UNIVERSIDADE DE CAXIAS DO SUL
CENTRO DE COMPUTAÇÃO E TECNOLOGIA DA INFORMAÇÃO
CURSO DE BACHARELADO EM SISTEMAS DE INFORMAÇÃO

ELISANGELA PUHL

ESTUDO DE PERFIL DE USUÁRIOS DE REDES SOCIAIS ATRAVÉS DA MINERAÇÃO DE DADOS

Profª. Draª. Helena G. Ribeiro
Orientadora

Caxias do Sul, Dezembro de 2011

"O conhecimento é o processo de acumular dados;,
a sabedoria reside na sua simplificação."
— Martin H. Fischer

AGRADECIMENTOS

A colaboração de muitas pessoas foi importante para a execução deste trabalho, dessa forma gostaria muito de destacar algumas pessoas em especial.

Agradeço a minha família, com um carinho muito especial a minha mãe, meu pai, meu irmão e minhas tias Lourdes Puhl e Helena Puhl, que sempre acreditaram em mim e nos momentos mais difícies, me ofereceram suporte para minha evolução pessoal e profissional.

Agradeço ao meu namorado pelo apoio e compreensão que teve durante o semestre da execução deste trabalho.

Agradeço a minha orientadora, Helena Grazziotin Ribeiro, pela paciência e que sempre atendeu-me prontamente e me orientou para que realizasse um bom trabalho.

Agradeço a coordenadora do curso, Iraci Cristina da Silveira De Carli, que durante minha graduação sempre foi uma ótima conselheira.

Por fim, gostaria de agradecer a todas as pessoas que de uma forma ou outra se fizeram presentes e contribuíram em algum momento para o cumprimento de minha jornada na universidade. Com certeza, cheguei a este momento com a colaboração de todos vocês.

SUMÁRIO

LISTA DE ABREVIATURAS E SIGLAS

API	*Application Programming Interface*
DBLP	*Digital Bibliography and Library Project*
BDBComp	Biblioteca Digital Brasileira de Computação
CSV	*Comma-separated values*
DCBD	Descoberta do Conhecimento em Banco de Dados
DHP	*Direct Hashing and Pruning*
FURB	Fundação Universitária Regional de Blumenau
GSP	*Generalized Sequential Pattern*
JDBC	*Java Database Connectivity*
JSON	*JavaScript Object Notation*
WEKA	*Waikato Environment for Knowledge Analysis*
KDD	*Knowledge Discovery in Databases*
KDT	*Knowledge Discovery in Text*
RNA	Redes Neurais Artificiais
SDK	*Software Development Kit*
TCC	Trabalho de Conclusão de Curso
XML	*Extensible Markup Language*
XSLT	*eXtensible Stylesheet Language for Transformation*

LISTA DE FIGURAS

LISTA DE TABELAS

RESUMO

Este trabalho tem como objetivo a extração de dados de redes sociais para aplicar sobre eles técnicas de mineração de dados apropriadas com o fim de mapear o perfil dos usuários para auxiliar a área de marketing de uma empresa de software. Ele descreve os conceitos relacionados a estes assuntos, e as principais etapas e como funcionam os algoritmos, métodos e técnicas utilizadas para este tipo de processo. Para corroborar estes conceitos e técnicas, foi realizado um estudo de caso na Procad Softwares que tem uma forte atuação nas redes sociais.

Palavras-chave: Mineração de dados, perfil de usuário, redes sociais.

1 INTRODUÇÃO

Com a crescente concorrência no mercado e o aumento da inclusão digital no Brasil, cada vez mais as empresas estão com suas atenções voltadas para as mídias sociais com o objetivo de estar mais conectados com o seu cliente.

As redes sociais tem sido objeto de estudos nesses últimos anos com o objetivo de analisar as interações entre as pessoas, a tendência e os padrões estruturais. As tendências nos últimos anos concentram-se nas redes sociais on-line, como o *Facebook, Linkedin e Myspace*. Essas redes tem mostrado crescimento, devido a não limitação geográfica comparada as redes sociais convencionais, em que suas interações são em grupos fechados.

A grande expansão das redes sociais on-line tem levado as empresas a mudarem suas estratégias de *marketing* para entender e atender da melhor forma esse cliente com esse novo perfil. Essas companhias tem criado perfis corporativos com o objetivo manter uma proximidade virtual com esse público, lançando promoções em *sites* de compras coletivas. Atualmente, os dados gerados nessas interações, como por exemplo *feedback* do cliente, número de vendas através de *e-commerce* e retorno sobre o investimentos em promoções em sites de compras coletivas são coletados e não estão sendo analisados, impossibilitando que as estratégias de *marketing* sejam direcionadas de forma assertiva.

