é na predição desses novos dados.

Um exemplo é uma base que armazena características de cliente baseados em seu histórico de transações. Os clientes poderão ser classificados em determinadas categorias, de acordo com suas características. Abaixo segue alguns exemplos de aplicações:

- ➔ Identificação de quando uma pessoa possa ser uma ameaça à segurança;
- ➔ Identificação do local de uma doença;
- ➔ Identificação de uma turma para um aluno (qual a mais indicada);
- ➔ Determinar quando uma transação de cartão de crédito possa ser uma fraude.

1.2.2 Agrupamento

A técnica de agrupamento consiste na congregação de dados similares em grupos, os quais são diferentes entre si. Existem diversos algoritmos de agrupamento, um exemplo deles é o K-Means, no qual os dados são postos em um plano R^n e têm suas distâncias para outros pontos específicos calculadas, delimitando, através de *n* iterações, a área de um grupo.

A diferença da classificação é que o agrupamento não pretende classificar dados em categorias a fim de estimar um valor e sim em apenas os separar em grupos de dados semelhantes. Outra diferença entre essas duas técnicas é no método de aprendizagem dos algoritmos, o qual da classificação é supervisionado, ou seja, é necessário um conjunto de treinamento, diferentemente do agrupamento, o qual possui aprendizado não supervisionado e, portanto, não precisa ser treinado.

Para a avaliação do resultado dos agrupamentos, normalmente é utilizado o coeficiente de silhueta, o qual realiza um cálculo específico utilizando as distâncias inter e intra-grupos, de modo a verificar se um objeto de um determinado grupo é mais similar aos objetos do seu próprio grupo do que objetos de outros grupos.

Abaixo segue alguns exemplos de aplicações de agrupamento:

- ➔ Separação de produtos em um supermercado de acordo com seu nicho;
- ➔ Separação de comportamentos suspeitos para a auditoria;
- ➔ Redução de um conjunto com atributos parecidos com centenas de atributos.

1.2.3 Associação

A tarefa de associação consiste em, através de análises, relacionar elementos em uma base de dados transacional (transações). Um exemplo de problema de associação é a aplicação em “Cesta de compra de supermercados”, onde o objetivo é encontrar dados comprados em conjunto, a fim de posicionar estoques e estratégias para atrair o consumidor (LEE et al., 2016).

Tabela 1 - Exemplo de transações de cesta de mercado

Identificador da transação	Itens
1	{Pão, Leite}
2	{Pão, Fraldas, Cerveja, Ovos}
3	{Leite, Fraldas, Cervejas, Refrigerante}
4	{Pão, Leite, Fraldas, Cerveja}
5	{Pão, Leite, Fraldas, Refrigerante}

Figura 02: “Tabela 1: Exemplo de transações de cesta de mercado”.

As regras de associação são interessantes para criar relacionamentos escondidos em grandes bases de dados. As regras descobertas podem ser analisadas como conjunto de itens frequentes ou na forma de regras associativas. Como exemplo, a seguinte regra pode ser extraída da Tabela 1.

{Fraldas -> Cerveja}

É possível verificar uma forte relação na venda de fraldas e cervejas, pois muitos clientes que compram cerveja, também compram fraldas. Deste modo, os varejistas podem utilizar essas informações para deixar esses produtos próximos nas prateleiras para atrair os consumidores.

O algoritmo de mineração de dados conhecido como APRIORI, encontra itens frequentes numa transação. Uma transação indica tuplas de uma base, enquanto os itens as colunas, ou seja, os atributos. Cada transação possui um subconjunto de itens, conhecido como *itemsets* (BASU et al., 2016).

Na criação de uma regra de associação, um conjunto de *itemsets* de uma transação deve ser analisada. Este *itemset* deve ser frequente numa base de dados, tendo uma frequência mínima (suporte) estabelecida pela regra (HAN; KAMBER; PEI, 2012).

Além do suporte, existe um outro índice que valida e limita o número de regras criadas, conhecido como *confiança*. A confiança mostra quantas vezes o suporte se mostrou verdadeiro. Se as regras obedecerem um suporte mínimo e uma confiança mínima, elas são consideradas regras interessantes (HAN; KAMBER; PEI, 2012).

Um suporte de 2% para a regra {Fraldas} →{Cerveja}, significa que 2% de todas as transações mostram que fraldas e cervejas são compradas juntas. Uma confiança de 60% significa que 60% das transações que compraram fraldas também compraram a cerveja.

Outro índice estatístico frequente em regras de associação é o *lift*. O lift de uma regra de associação representado por A Þ B indica quanto frequente é B se A ocorre. Esta medida é computada por: Lift(A Þ B) = Conf(A Þ B) ÷ Sup(B).

