UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ
DAINF - DEPARTAMENTO ACADÊMICO DE INFORMÁTICA
CURSO DE BACHARELADO EM CIÊNCIAS DA COMPUTAÇÃO

GABRIEL LEVIS ZAWALSKI

APLICAÇÃO DO PROCESSO DE DESCOBERTA DE CONHECIMENTO EM UMA BASE DE DADOS DE REDE SOCIAL

TRABALHO DE CONCLUSÃO DE CURSO

PONTA GROSSA
2018

LISTA DE QUADROS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

KDD	*Knowledge Discovery in Databases*

SUMÁRIO

1 INTRODUÇÃO

Segundo o *site* <https://www.internetworldstats.com/stats.htm>, 54% da população mundial está conectada na *internet*, somando um total de mais de 4 bilhões de pessoas *online* no mundo todo. Desse total, segundo pesquisas da (DOSSIER STATISTA) 2.62 bilhões se conectam através de redes sociais e a previsão é de 3.02 até o fim de 2021.

Com a crescente presença de serviços digitais na vida cotidiana da população, como o uso de redes sociais, serviços de *streaming*, lojas e compras *online* há uma quantidade enorme de dados está sendo armazenada a todo momento em todos os lugares do mundo (CITAR DOSSIER STATISTA PARA NUMERO DE USUÁRIOS). Acumular dados não é suficiente por si só, tão importante quanto é extrair informações destes. Porém, analisar os dados de maneira manual requer interpretação por parte do analista, sendo muitas vezes ineficientes, demorados, custosos além de propensos à subjetividade do analista. (REFERENCIAR ESSA PARTE)

(REFERENCIA FAYYAD) diz que neste cenário há uma grande necessidade de novas teorias e ferramentas computacionais a fim de auxiliar na extração de conhecimento dos grandes e cada vez maiores volumes de dados. (FINALIZAR PENSAMENTO)

O processo de Descoberta de Conhecimento em Bases de Dados (*Knowledge Discovery in Databases* - KDD), termo cunhado por (REFERENCIA FAYYAD) na realização do primeiro *workshop* a respeito do tema, foi uma resposta aos problemas causados por esse grande acumulo de dados. (REFERENCIAR KDD)

O KDD é um processo iterativo não trivial dividido em 9 etapas dependentes de interações do usuário. O processo tem por objetivo identificar informações válidas, novas, potencialmente úteis e compreensíveis em um grupo de dados, segundo (REFERENCIA FAYYAD). Essas informações extraídas também precisam ser, até certo grau de certeza, válidas para um novo conjunto de dados.

O processo de KDD envolve desde a compreensão dos dados armazenados na base, sua seleção e pré-processamento, passando por etapas de transformação e aplicação de técnicas de mineração de dados, terminando nas etapas de identificação dos padrões gerados que se enquadram nos requisitos de conhecimento do processo.

Entre as etapas do processo de KDD, a mineração de dados é uma das que possui grande ênfase tanto nas áreas acadêmicas quanto nas áreas práticas devido as grandes quantidades de métodos e resultados disponíveis e testados. O processo de mineração de dados envolve a descoberta de padrões a partir de dados e a adaptação de modelos para melhor acomodar os dados existentes, utilizando-se de técnicas de muitas áreas diversas, como aprendizagem de máquina, reconhecimento de padrões e estatística. (REFERENCIAR ESSA PARTE)

Dentro dos diversos métodos de mineração de dados, sendo os mais comuns os de classificação, clusterização e regressão, existem uma extensa gama de algoritmos, porém todos com funcionamentos e fundamentos parecidos. (REFERENCIA FAYYAD) diz que os

métodos de mineração de dados podem ser compostos de três algoritmos diferentes: modelo de representação, modelo de avaliação, e busca.

Como foi citado anteriormente, somente ter acesso a determinados dados não garante a completa compreensão do problema, porém, com a aplicação de processos como o KDD, pode-se tirar conclusões úteis a respeito dos dados tratados. Este trabalho tem como objetivo utilizar o processo do KDD numa base de dados retirada de uma rede social, a fim de encontrar um modelo de previsão para as métricas de avaliação estabelecidas pelo *Facebook*.