Há estudos que utilizam técnicas de mineração de dados para a identificação de perfil de usuário em diversas áreas, como em (JUNIOR; PEREZ, 2006), que para identificar o perfil de usuários inadimplentes de serviços de telefonia, foram testadas técnicas de mineração de dados como redes neurais artificiais (RNA) e árvore de decisão. Em (BRESOLIM, 2011) foi utilizada árvores de decisão para a aplicação em um sistema de recomendação, baseado no histórico de acesso a um portal de vídeo locadora.

Mineração de Dados é um ramo da computação que teve início nos anos 80, quando os profissionais das empresas e organizações começaram a se preocupar com os grandes volumes de dados estocados e inutilizados dentro da empresa. Através das Técnicas de Mineração de Dados, que conforme descrito por (AMO, 2011), consiste

essencialmente em extrair informação de gigantescas bases de dados da maneira mais automatizada possível. Atualmente, Mineração de Dados consiste, sobretudo na análise dos dados após a extração, buscando-se por exemplo levantar as necessidades reais e hipotéticas de cada cliente para realizar campanhas de *marketing*.

Analisada a importância da mineração de dados no contexto de estudar o perfil desse novo consumidor, o enfoque deste trabalho se dará em estudar as técnicas de mineração de dados para apoiar o processo de *marketing* nas organizações.

A organização do texto está estruturada conforme descrito a seguir. O capítulo 2 apresenta os conceitos e as técnicas de mineração de dados, as etapas do processo de descoberta do conhecimento e uma introdução a mineração de textos. O capítulo 3 apresenta os conceitos de redes sociais, mídias sociais e gráfico social, e também descreve a relação entre *marketing* e redes sociais. O capítulo 4 será descrito o estudo de caso que foi aplicado na empresa Procad, utilizando a ferramenta WEKA com a base de dados extraída do Facebook. Por fim, no capítulo 5, é apresentada as considerações finais do estudo de caso desenvolvido e os sobre os resultados obtidos, bem como as sugestões de trabalhos futuros.

2 MINERAÇÃO DE DADOS

A mineração de dados surgiu da necessidade do emprego de técnicas e ferramentas que permitem transformar, de maneira inteligente e automática, os dados disponíveis em informações úteis. Em virtude da evolução da tecnologia computacional, a capacidade de coletar e armazenar dados superou a capacidade de analisar e extrair conhecimento destes dados causando o problema da superabundância de dados. A fase de Mineração de Dados caracteriza-se pelo emprego de um algoritmo que busca extrair o conhecimento implícito e potencialmente útil dos dados. A mineração de dados, portanto, é uma descoberta eficiente de informações válidas e não óbvias de uma grande coleção de dados.

2.1 Descoberta de Conhecimento em Banco de Dados (DCBD)

A descoberta do conhecimento em bancos de dados (DCBD), ou *Knowlegde of Discovery in Databases(KDD)* possibilita a descoberta eficiente do conhecimento sobre uma grande coleção de dados. Dentre muitas definições existentes para (KDD), podemos citar (FAYYAD et al., 1996) *"... o processo não trivial de identificar, em dados, padrões válidos, novos, potencialmente úteis e ultimamente compreensíveis"*.

A DCBD é um processo contínuo e cíclico que permite que os resultados sejam alcançados e melhorado ao longo do tempo. As etapas do processo de descoberta do conhecimento compreendem: entendimento do negócio, seleção dos dados, limpeza de dados e pré-processamento dos dados, transformação ou codificação de dados, mineração de dados, avaliação e consolidação do conhecimento, conforme figura 2.1.

Embora as etapas devam ser seguidas na ordem que são apresentadas, segundo (FAYYAD et al., 1996) o processo de descoberta do conhecimento é interativo e iterativo, ou seja, vários decisões são tomadas pelo usuário no decorrer do processo e o mesmo processo podendo ser repetido em alguma etapa.