2. DESCRIÇÃO DO PROBLEMA

Uma empresa de venda de jogos quer personalizar a página do site de acordo com a localização do usuário, levando em consideração os jogos mais vendidos na presente localidade.

Neste trabalho, iremos analisar as plataformas, as editoras e os gêneros mais vendidos em cada localização e em escala global em um aspecto temporal.

2.1 Atributos

A base de dados utilizada no desenvolvimento do trabalho é uma base transacional e possui os seguintes atributos:

- ➔ Platform: plataforma;
- ➔ Year: ano de lançamento;
- ➔ Genre: gênero;
- ➔ Publisher: editora;
- ➔ NA_Sales: vendas na América do Norte;
- ➔ EU_Sales: vendas na Europa;
- ➔ JP_Sales: vendas no Japão;
- ➔ Other_Sales: vendas nas demais localidades;
- ➔ Global_Sales: vendas globais (somatória de todas as outras).

2.2 Pré-processamento

Um dos fatores motivadores do pré-processamento é o princípio GIGO (Garbage In, Garbage Out), onde dados errados tendem a atrapalhar a execução do algoritmo, ocasionando em resultados indesejados. O pré-processamento tem como objetivo monitorar dados brutos a fim de que possa ser extraído daquela base um conhecimento condizente/válido, eliminando anomalias e inconsistências antes da aplicação de um algoritmo de mineração.

Muitas bases de dados podem estar com valores ausentes, dados repetidos, dados errados e entre outros. Tarefas de pré-processamento, como a limpeza, integração e transformação buscam resolver estes problemas e preparar a base para a execução de algoritmos de mineração de dados.

2.2.1 Limpeza

Nesta etapa, o intuito é imputar valores ausentes e eliminar inconsistências dos dados. Segue algumas técnicas para imputação de valores ausentes:

1. Média e Moda → a média dos atributos numéricos e a moda dos atributos nominais são calculados, sendo colocados nos valores ausentes.
2. Ignorar objeto → o objeto com valores ausentes é excluído da base, podendo ocasionar na perda de muitos dados.
3. Hot Deck → os objetos similares com o objeto que contém valores ausentes são aleatoriamente selecionados, substituindo nos dados faltantes.

‘ Dados inconsistentes são dados com formas distintas em determinadas situações, por exemplo: dados duplicados, diferentes categorias para o mesmo caso(criança e infantil, por exemplo), valores muito discrepantes em comparação com o restante de registros, atributos numéricos e categóricos, etc.

Nestes casos, onde não há uma padronização dos dados, é dificultada a aplicação de um algoritmo de mineração de dados. Para tanto, é necessária a correção deste cenário, a qual normalmente tem que ser feita manualmente, sendo necessário uma análise profunda da base de dados.

2.2.2 Integração

Na integração é gerada uma massa de dados, ou seja, os dados são integrados numa mesma base. Esses dados podem ser de outras bases, arquivos, planilhas, data warehouses, vídeos, imagens, entre outras (CAMILO, 2009). É interessante esta técnica pois vários dados podem estar em locais distintos, sendo necessário uni-los em só local para a análise e estudo do cenário como um todo.

2.2.3 Transformação

Alguns dados podem estar com valores em maiúsculo e outros em minúsculo. A transformação tem como objetivo padronizar esses dados. Não existe um critério único para transformação dos dados e diversas técnicas podem ser usadas de acordo com os objetivos pretendidos. Algumas das técnicas empregadas nesta etapa são: suavização (remove valores errados dos dados), agrupamento (agrupa valores em faixas sumarizadas), generalização (converte valores muito específicos para valores mais genéricos), normalização (colocar as variáveis em uma mesma escala) e a criação de novos atributos (gerados a partir de outros já existentes) (CAMILO, 2009).