1.1 DESCRIÇÃO DO PROBLEMA E MOTIVAÇÃO

O grande aumento do volume de dados acumulados por diversos tipos de aplicações pode gerar um grande problema, uma vez que nem sempre somente tê-los significa ter algum conhecimento a respeito do problema em questão. Erros podem ser cometidos sem uma devida análise rigorosa e metódica a fim de se extrair informações relevantes, levando a conclusões erradas e na base de achismos.

Aplicar o processo de KDD se torna cada vez mais interessante nesse cenário super-populoso de dados uma vez que estabelece um processo que pode ser seguido e replicado em diversos tipos de situações diferentes. A base de dados utilizada no trabalho foi escolhida por possuir dados de uma situação real, além de serem relativamente recentes e com poucos trabalhos explorando os diversos tratamentos que os dados podem receber, permitindo assim uma grande gama de abordagens diferentes possíveis.

1.2 OBJETIVOS

Esta Seção apresenta o obejtivos geral e os objetivos específicos deste trabalho. Na Subseção 1.2.1 se encontra o objetivo geral e na Subseção 1.2.2 se encontram os objetivos específicos.

1.2.1 Objetivos gerais

O objetivo geral deste trabalho é aplicar o processo de KDD numa base de dados de domínio público a fim de criar um modelo de previsão de sucesso de postagem.

1.2.2 Objetivos específicos

Como objetivos específicos têm-se:

- Compreender o funcionamento do processo de KDD;
- Analisar as etapas do KDD, identificado técnicas que podem ser aplicadas;
- Aplicar o processo de KDD na base de dados;
- Realizar experimentos e analisar os resultados obtidos por meio de comparação estatística com outros trabalhos da área.

1.3 ORGANIZAÇÃO DO TRABALHO

Este trabalho encontra-se dividido nos seguintes sete capítulos:

- Capítulo 1: Capítulo introdutório de contextualização do trabalho, apresentando em linhas gerais a situação atual, a motivação que levou a idealização deste trabalho e os objetivos a serem alcançados.
- Capítulo 2: Este capítulo aborda os conceitos necessários para compreender o processo de KDD. Contém uma descrição das etapas que o constituem como um todo e entra mais a fundo em conceitos importantes da parte de mineração de dados.
- Capítulo 3: O capítulo apresenta conceitos e definições a respeito de redes sociais, classificando-as e descrevendo sua relevância fora e dentro da área acadêmica. Também contém uma explicação a respeito de termos da rede social Facebook, sendo esta o local onde os dados utilizados neste trabalho foram retirados.
- Capítulo 4: Neste capítulo encontra-se uma explicação a respeito do método de revisão metodológica, ressaltando os pontos importantes do processo e a aplicação deste em bases de artigos acadêmicos, permitindo a identificação de trabalhos correlatos que serviram de referência para a elaboração deste.
- Capítulo 5: O capítulo descreve a base de dados utilizada para a realização dos experimentos deste trabalho, bem como as técnicas e métodos a serem aplicados na mesma, seguindo o processo definido no capítulo 2.
- Capítulo 6: Este capítulo contém a descrição e análise dos resultados obtidos após a aplicação dos métodos e técnicas descritos no capítulo 5 sobre a base de dados deste trabalho.
- Capítulo 7: O capítulo contém possibilidades de continuações deste trabalho, bem como as considerações finais a respeito dos resultados obtidos e do cumprimento dos objetivos propostos.

2 DESCOBERTA DE CONHECIMENTO EM BASES DE DADOS

Este capítulo aborda o processo de Descoberta de Conhecimento em Bases de dados, abreviado pela sua silga em inglês KDD. A Seção 2.1 apresenta o histórico e informações básicas a respeito do processo como um todo, oferecendo uma visão geral de suas partes principais e de sua importância. A Seção 2.2 contém uma explicação mais detalhada dos principais agrupamentos de etapas do processo, cada uma explicada nas Seções secundárias 2.2.1, 2.2.2 e 2.2.3. Dentro da Seção secundária 2.2.2 uma Seção terciária 2.2.2.1 entra em mais detalhes a respeito das tarefas associadas à mineração de dados.

2.1 CONCEITOS FUNDAMENTAIS DO KDD

(BASEAR SEÇÃO EM Data Mining A Knowledge Discovery Approach)

KDD é um processo não-trivial de identificação de padrões validos, novos, potencialmente uteis e compreensíveis de dados. (citar KDD unifying framework)

KDD é um processo consistente de diversas etapas.