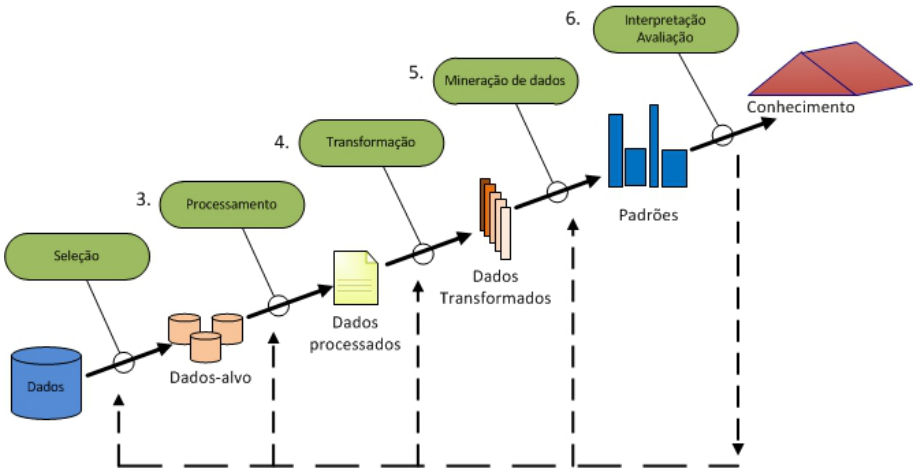


Figura 2.1: Visão geral das etapas do processo de KDD (FAYYAD et al., 1996)

2.1.1 Entendimento do Negócio

O primeiro passo de um projeto de descoberta do conhecimento em banco de dados (DCBD) deve ser sobre o ponto de vista do negócio. O objetivo desta fase é obter uma compreensão mais profunda dos objetivos do projeto e outras circunstâncias do ponto de vista do negócio. A informação resultante desta fase inicial é transformada na definição do problema de mineração de dados.

2.1.2 Seleção dos Dados

Segundo (FAYYAD et al., 1996) nesta fase é escolhido o conjunto de dados que será utilizada para análise.

Este processo pode ser complexo, visto que os dados podem vir de diversas bases diferentes, como por exemplo *data warehouse*, planilhas e outros sistemas e podem possuir diferentes formatos. Esse passo possui impacto significativo sobre a qualidade do resultado do processo.

2.1.3 Limpeza e Pré-Processamento dos Dados

Essa é uma etapa muito importante do processo, visto que a qualidade dos dados irá determinar a eficiência dos algoritmos de mineração de dados. Nesta etapa são realizadas tarefas que eliminam dados redundantes e inconsistentes, recuperam dados incompletos e avaliam possíveis dados discrepantes ao conjunto, ou fora do padrão (*outliers*).

Nesta fase também são utilizados métodos de redução ou transformação para diminuir o número de variáveis envolvidas no processo, visando com isto melhorar o desempenho do algoritmo de mineração (FAYYAD et al., 1996).

2.1.4 Transformação dos Dados

Após os dados serem selecionado, limpos e pré-processados estes necessitam ser armazenados e formatados adequadamente para que os algoritmos de mineração possam ser aplicados.

Esta fase abrange tanto os aspectos sintáticos para aplicação do algoritmo quanto os aspectos semânticos, como registro da tabela, e seleção de atributos. Por último e não menos importante, também pode incluir novos atributos decorrentes que contêm maior quantidade de informação contida implicitamente nos dados (FAYYAD et al., 1996).

2.1.5 Mineração de dados

A mineração de dados consiste na efetiva aplicação do algoritmo escolhido sobre os dados a serem analisados com o objetivo de localizar padrões desejados.

Nos seções seguintes, serão descritos as tarefas desempenhadas e suas técnicas que dão suporte a escolha da aplicação do melhor algoritmo para a resolução do problema.

Conforme afirma (JESUS; MOSER; OGLIARI, 2011), cada técnica de mineração de dados ou cada implementação específica de algoritmos que são utilizados para conduzir as operações de mineração de dados adaptam-se melhor a alguns problemas que a outros, ou seja, os métodos são escolhidos e aplicados com base em que problema necessita-se resolver, não tendo um método universalmente melhor, depende do problema e também do objetivo que se busca.

2.1.6 Avaliação

A avaliação do processo visa garantir que o modelo gerado atenda às expectativas da organização. Os resultados do processo de descoberta do conhecimento podem ser mostrados de diversas formas. Porém, estas formas devem possibilitar uma análise criteriosa para identificar a necessidade de retornar a qualquer um dos estágios anteriores do processo de mineração.

2.1.7 Consolidação do Conhecimento

Incorporar o conhecimento, documentar e reportar as partes interessadas, isto também inclui a resolução dos problemas com o conhecimento obtido, esta é a definição dessa fase por (FAYYAD et al., 1996).

A criação do modelo não é o fim do projeto. Mesmo se a finalidade do modelo for apenas aumentar o conhecimento dos dados, o conhecimento ganho necessitará ser organizado e apresentado em uma maneira que o cliente possa usar. Dependendo das exigências, a fase de execução pode ser tão simples quanto a geração de um relatório, ou tão complexo quanto executar processos de mineração de dados repetidamente.