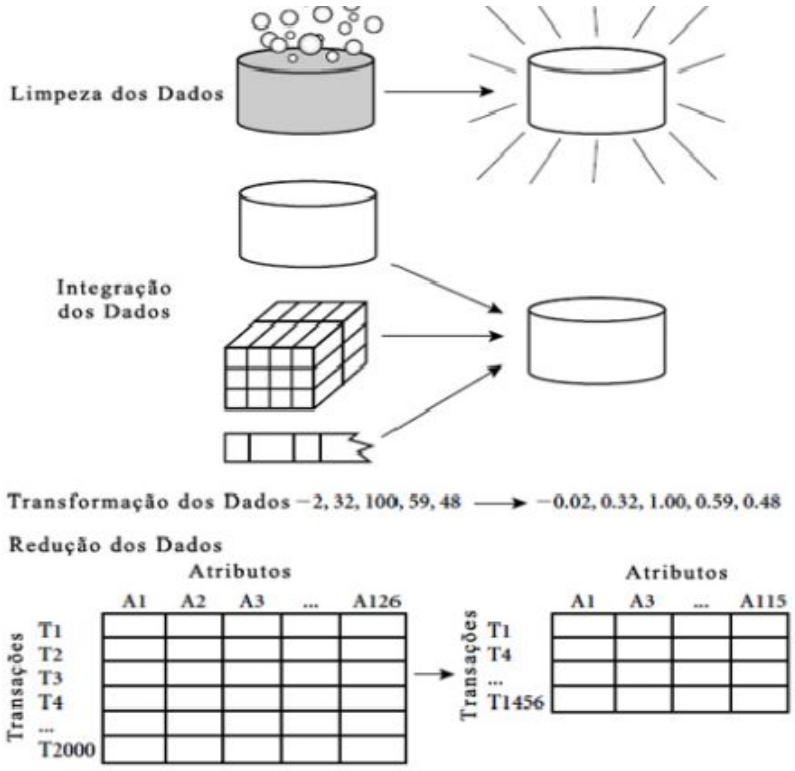

Figura 03: Processo de Pré-Processamento

Para o desenvolvimento do trabalho, a base que, ao início, contava com aproximadamente 1660 registros, foi modificada na fase de pré-processamento.

Em primeira instância, tendo em vista o objetivo desta mineração (personalização da página inicial de um site de acordo com as maiores vendas) e a condição da base (a qual era ranqueada, ou seja, os jogos mais vendidos ficavam ao topo) foi decidido e consentido o estudo ser realizado somente com os primeiros 4 mil registros.

Alguns registros contavam com *Year* = “N/A”. Nestes casos, analisando a quantidade de vendas, alguns deles foram removidos e outros dados inputados através de pesquisa manual.

Alguns outros atributos, como *Platform* e *Publisher* se apresentavam ausentes em alguns poucos registros. Neste caso, o procedimento adotado foi semelhante ao do *Year*.

Na associação, valores numéricos não são indicados, portanto foi necessário categorizar os atributos de Venda e Ano, através dos seguintes critérios:

Categorias de Venda:

(em milhões)	Baixíssimo	Baixo	Médio	Médio-Alto	Alto	Altíssimo
NA_Sales	menor que 0,5	0,5 a 1	1 a 2,5	2,5 a 5	5 a 15	maior que 15
EU_Sales	menor que 0,5	0,5 a 1	1 a 2,5	2,5 a 5	5 a 15	maior que 15
JP_Sales	menor que 0,5	0,5 a 1	1 a 2,5	2,5 a 5	5 a 15	maior que 15
Other_Sales	menor que 0,5	0,5 a 1	1 a 2,5	2,5 a 5	5 a 15	maior que 15
Global_Sale	menor que 0,5	0,5 a 1	1 a 5	5 a 10	10 a 20	maior que 20

Categorias de Ano:

	Muito Antigo	Antigo	Médio	Novo
Ano	Antes de 2000	2000 a 2007	2007 a 2013	2013 a 2016

A base de dados utilizada no desenvolvimento do trabalho possuía pouquíssimos dados incoerentes ou faltantes (ruídos), os quais foram corrigidos sem muitos esforços. Portanto, à pedido da professora, foram integrados à base de dados cerca de 10% de registros contendo um ou mais dados incompletos, a fim de atestar sua influência no resultado.

Para isso, foi desenvolvido um código em C++ para gerar uma pequena base de dados com ruídos, a qual foi integrada à base original para testes. O código em questão está disponível em:
https://drive.google.com/file/d/1I2-KElg0ed3TEACLiMYmSmH-dohJqJ1l/view?usp=sharing

2.3 Tarefa utilizada

A tarefa escolhida para a solução do problema foi a associação. Ela possibilitará uma análise profunda sobre cada atributo da base de dados, podendo prever determinadas regras a fim de promover as vendas de jogos.

O algoritmo utilizado será o APRIORI, visto que ele permite encontrar itens frequentes numa transação e é um dos mais populares e com melhor eficiência.

A aplicação desta tarefa pode ajudar a prever quais plataformas são mais preteridas em determinadas áreas. Isso ajuda na análise de quais jogos promover com abundância e divulgação em locais específicos, já que existem jogos unificados para determinadas plataformas, além de um padrão de gênero.

Os gêneros dos jogos também podem ser identificados com mais frequência de ocorrência em determinados locais, permitindo assim promover jogos com gêneros específicos para estas áreas.

Dessa forma, as regras de associação podem fornecer indícios de possíveis relações existentes entre os atributos envolvidos nas vendas de jogos. Tais relações podem ajudar a compreender e prever quais as áreas para vendas de determinados jogos, além das análises das mudanças de costumes e gostos em relação a determinados games ao passar dos anos.