Os dados passam por tratamentos e técnicas de extração de informação a fim de encontrar informações.

Estas informações precisam atender a requisitos, como serem novas, potencialmente úteis e válidas para vários casos.

O KDD é importante para criar um padrão de análise de dados que pode ser seguido.

Garante que os dados possam ser reutilizados e testes replicados de maneira consistente.

Em suma, extrai uma abstração de alto nível dos dados.

2.2 ETAPAS DO KDD

O processo de KDD envolve vários passos incluindo etapas com input de decisões, ou seja, ele é interativo e iterativo.

O processo de KDD para extração de conhecimento em ambito acadêmico foi definido por Fayyad (citar Data Mining A Knowledge Discovery Approach) e é dividido em 9 etapas.

1. Entendimento do domínio da aplicação
2. Seleção de dados
3. Limpeza dos dados
4. Redução de dados e projeção
5. Escolha da tarefa de mineração
6. Escolha do algoritmo de mineração
7. Mineração
8. Interpretação dos padrões minerados
9. Consolidação do conhecimento extraído

O processo pode conter loops entre quaisquer pontos do processo, porém o fluxo básico é ilustrado na Figura (Figura do processo).

A maioria dos trabalhos focam-se exlusivamente no passo sete (CITAR FAYYAD NISSO), porém todos os passos são importantes.

Para as delimitações deste trabalho, dividiremos os 9 passos do processo em 3 grupos, sendo o primeiro chamado pré-processamento englobando os passos 1 ao 4, o segundo de mineração de dados engolbando os passos 5 ao 7 e o ultimo grupo denominado de pós-processamento engolbando os restantes passos 8 e 9.

2.2.1 Pré-processamento

O pré-processamento engloba dos passos 1 ao 4 do processo completo do KDD.

(Data Mining A Knowledge Discovery Approach) 1. Developing and understanding the application domain. This step includes learning the relevant prior knowledge and the goals of the end user of the discovered knowledge. 2. Creating a target data set. Here the data miner selects a subset of variables (attributes) and data points (examples) that will be used to perform discovery tasks. This step usually includes querying the existing data to select the desired subset. 3. Data cleaning and preprocessing. This step consists of removing outliers, dealing with noise and missing values in the data, and accounting for time sequence information and known changes. 4. Data reduction and projection. This step consists of finding useful attributes by applying dimension reduction and transformation methods, and finding invariant representation of the data.

(KDD unifying framework) 1. Developing an understanding of the application domain and the relevant prior knowledge, and identifying the goal of the KDD process from the customer's viewpoint. 2. Creating a target data set: selecting a data set, or focusing on a subset of variables or data samples, on which discovery is to be performed. 3. Data cleaning and preprocessing: basic operations such as the removal of noise if appropriate, collecting the necessary information to model or account for noise, deciding on strategies for handling missing data fields, accounting for time sequence information and known changes. 4. Data reduction and projection: finding usefid features to represent tile data depending on tile goal of tile task. Using dimensionality reduction or transformation methods to reduce tile effective number of variables under consideration or to find invariant representations for the data.

2.2.2 Mineração de dados

Engloba as etapas de 5 a 7 do KDD e área de grande enfoque de pesquisa científica.

(Data Mining A Knowledge Discovery Approach) 5. Choosing the data mining task. Here the data miner matches the goals defined in Step 1 with a particular DM method, such as classification, regression, clustering, etc. 6. Choosing the data mining algorithm. The data miner selects methods to search for patterns in the data and decides which models and parameters of the methods used may be appropriate. 7. Data mining. This step generates patterns in a

particular representational form, such as classification rules, decision trees, regression models, trends, etc.

(KDD unifying framework) 5. Matching tile goals of tile KDD process (step 1) to particular data mining method: e.g., summarization, classification, regression, clustering, etc. Methods are described in Section 5.1, and in more detail in (Fayyad, Piatetsky-Shapiro, : Smyth 1996). 6. Choosing the data mining algorithm(s): selecting method(s) to be used for searching for patterns the data. This includes deciding which models and parameters may be appropriate (e.g. models for categorical data are different than models on vectors over the reals) and matching a particular data mining method with the overall criteria of the KDD process (e.g., the end-user may be more interested in understanding the model than its predictive capabilities- see Section 5.2). 7. Data mining: searching for patterns of interest in a particular representational form or a set of such representations: classification rules or trees, regression, clustering, and so forth. The user can significantly aid the data mining method by correctly performing the preceding steps.