2.2 Tarefas da Mineração de Dados

O termo "minerar"significa extrair (minérios) da mina. No caso de uma base de dados o termo "minerar dados"significa extrair o conteúdo, conhecimento de um conjunto de dados. Em geral, essa extração tem como objetivo descrever e prever o comportamento futuro de um fenômeno. Descrever tem como foco encontrar algo que faça sentido e que consiga explicar os resultados ou valores obtidos em determinados dados ou negócios. Prever, por outro lado, tem como foco antecipar o comportamento ou o valor futuro de um fenômeno ou variável de interesse, com base no conhecimento de valores do passado.

Na busca de tais objetivos, diferentes estratégias podem ser utilizadas para minerar as bases de dados disponíveis. Portanto, a mineração é a etapa da DCBD que consiste na aplicação de algoritmos específicos que extraem padrões a partir desses dados (FAYYAD et al., 1996). Dessa forma, de acordo com o objetivo da extração do conhecimento deve-se escolher o algoritmo apropriado para o processo. As principais tarefas da mineração de dados são: associação, classificação, regressão, análise de agrupamentos, sumarização e detecção de desvio.

2.2.1 Associação (*association*)

A associação é uma tarefa que tem como objetivo encontrar relacionamentos ou padrões frequentes que ocorrem em um conjunto de dados, ou seja é a busca por itens que ocorrem de forma simultânea. Esta estratégia é muito usada nas redes varejista para definir o perfil do consumidor e distribuição dos produtos nas gôndolas. Por exemplo, uma vez observado que dois produtos são frequentemente adquiridos juntos, muito provavelmente estarão próximos, exemplo cervejas e salgadinhos.

Na mineração de textos, termos simultâneos e frequentes podem auxiliar na remoção de ambiguidades e na caracterização de contextos. Um exemplo desta tarefa é na área de atendimento ao cliente, considerando-se que após o lançamento de um novo produto diversos clientes utilizaram o site da empresa para relatar seu nível de satisfação com tal produto. Palavras frequentes e simultâneas podem auxiliar na identificação de pontos fortes e fracos do produto, levando a novas ações de marketing ou até mesmo a reformulação da produção. Algoritmos tais como o Apriori, GSP, DHP, entre outros, são exemplos que implementam essa tarefa (BEZERRA; GOLDSCHMIDT, 2006).

2.2.2 Classificação (*classification*)

A classificação é o processo de encontrar um conjunto de modelos (funções) que descrevem e distinguem classes ou conceitos, com o propósito de utilizar o modelo para predizer a classe de objetos que ainda não foram classificados (AMO, 2011).

Estudos de caso pesquisados para esse trabalho demonstram que a classificação é tarefa essencial quando se trata de mapear o perfil de um cliente como no caso do estudo de perfil de usuários de serviços de telefonia em (JUNIOR; PEREZ, 2006) e no estudo de caso com o objetivo de mapear o comportamento do usuário da Web em (BRUSSO, 2010).

Como exemplo da tarefa de classificação, podemos citar uma empresa que deseja separar notícias em função do segmento ao qual pertençam (esportes, políticas, economia, etc). Uma aplicação da tarefa de classificação consiste em descobrir uma função que mapeie corretamente os textos, a partir do seu conteúdo em uma dessas classes, tal função uma vez descoberta, pode ser utilizada para prever a alocação de novos textos recebidos. Na implementação desta tarefa podem ser utilizados algoritmos tais como: Rede Neurais Artificiais, Algoritmos Genéricos e Lógica Indutiva (BEZERRA; GOLDSCHMIDT, 2006).

A figura 2.2 demonstra a tarefa de classificação, onde está classificando em conjuntos de dados distintos empréstimos e a não ocorrência de empréstimo (FAYYAD et al., 1996).

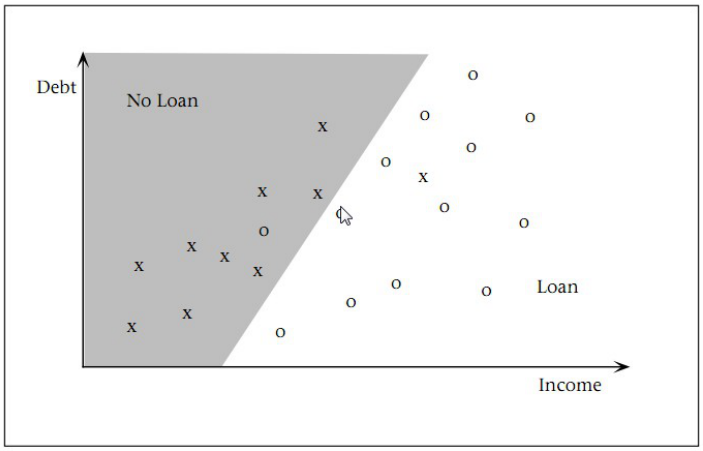


Figura 2.2: Representação da tarefa de classificação.