O software utilizado para a tarefa de associação foi o **Weka**, disponível em: https://www.cs.waikato.ac.nz/ml/weka/

3. EXPERIMENTOS

Nesta seção são descritos os experimentos realizados durante o desenvolvimento do trabalho. Cada subseção traz os experimentos referentes ao período descrito e, para cada experimento, as informações: os atributos utilizados, níveis mínimos de suporte e confiança e as regras mais promissoras encontradas. Ao final desta seção, foi apresentada uma discussão dos experimentos.

3.1 Plataformas

Neste experimento foi levado em consideração apenas as plataformas, desconsiderando as vendas da Europa, Japão, América do Norte, vendas globais e de outros lugares do mundo. Os outros atributos como a ano, gênero e editora também foram utilizados.

Experimento 01:
Níveis mínimos de suporte: 7%
Níveis mínimos de confiança: 10%

Com 11% de suporte e 76% de confiança, a regra abaixo relata que a plataforma PS2 foi mais frequente no período antigo, sendo comprovada tal regra visto na inexistência desta plataforma nos períodos mais recentes.

{Platform = PS2} -> {Year = antigo}
Suporte: 11% Confiança: 76% Lift: 2.9

Com 9% de suporte, 83% e 81% de confiança respectivamente, as plataformas Xbox 360 e PS3 foram mais vendidas no período médio (2006 a 2013). É possível perceber uma mudança nos gostos das plataformas pelas pessoas ao passar dos anos, além da boa adaptação com os lançamentos dos consoles mais modernos. O Xbox 360 foi lançado em 2005 e o PS3 em 2006, e logo no lançamento e nos anos seguintes verificamos a alta quantidade de vendas e adaptação das pessoas ao consoles mais novos.

{Platform = XB360} -> {Year = medio}
Suporte: 9% Confiança: 83% Lift: 1.9
{Platform = PS3} -> {Year = medio}
Suporte: 9% Confiança: 81% Lift: 1.85

Com suporte de 7% e confiança de 95%, a plataforma Wii foi mais frequente no período médio (2006 a 2013).

{Platform = Wii} -> {Year = medio}
Suporte: 7% Confiança: 95% Lift: 2.19

Com suporte de 7% e confiança de 79%, a plataforma PS1 foi mais vendida no período muito antigo (antes de 2000). Isso pode ser comprovado por ser o primeiro video game lançado pela Sony justamente nessa época.

{Platform = PS1} -> {Year = muito antigo}
Suporte: 7% Confiança: 79% Lift: 4.57

3.2 Vendas globais

Neste experimento foi levado em consideração apenas as plataformas e as vendas globais.

Experimento 02:
Níveis mínimos de suporte: 4%
Níveis mínimos de confiança: 10%

É possível analisar nas regras abaixo que a plataforma PS2 obteve mais jogos vendidos (unitários) do que as plataformas Wii, Xbox 360 e PS3. É possível relacionar toda essa popularidade das plataformas com o grande tempo em que elas se mantiveram como as principais no mundo aproximadamente entre 2005 e 2016.

{Platform = PS2} -> {Global_Sales = Medio}
Suporte: 7% Confiança: 48% Lift: 1.05
{Platform = XB360} -> {Global_Sales = Medio}
Suporte: 5% Confiança: 45% Lift: 0.99
{Platform = PS3} -> {Global_Sales = Baixo}
Suporte: 5% Confiança: 49% Lift: 1.04
{Platform = Wii} -> {Global_Sales = Medio}
Suporte: 4% Confiança: 50% Lift: 1.06

3.3 Vendas e plataformas

Neste experimento foi levado em consideração apenas as plataformas e todas as vendas. Os testes foram feitos separadamente, testando as plataformas com as vendas de determinada região, assim sucessivamente. Os registros da base foram diminuídos com o intuito de encontrar mais facilmente quais as plataformas mais vendidas em determinados locais.

Experimento 03:
Total de registros: 999
Níveis mínimos de suporte: 1%
Níveis mínimos de confiança: 10%

Nas regras abaixo é possível verificar que a plataforma PS2 foi predominante nos locais da Europa, América do Norte e vendas globais. A plataforma NES foi predominante no Japão, no local onde foi criada. Já o PS3 foi escolhido pelos outros locais do mundo como a plataforma preferida.

{Platform = PS2} -> {NA_Sales = medio}
Suporte: 9% Confiança: 57% Lift: 1.07
{Platform = PS2} -> {EU_Sales = medio}
Suporte: 6% Confiança: 41% Lift: 1.19
{Platform =NES} -> {JP_Sales = medio}
Suporte: 1% Confiança: 53% Lift: 4.33
{Other_Sales = medio} -> {Platform: PS3}
Suporte: 1% Confiança: 30% Lift: 2.45
{Platform = PS2} -> {Global_Sales: medio}
Suporte: 13% Confiança: 89% Lift: 1.12

3.4 Vendas e gêneros

Neste experimento foi levado em consideração apenas os gêneros e todas as vendas. Os testes foram feitos separadamente, testando os gêneros com as vendas de determinada região, assim sucessivamente. Os registros da base foram diminuídos com o intuito de encontrar mais facilmente quais as plataformas mais vendidas em determinados locais.