O item 5 faz a escolha entre as tarefas de mineração a seguir.

Dentro de cada tarefa de mineração existe uma grande quantidade de algoritmos diferentes que realizam a tarefa em específico em condições diferentes e com resultados diferentes.

2.2.2.1 Tarefas da mineração de dados

As tarefas da mineração de dados são divididas em 3 grandes grupos: Agrupamento, Classificação e Regressão

Agrupamento é utilizar de padrões entre os dados para criar conjuntos de elementos parecidos

Classificação é utilizado para identificar novos elementos como pertencentes a grupos já definidos

Regressão consiste em inferir valores de classificação discretos em um intervalo de valores a fim de classificar os dados em uma escala graduada

2.2.3 Pós-processamento

Engolba as etapas restantes do KDD. Envolve a avaliação das informações extraídas dos dados.

(Data Mining A Knowledge Discovery Approach) 8. Interpreting mined patterns. Here the analyst performs visualization of the extracted patterns and models, and visualization of the data based on the extracted models. 9. Consolidating discovered knowledge. The final step consists of incorporating the discovered knowledge into the performance system, and documenting and reporting it to the interested parties. This step may also include checking and resolving potential conflicts with previously believed knowledge

(KDD unifying framework) 8. Interpreting mined patterns, possibly return to any of steps 1-7 for further iteration. This step can also involve visualization of the extracted patterns/models, or visualization of the data given the extracted models. 9. Consolidating discovered knowledge: incorporating this knowledge into another system for filrther action, or simply documenting it and reporting it to interested parties. This also includes checking for and resolving potential conflicts with l)reviously believed (or extracted) knowedge.

3 REDES SOCIAIS

Desde seu surgimento, *sites* de redes sociais, como *Facebook* e *Twitter*, vêm atraindo milhões de usários ao redor do mundo. Segundo estatísticas do site Dossier Statista, em 2017 haviam 2.46 bilhões de usários de redes sociais ao redor do mundo e é estimado que em 2019 esse número suba para 2.77 bilhões.

Tentar definir em termos precisos o que seria uma rede social se prova difícil devido a grande variedade de serviços independentes e integrados de serviços de comunicação. Simplesmente definir como serviços que permitem aproximar pessoas de forma digital se torna uma definição muito ampla.

3.1 CONCEITOS FUNDAMENTAIS

Revisões de literatura permitiram extrair 4 características principais recorrentes nos serviços de *networking* social.

(REFERENCIAR SOCIAL MEDIA DEFINITION PEGAR ARTIGO NO SCIENCE DIRECT: https://www.sciencedirect.com/science/article/abs/pii/S0308596115001172)

1. social networking services are interactive Web 2.0 Internet-based applications,
2. user-generated content (UGC), such as user-submitted digital photos, text posts, "tagging", online comments, and diary-style "web logs"(blogs), is the lifeblood of the SNS organism,
3. users create service-specific profiles for the site or app that are designed and maintained by the SNS organization, and
4. social networking services facilitate the development of social networks online by connecting a user's profile with those of other individuals or groups.

Breve histórico de algumas redes sociais famosas.

Tipos de redes sociais brevemente explicados.

1. *Networking*
2. *Blogging*
3. *Microblogging*
4. *Photo Sharing*
5. *Video Sharing*

Usos diferentes de redes sociais

1. *Real time*
2. *Location based*
3. Mercados de nicho
4. Ciência
5. Educação
    a) Profissional

   b) Currículo
   c) Aprendizado
6. Procuro de emprego
7. Serviços de *hosting*
8. Trocas

3.2 IMPACTOS E RELEVÂNCIA

Alguns benefícios

Alguns problemas

Relevância dos estudos na área

3.3 *FACEBOOK*

Breve histórico *Facebook*.

Facebook utilizado em negócios para empresas.