2.2.3 Regressão

A tarefa de regressão é semelhante à classificação, na qual a regressão analisa valores numéricos tendo como principal objetivo apresentar uma previsão a partir dos dados históricos contidos em uma base de dados, ou seja, compreende a busca por uma função que mapeie os registros de uma banco de dados em valores reais. Estatística, Redes Neurais, dentro outras áreas, oferecem ferramentas para implementação da tarefa de regressão (FAYYAD et al., 1996).

A figura 2.3 demonstra a tarefa a representação da tarefa de regressão linear, onde através de valores numéricos é traçada uma função linear que separa os dados por valores. (FAYYAD et al., 1996)

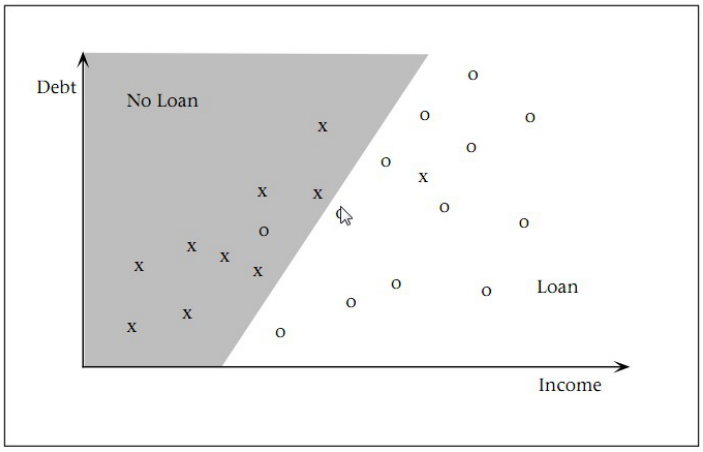


Figura 2.3: Representação da tarefa de regressão linear.

2.2.4 Análise de Agrupamentos ou Clusterização

A análise de agrupamentos consiste na busca de similaridade entre os dados tal que permita definir um conjunto finito de classes ou categorias que os contenha e os descreva. A principal diferença dessa abordagem para a classificação é que na análise de agrupamento não se tem o conhecimento prévio sobre o número de classes. Descobrir grupos homogêneos de clientes é uma das possíveis aplicações e pode ser usada para ajudar na definição da estratégia de marketing a ser adotada (FAYYAD et al., 1996).

A mineração de textos é utilizada para separar os documentos de uma base de textos em subconjuntos ou *clusters*, de tal forma que os documentos de um *cluster* compartilhem de propriedades comuns que os distingam de documentos em outros *clusters*. O objetivo nesta tarefa é maximizar a similaridade intracluster e minimizar a similaridade intercluster. Por exemplo: uma escola pode realizar um processo de clusterização de sua base de documentos para agrupar aqueles que compartilhem o mesmo perfil de conteúdo (BEZERRA; GOLDSCHMIDT, 2006).

A figura 2.4 mostra um gráfico exemplificando o processo de análise de agrupamentos, com a separação de três *clusters* ou subconjuntos conforme sua distribuição ou similaridade.

Na implementação desta tarefa podem ser utilizados algoritmos tais como: K-Means, K-Modes, K-Prototypes, K-Medoids, Kohonen, dentre outros (BEZERRA; GOLDSCHMIDT, 2006).

2.2.5 Sumarização

A sumarização consiste em procurar identificar e indicar características comum entre conjuntos de dados. Esta tarefa é muito comum em mineração de textos. Como exemplo, pode-se considerar um banco de textos da tarefa anterior de clusterização,

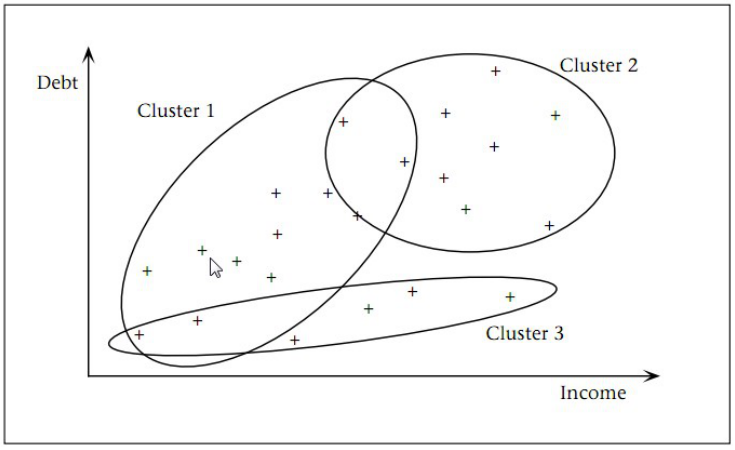


Figura 2.4: Representação da análise de agrupamentos ou clusterização.

uma prática usual é utilizar um algoritmo de sumarização de textos que permita descrever de forma resumida o conteúdo dos documentos em cada cluster (BEZERRA; GOLDSCHMIDT, 2006).