Experimento 04:
Total de registros: 999
Níveis mínimos de suporte: 0.03%

Níveis mínimos de confiança: 49%

Nas regras abaixo é possível verificar que o gênero Sports foi predominante nos locais da Europa. O gênero Action,Sports e Shooter não tem grandes vendas no japão, porém o gênero Role Playing tem mais vendas quando as vendas são no Japão, na América do norte não foram observadas regras devido a boa distribuição das vendas dos diversos gêneros.

{EU_Sales=Altissima } -> {Genre=Sports}
Suporte 0.03% Confiança: 100% Lift: 6.07
{JP_Sales=Medio alto} -> {Genre=Role-Playing }
Suporte: 0.6% Confiança: 49% lift: 5.06
{Genre=Action} -> {JP_Sales=Baixissimo}
Suporte:15% Confiança:93% Lift: 1.08
{Genre=Sports} -> {JP_Sales=Baixissimo}
Suporte 15% Confiança: 91% Lift: 1.06

Pode se observar também que nas vendas mundiais os gêneros que na média são mais vendidos são o de esportes e ação.

{Genre=Action} ->{ Global_Sales=Medio}
Suporte: 7% Confiança: 46% lift: 1
{Genre=Sports} -> {Global_Sales=Medio}
Suporte: 7% Confiança:42% lift: 0.93

Foi possível observar que no Atributo Other_Sales todos os gêneros tem vendas associadas com baixíssimas sendo o suporte mínimo 0,2% com uma confiança mínima de 91%, ou seja, não são comparadas com as vendas nos principais polos, podendo concluir que não há um gênero de preferência.

3.5 Anos e o gêneros

Neste experimento foi levado em consideração apenas os gêneros e todos os anos. Os registros da base foram diminuídos com o intuito de encontrar mais facilmente quais os gêneros preferidos em determinados anos.

Experimento 05:
Total de registros: 999
Níveis mínimos de suporte: 0,1%

Níveis mínimos de confiança: 10%

{Genre = Action} -> {Year = medio}
Suporte: 8% Confiança: 41% Lift: 1.02
{Genre = Sports} -> {Year = medio}
Suporte: 6% Confiança: 45% Lift: 1.13
{Genre = Misc} -> {Year = medio}
Suporte: 5% Confiança: 57% Lift: 1.43
{Genre = Role_Playing} -> {Year = medio}
Suporte: 4% Confiança: 45% Lift: 1.12
{Genre = Shooter} -> {Year = medio}
Suporte: 4% Confiança: 43% Lift: 1.07
{Genre = Strategy} -> {Year = muito_antigo}
Suporte: 0,8% Confiança: 67% Lift: 3.25
{Genre = Adventure} -> {Year = muito_antigo}
Suporte: 0,1% Confiança: 37% Lift: 1.79
{Genre = Fighting} -> {Year = antigo}
Suporte: 1% Confiança: 35% Lift: 1.42
{Genre = Simulation} -> {Year = antigo}
Suporte: 1% Confiança: 35% Lift: 1.41
{Genre = Shooter} -> {Year = novo}
Suporte: 1% Confiança: 29% Lift: 1.98
{Year = novo} -> {Genre = Action}
Suporte: 1% Confiança: 27% Lift: 1.31

Nesta análise, verificamos que os gêneros Action, Sports, Role-Playing, Misc e Shooter foram preferidos nos anos médios. Os gêneros Fighting e Simulation foram mais frequentes no período antigo, enquanto que Adventure e Strategy no período muito antigo. Os gêneros Action e Shooter são mais frequentes no período novo. Isso pode ser comprovado com o surgimento de campeonatos destas modalidades. É possível compreender a evolução das preferências dos games ao passar dos anos.

3.6 Vendas e editoras

Neste experimento foi levado em consideração apenas as editoras e todas as vendas. Os testes foram feitos separadamente, testando as editoras com as vendas de determinada região, assim sucessivamente. Os registros da base foram

diminuídos com o intuito de encontrar mais facilmente quais os gêneros preferidos em determinados anos.