Explicação das métricas de avaliação de desempenho de postagem pelo *Facebook*. Se encontram no quadro da metodologia

4 REVISÃO SISTEMATICA DA LITERATURA

Revisão dos trabalhos relacionados da área

4.1 MÉTODO DE REVISÃO SISTEMÁTICA

Explicação do método de revisão sistemática

4.2 APLICAÇÃO DO MÉTODO

Resultados da aplicação da revisão sistemática.

Quadro 1 – Bases de dados pesquisadas.

Base de dados	**Sites**
arXiv.org	<https://arxiv.org/>
Emerald Insight	<https://www.emeraldinsight.com/>
IEEEXplore	<https://ieeexplore.ieee.org>
Science Direct	<https://www.sciencedirect.com/>
Springer	<https://link.springer.com/>

Tabela 1 – Resultado das buscas.

Base de dados	**Resultados**
arXiv.org	21
Emerald Insight	4
IEEEXplore	752
Science Direct	266
Springer	902

5 METODOLOGIA

Metodologia, ferramenta e base utilizada na pesquisa.

5.1 PRÉ-PROCESSAMENTO

Tudo que foi aplicado na base de dados.

5.2 MINERAÇÃO DE DADOS

Tudo que foi aplicado para mineração, svm e regressão.

5.3 PÓS-PROCESSAMENTO

Tudo que foi feito após o proc de datamining.

Quadro 2 – Descrição dos atributos da base de dados utilizada no trabalho.

Atributo	**Descrição**
Page total likes	Quantidade de likes totais da página quando feita a postagem.
Type	Tipo de postagem, entre Fotos, *Link*, Status e Vídeo.
Category	Categoria de tipo de propaganda utilizada internamente pela empresa, adicionado de forma manual na base de dados.
Post Month	Mês em que a postagem foi feita, retirado da data da postagem.
Post Weekday	Dia da semana em que a postagem foi feita, retirado da data da postagem.
Post Hour	Hora do dia em que a postagem foi feita, retirado da data da postagem.
Paid	Se o post usou os serviços de anúncios pagos do *Facebook*.
Lifetime Post Total Reach	Quantidade de usuários únicos que a postagem alcançou, independente da forma com que a postagem chegou até o usuário.
Lifetime Post Total Impressions	Quantidade total de vezes que uma postagem foi vista. Esse número pode ser maior que o *Total Reach* pois um mesmo usuário pode ver a postagem diversas vezes.
Lifetime Engaged Users	Quantidade de usuários que clicaram na postagem de forma que geram ou não *Stories*. *Stories* são tipos de interações que fazem com que a postagem seja propagada para outros usuários, como, por exemplo, Compartilhar. Dessa forma essa estatística conta a quantidade total de usuários que clicaram na postagem de uma forma qualquer. (RETIRAR EXPLICAÇÃO DE STORIES DAQUI)
Lifetime Post Consumers	Quantidade de usuários que clicaram na postagem de forma que não geram *Stories*. Essa estatística é diferente da *Lifetime Engaged Users* pois só conta os usuários que clicaram na postagem de forma a não espalhar a postagem, dessa forma contando somente clique dentro do conteúdo em si, como tocar o vídeo ou ampliar a foto, por exemplo.
Lifetime Post Comsumptions	Quantidade total de cliques que não geram *Stories*. Essa estatística conta a quantidade de clique feitos pelos *Post Consumers*, tentando aproximar a quantidade de vezes que a postagem foi consumida, seja quantidade de vezes que o vídeo foi visto ou quantos clicaram no link compartilhado.
Lifetime Post Impressions by people who have liked your Page	Quantidade total de vezes que uma postagem foi vista por usuários que deram *like* na página.
Lifetime Post reach by people who like your Page	Quantidade de usuários únicos que a postagem alcançou, independente da forma com que a postagem chegou até o usuário, porém

6 ANÁLISE E DISCUSSÃO DOS RESULTADOS

6.1 RESULTADOS DO PROCESSO DE KDD

Conhecimento retirado do processo de kdd dos dados

6.2 ANÁLISE DOS RESULTADOS

Comparação dos resultados com os dados reais da base e de oturos trabalhos

6.2.1 Comparativo entre dados reais coletados

6.2.2 Comparativo entre outros trabalhos relacionados

7 CONCLUSÃO

Resultados obtidos da regressão

7.1 TRABALHOS FUTUROS

Continuações do trabalho

7.2 CONSIDERAÇÕES FINAIS

Considerações finais