2.2.6 Detecção de Desvios

A detecção de desvios consiste em procurar identificar conjunto de dados cujas características sejam divergentes de um conjunto de dados. Esses dados são denominados *outliers* (GOLDSCHMIDT; PASSOS, 2007).

2.3 Métodos de Mineração de Dados

Diversos métodos de mineração de dados podem ser utilizados nas diferentes técnicas de mineração de dados. Será descrito a seguir os dois que conforme os estudos de caso pesquisados podem ser utilizados para a aplicação no estudo de caso.

2.3.1 Árvore de Decisão

O método de Árvore de Decisão é baseado em probabilidades condicionais que geram regras. O mecanismo de funcionamento de uma árvore de decisão é uma estrutura de árvore ou fluxograma onde: cada nó interno é um atributo de banco de dados de amostra, diferente do atributo-classe, as folhas são valores do atributo-classe, cada ramo ligando um nó-filho a um nó-pai é etiquetado com um valor do atributo contido no nó-pai. Existem tantos ramos quantos possíveis valores para este atributo. Um atributo que aparece num nó não pode aparecer em seus avós descendentes.

A figura 2.5 exemplifica uma árvore de decisão que procura descobrir se uma pessoa é rica ou não com base nos seus preditivos, se possui doutorado e tem mais

de 30 anos é rica.

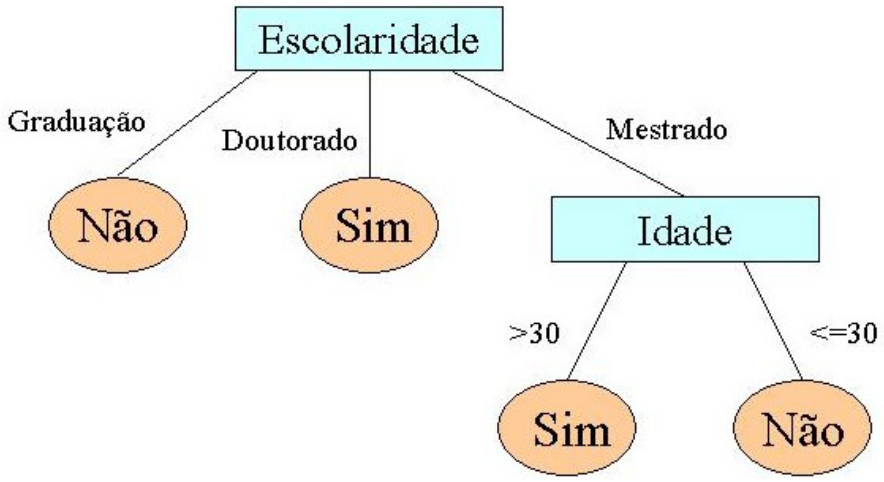


Figura 2.5: Representação de análise de árvore de decisão.

2.3.2 Redes Neurais Artificiais

As redes neurais são sistemas computacionais formados pela integração de inúmeros elementos de processamento, funcionalmente muito simples, altamente interconectados e trabalhando paralelamente. Concebidas com base no estudo do cérebro humano, redes neurais são totalmente diferentes de outras técnicas. São programas que implementam detecções sofisticadas de padrões e algoritmos de aprendizado de máquina, para construir modelos, principalmente, de prognóstico de grandes bancos de dados. Existem duas estruturas principais: nó, que corresponde ao neurônio e o link que corresponde as conexões entre os neurônios.

No artigo (MACEDO; MATOS, 2010) cita Simon Haykin para definir redes neurais:

"Uma rede neural é um processador maciçamente paralelamente distribuído constituído de unidades de processamento simples, que têm a propensão natural para armazenar conhecimento experimental e torná-lo disponível para uso. Ele se assemelha ao cérebro em dois aspectos":

1. O conhecimento é adquirido pela rede a partir de seu ambiente através de um processo de aprendizagem;

2. Forças de conexão entre neurônios, conhecidos como pesos sinápticos, são utilizados para armazenar o conhecimento adquirido.