Experimento 06:
Total de registros: 999
Níveis mínimos de suporte: 0,1%
Níveis mínimos de confiança: 10%

{Global_Sales = Altissima} -> {Publisher = Nintendo}
Suporte: 100% Confiança: 85% Lift: 3.7
{Other_Sales = Medio alto} -> {Publisher = Nintendo}
Suporte: 50% Confiança: 63% Lift: 2.72
{NA_Sales = Alta} -> {Publisher=Nintendo}
Suporte: 5% Confiança: 50% Lift: 2.18
{EU_Sales = Alta} -> {Publisher=Nintendo}
Suporte: 0,8% Confiança: 54% Lift :2.35
{JP_Sales = Alta} -> {Publisher = Nintendo}
Suporte: 0,8% Confiança: 100% Lift: 4.36

Através desse experimento, foi completamente possível observar que a Nintendo é a editora que mais vendeu ao redor do mundo. Portanto, seria muito viável à WumpusGames, sempre reservar um espaço à essa editora em sua tela inicial.

3.7 Vendas, editoras e anos

Neste experimento foi levado em consideração apenas as editoras, todas as vendas e todos os anos. Os testes foram feitos separadamente, testando as editoras com as vendas de determinada região, assim sucessivamente. Os registros da base foram diminuídos com o intuito de encontrar mais facilmente quais os gêneros preferidos em determinados anos.

Experimento 07:
Total de registros: 999
Níveis mínimos de suporte: 0,1%
Níveis mínimos de confiança: 10%

{Year = Muito antigo, NA_Sales = Alta} -> {Publisher = Nintendo}
Suporte: 1% Confiança: 92% Lift: 4.02
{Publisher = Activision, NA_Sales = Medio alto} -> {Year = Medio}

Suporte: 1% Confiança: 50% Lift: 1.25

{Year = Muito antigo, EU_Sales = Alta} -> {Publisher = Nintendo}

Suporte: 0,3% Confiança: 100% Lift: 4.36

{Publisher = Take-Two Interactive, EU_Sales = Alta} -> {Year=Novo}

Suporte: 0,3% Confiança: 75% Lift: 5.13

{Year = Medio, JP_Sales = Alta} -> {Publisher = Nintendo}

Suporte: 0,3% Confiança: 100% Lift: 4.36

{Publisher = Nintendo, Other_Sales = Medio} -> {Year = Medio}

Suporte: 0,3% Confiança: 63% Lift: 1.56

{Year = Novo, Other_Sales=Medio} -> {Publisher = Electronic Arts}

Suporte: 0,3% Confiança: 42% Lift: 2.79

{Publisher=Ubisoft, Global_Sales = Medio alto} -> {Year = Medio}

Suporte: 0,3% Confiança: 100% Lift: 2.5

{Year = Antigo, Global_Sales = Altissima} -> {Publisher = Nintendo}

Suporte: 0,3% Confiança: 75% Lift: 3.27

{Year = Novo, Publisher = Capcom} -> {JP_Sales = Medio alto}

Suporte: 0,1% Confiança: 100% Lift: 20.37

Com o experimento acima realizado, é possível notar que as vendas realizadas da Nintendo, em sua maioria, eram nos tempos antecessores a ‘médio’, mostrando que ela dominava o mercado antes. Entretanto, o cenário aparenta estar mudando, à medida que outras editoras aparecem nas regras atuais, possuindo vendas médias altas ou altas.

Capcom aparece com boas vendas no Japão ultimamente, assim como a Eletronic Arts tem um desempenho considerável ao redor do mundo e a Take-Two Interactive, a qual apresenta altas vendas na Europa. Para tanto, seria interessantíssimo à WumpusGames disponibilizar um espaço aos conteúdos destas editoras na página inicial de seu site.

4. EXPERIMENTOS COM RUÍDOS

Nesta seção serão realizados alguns mesmos experimentos realizados no capítulo anterior, a fim de comparar os resultados obtidos e identificar a influência que os ruídos têm sobre o conhecimento gerado a partir da mineração de dados.

Cada subseção traz os experimentos referentes ao período descrito e, para cada experimento, as informações: os atributos utilizados, níveis mínimos de suporte e confiança e as regras mais promissoras encontradas.

4.1 Plataformas *(referência à subseção 3.1)*

Neste experimento foi levado em consideração apenas as plataformas, desconsiderando as vendas da Europa, Japão, América do Norte, vendas globais e de outros lugares do mundo. Os outros atributos como a ano, gênero e editora foram utilizados.

Experimento 08:
Níveis mínimos de suporte: 5%
Níveis mínimos de confiança: 70%

É possível observar que a plataforma mais utilizada no período antigo foi o PS2, e a plataforma mais utilizada no período médio é o PS3, não foi observada uma plataforma no período atual, provavelmente devido a grande quantidade das tais nesse período.

{Platform = PS2} -> {Year = antigo}
Suporte: 10% Confiança: 74% Lift: 2.93
{Platform = PS3} -> {Year = medio}
Suporte: 9% Confiança: 79% Lift: 1.92
{Platform = XB360} -> {Year = medio}
Suporte: 8% Confiança: 80% Lift: 1.95

Nessa análise não foi observada mudança nas plataformas devido ao ruído, ou seja deve haver uma grande quantidade de registros com essas plataformas o que aumenta a confiabilidade dos resultados, a única alteração foi nos valores de confiança e suporte.

4.2 Vendas globais *(referência à subseção 3.2)*

Neste experimento foi levado em consideração apenas as plataformas e as vendas globais.

Experimento 09:
Níveis mínimos de suporte: 4%
Níveis mínimos de confiança: 10%

Foi notado que as plataformas que tiveram mais vendas no mundo no decorrer dos anos são PS2, XB360 e PS3.

{Platform = PS2} -> {Global_Sales = Medio}
Suporte: 7% Confiança: 53% Lift: 1.14
{Platform = XB360} -> {Global_Sales = Medio}
Suporte: 5% Confiança: 48% Lift:1.03
{Platform = PS3} -> {Global_Sales = Medio}
Suporte: 5% Confiança: 46% Lift: 0.99

Com a execução com ruídos percebe-se divergência nos resultados, onde as confianças apresentam valores diferentes, além da mudança da regra do PS3, onde, no experimento sem ruídos, apresenta as Vendas Globais como “Baixo” diferentemente deste experimento, o qual aponta que suas vendas foram médias. Outra diferença observável é que não foi possível encontrar regras para o Wii.

Este experimento evidencia a importância da limpeza de dados, pois apresentou diferenças que poderiam influenciar as tomadas de decisão.

4.3 Vendas e plataformas *(referência à subseção 3.3)*

Neste experimento foi levado em consideração apenas as plataformas e todas as vendas.

Experimento 10:
Níveis mínimos de suporte: 1%
Níveis mínimos de confiança: 10%

Neste experimento, não foi possível encontrar regras interessantes devido aos ruídos.

4.4 Vendas e gêneros *(referência à subseção 3.4)*

Neste experimento foi levado em consideração apenas os gêneros e todas as vendas.

Experimento 11:
Níveis mínimos de suporte: 0.03%
Níveis mínimos de confiança: 49%

{Genre = Action} --> {Global_Sales = Medio}
Suporte: 7% Confiança: 50% Lift: 1.07
{Genre = Action} --> {JP_Sales = Baixissimo}
Suporte: 15% Confiança: 89% Lift: 1.15
{Genre = Sports} --> {JP_Sales = Baixissimo}
Suporte: 13% Confiança 88% Lift: 1.13

Neste experimento podemos analisar alguns resultados inconsistentes, sendo completamente diferentes para cada localidade, devido a comparação com a análise feita sem ruídos. A regra {EU_Sales = Altissima } -> {Genre = Sports} e {JP_Sales = Medio alto} -> {Genre = Role-Playing} não foram encontradas na base com ruídos, podendo levar à uma interpretação inconsistente.

4.5 Anos e o gêneros *(referência à subseção 3.5)*

Neste experimento foi levado em consideração apenas os gêneros e todos os anos. Os registros da base foram diminuídos com o intuito de encontrar mais facilmente quais os gêneros preferidos em determinados anos.

Experimento 12:
Total de registros: 1100 (100 registros têm ruído)
Níveis mínimos de suporte: 0,1%
Níveis mínimos de confiança: 10%

{Genre = Action} -> {Year = medio}
Suporte: 7% Confiança: 40% Lift: 1.02

{Genre = Sports} -> {Year = medio}
Suporte: 5% Confiança: 43% Lift: 1.13
{Genre = Misc} -> {Year = medio}
Suporte: 4% Confiança: 54% Lift: 1.43

{Genre = Role_Playing} -> {Year = medio}
Suporte: 4% Confiança: 45% Lift: 1.12
{Genre = Shooter} -> {Year = medio}
Suporte: 4% Confiança: 41% Lift: 1.07
{Genre = Strategy} -> {Year = muito_antigo}
Suporte: 0,8% Confiança: 45% Lift: 3.25
{Genre = Adventure} -> {Year = muito_antigo}
Suporte: 0,1% Confiança: 40% Lift: 1.79
{Genre = Fighting} -> {Year = antigo}
Suporte: 1% Confiança: 33% Lift: 1.42
{Genre = Simulation} -> {Year = antigo}
Suporte: 1% Confiança: 38% Lift: 1.41
{Genre = Shooter} -> {Year = medio}
Suporte: 1% Confiança: 41% Lift: 1.98
{Year = novo} -> {Genre = Action}
Suporte: 1% Confiança: 25% Lift: 1.31