A figura 2.6 mostra as camadas que podem ser geradas num processo de rede neural. As camadas intermediárias representam os diferentes níveis do conhecimento que são adquiridos no seu processamento, na tentativa de se aproximar do cérebro humano.

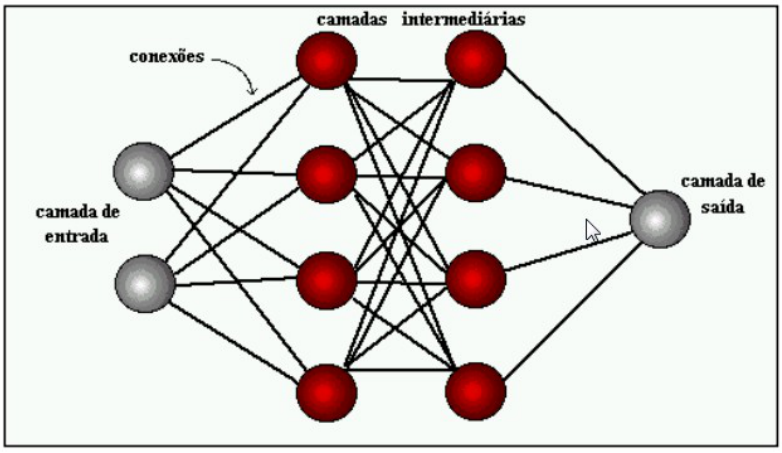


Figura 2.6: Representação de uma rede neural.

2.4 Mineração de Textos

A Mineração de Textos, ou *Text Mining*, procura abstrair padrões, relações e regras (conhecimentos) a partir de dados textuais. Esses dados textuais podem se apresentar em linguagem natural (inglês, português) ou em formato semi-estruturado, como por exemplo, documentos em XML, documentos de e-mail, etc.

Segundo (BEZERRA; GOLDSCHMIDT, 2006), há três tipos de processamento quando se trata de mineração de textos, que são:

- Processamento Léxico e Sintático: Envolve o reconhecimento de *tokens*(termos), a normalização de termos e a construção de linguagem;
- Processamento Semântico; Envolve a extração do significado inerente aos textos. Requer a extração de entidades nomeadas tais como nomes de pessoas, nomes de organizações, locais, etc.
- Processamento de Características Extra Semânticas: Mais complexo, envolve a identificação de sentimentos nos textos analisados Por exemplo: sarcasmo, melancolia, alegria, etc.

2.5 Estudos de Casos com Mineração de Dados

Foram pesquisados alguns estudos de caso que utilizam de técnicas de mineração de dados para mapear o perfil de usuários, abaixo uma breve descrição da problemática desses estudos e as técnicas aplicadas.

No artigo (JESUS; MOSER; OGLIARI, 2011) foi desenvolvido um sistema de recuperação e disseminação de informações, personalizado segundo cada perfil de usuário da Biblioteca Central da Universidade Regional de Blumenau (FURB). Para a extração do conhecimento nos dados da biblioteca sobre o perfil de usuários foi

utilizado a técnica de classificação e após feito isso os dados foram armazenados em um *Data Mart*.

No estudo de caso (JUNIOR; PEREZ, 2006) tem como objetivo mapear o perfil de usuários inadimplentes nos serviços de telefonia, para esse estudo após a limpeza dos dados, estes foram codificados para criar padrões. Foram testados dois métodos: Árvore de Decisão e Redes Neurais. Segundo o autor, os métodos utilizados mostraram-se eficientes na realização do trabalho, porém apresentaram alguns detalhes consideráveis.

A técnica de árvore de decisão apresentou dados que indicam um padrão de comportamento, porém gerou um grande número de subdivisões tornando a leitura do resultado pouco ágil, mas de fácil compreensão. Também permite a geração das regras que definiram o padrão, isso facilita a busca por novos padrões em um novo volume de dados submetidos a esta técnica.

As RNAs apresentaram um resultado um pouco opaco não dando condições de avaliar a técnica e quais os critérios utilizados para chegar ao resultado final. Desta forma, não se pode especificar qual o critério que definiu o padrão dos usuários inadimplentes, isso é de conhecimento somente da RNA.

No artigo (FREITAS; NEDEL, 2008) apresenta um estudo de caso de extração do conhecimento e análise visual de redes sociais, o objetivo deste artigo é apresentar um visão geral dos problemas envolvidos na área de descoberta de conhecimento e visualização de informações de redes sociais. Os dados utilizados foram extraídos das bibliotecas digitais BDBComp Biblioteca Digital Brasileira de Computação e DBLP *(Digital Bibliography and Library Project)*.