A introdução de ruídos neste experimento não afetou grandemente nas regras obtidas, causando algumas mudanças de confiança 2 ou 3% acima ou abaixo da anterior e baixando alguns níveis de suporte, como é possível observar na ausência de quaisquer regras com suporte igual a 8%. Entretanto não pode-se ignorar tais divergências, a medida que trata-se de

5. CONCLUSÃO

O projeto trouxe várias regras passíveis de serem analisadas. Foi possível analisar que a plataforma PS2 foi a que trouxe mais benefícios para o mundo dos games, sendo um dos fatores para tal evolução. O PS2 predominou como a plataforma mais vendida em todo o período, estando logo atrás o Xbox 360 e o PS3. Na análise dos gêneros, vimos que o Sports predominou na Europa, enquanto que Shooter, Action e Sports foram ruins no Japão. Esta análise possibilita um aumento do marketing de determinado jogo preterido pelo local, além do aumento do desenvolvimento de jogos do gênero não preteridos, visto na comprovada alta venda nesta área.

No experimento 3.3 foi possível analisar quais plataformas foram mais vendidas em determinadas áreas. Uma regra interessante é que o PS3 foi preferido por outros locais do mundo, podendo assim aumentar a divulgação de jogos e consoles a fim de elevar as vendas.

No experimento 3.1, verificamos uma mudança nas preferências das plataformas. Isso mostra como o mundo se adaptou bem ao lançamento dos novos consoles, além dos novos jogos.

No experimento 3.5, verificamos que os gêneros Action, Sports, Role-Playing, Misc e Shooter foram preteridos nos anos médios. Os gêneros Fighting e Simulation foram mais frequentes no período antigo, enquanto que Adventure e Strategy no período muito antigo. Os gêneros Action e Shooter são mais frequentes no período novo. Isso pode ser comprovado com o surgimento de campeonatos destas modalidades. É possível compreender a evolução das preferências dos games ao passar dos anos.

Nos experimentos 3.6 e 3.7 é possível analisar que a Nintendo predominou como a editora mais vendida em todo o mundo por vários anos, principalmente pelo período antigo. Já nos anos mais recentes a Nintendo perdeu para outras editoras como a Capcom, Eletronic Arts e Take-Two Interactive, mostrando a evolução dos jogos.

Nos resultados com ruídos, analisamos que alguns experimentos as regras foram muito inconsistentes e ineficientes. Em outros como no 4.1 e 4.2 os resultados variaram pouco, porém a plataforma PS3 foi considerada baixa nas vendas globais nos experimentos sem ruídos, ao contrário do experimento com ruídos, onde ela foi considerada média. Isso mostra como é importante limpar a base antes de aplicar qualquer técnica de mineração de dados.

Este trabalho possibilitou aos integrantes a análise da importância da mineração de dados e o poder que ela possui na análise e extração de regras. Essas regras podem ajudar muito as empresas, desde o aumento da campanha de marketing até o posicionamento de produtos nas prateleiras e muitos outros fatores.

5 . REFERÊNCIAS

BASU, Chumki et al. Association rule mining to understand GMDs and their effects on power systems. 2016 Ieee Power And Energy Society General Meeting (pesgm), [s.l.], p.1-6, jul. 2016. IEEE. http://dx.doi.org/10.1109/pesgm.2016.7741752. Acesso em 12 de maio de 2018.

Figura 2. "Processo de pré-processamento". **Fonte:** http://www.portal.inf.ufg.br/sites/default/files/uploads/relatorios-tecnicos/RT-INF_001-09.pdf. Acesso em 10 de maio de 2018

HAN, Jiawei; KAMBER, Micheline; PEI, Jian. Data Mining: Concepts and Techniques. 3. ed. Morgan Kaufmann, 2012. (The Morgan Kaufmann Series in Data Management Systems). Acesso em 12 de maio de 2018.

LEE, Dahye et al. Discovering Relationships between Factors of Round-trip Car Sharing by Using Association Rules Approach. In: Would Multidisciplinary Civil Engineering-Architecture- Urban Planning Symposium – WMCAUS, 161., 2016, Praga. Procedia Engineering. Elsevier, 2016, p.1282 - 1288. Acesso em 12 de maio de 2018.

VASCONCELOS, L. M. R. de; CARVALHO, C. L. de. Aplicação de Regras de Associação para Mineração de dados na Web. Goiás: Instituto de Informática Universidade Federal de Goiás,2004. Acesso em 11 de maio de 2018.

SILVA, Luiz Paulo Moreira. "Dimensões do espaço"; Brasil Escola. Disponível em . Acesso em 14 de fevereiro de 2018.

Solange O. Rezende, Ricardo M. Marcacini, Maria F. Moura. "O uso da Mineração de Textos para Extração e Organização Não Supervisionada de Conhecimento". Revista de Sistemas de Informacao da FSMA n. 7 (2011) pp. 7-21