O processo foi composto por quatro componentes principais: Gerenciamento dos Dados - os dados foram estruturados em XML; Descoberta do Conhecimento - foram utilizadas a descrição e predição, os padrões descritivos são classificados em agrupamentos, regras de associação e padrões sequencias e os padrões preditivos foram utilizadas a regressão e classificação; Aprendizagem de Máquina - em conjunto com as técnicas de descoberta de conhecimento, este componente atua racionalmente no ambiente com o objetivo de incrementar o desempenho em tarefas futuras; Técnicas de Visualização - técnicas de visualização de informações que permitem a análise visual e interativa das redes sociais apoiadas pelos processos de descoberta de conhecimento e aprendizagem de máquina.

3 AS REDES SOCIAIS

As redes sociais são objeto de estudo há muitos anos, onde são encontrados trabalhos publicados por Jacob Moreno em 1934, no qual ele propõe estudar a maneira como se conectam as pessoas e a que grupo pertence, descreve lugares de centralidade pelo qual algum membro é diferenciado do grupo e propõe formas sociogramas. Segundo (MENESES, 2007), as redes sociais são um sistema aberto em permanente construção, que se constroem individual e coletivamente. Utilizam o conjunto de relações que possuem uma pessoa e um grupo, e são fontes de reconhecimento, de sentimento de identidade do ser, da competência, da ação. Estão relacionadas com os papéis desempenhados nas relações com outras pessoas e grupos sociais, constituindo-se nas práticas sociais que no cotidiano não se aproveitam em sua totalidade.

3.1 As Redes Sociais Online

As redes sociais online são serviços baseados na web que permitem que indivíduos criem perfis públicos ou semi públicos dentro de um sistema limitado, tenham uma lista com outros usuários para compartilhar informação e pesquisar sua lista de conexões e as executadas por outros usuários.

De acordo com a definição acima, o primeiro site de rede social foi lançado em 1997, chamado de SixDegrees.com que permitia que os usuários a criassem perfis, lista de amigos e, a partir de 1998 pudessem navegar na lista de amigos De 1997 até 2001, uma série de ferramentas de comunidades começou a apoiar várias combinações de perfis. O *AsianAvenue, BlackPlanet, e MiGente* permitiam que os usuários criassem perfis pessoais e profissionais, os usuários podiam buscar amigos através de seus perfis sem a aprovação dos mesmos. Após tiveram outros sites como*LiveJournal*, onde as pessoas marcavam seus amigos para seguir seus diários e gerenciar sua privacidade.

A próxima onda de sites de redes sociais começou quando o Ryze.com foi lançado em 2001 para ajudar as pessoas a alavancar suas redes de negócios. O fundador da

Ryze introduziu pela primeira vez o site para seus amigos, principalmente os membros da sua empresa em São Francisco e a comunidade tecnológica, incluindo os empresários e investidores de muitos sites de redes sociais. Havia também o *Tribe.net, Linkedin, Friendster* que estavam unidos em um objetivo comum. No final, a Ryze não adquiriu popularidade, a *Tribe.net* cresceu para atrair um público de nicho específico (tribos), o *Linkedin* se tornou um serviço de negócio poderoso e o *Friendster* se tornou uma das maiores decepções da história da internet, visto que havia conseguido conquistar um grande número de usuários, mas acabou fracassando devido a problema de infraestrutura e com questões relacionadas com políticas de utilização, o que acabou causando descontentamento por parte dos usuários (BOYED; ELLISON, 2007).

O sucesso do *Friendster* fez com que várias empresas surgissem buscando aproveitar a onda de fascínio que as redes socais haviam criado, foi a partir de 2003 que uma nova indústria estava sendo construída. Foi nesse ano que como um clone do *Friendster* surgiu o *MySpace*, que tinha uma proposta mais aberta que o *Frindster*.

Como uma rede privada de estudantes de universidade de elite, em 2004 foi lançado o *TheFacebook*, enquanto isso também estava sendo lançado o *Orkut*, que no início prosperou nos Estados Unidos e se manteve firme diante do *MySpace*. Atualmente o *Orkut* é utilizado na maioria por brasileiros, e perdeu espaço no mercado americano.

Hoje em dia, as redes sociais estendem-se por todo o planeta. O *Facebook* é a maior dentre eles. Começando com *SixDegrees*, passando pelo Friendster até o Facebook, as redes sociais tornam-se uma parte familiar e onipresente da internet (KIRKPATRICK, 2011).

A figura 3.1, apresenta uma linha do tempo com o ano que cada rede social foi lançada.

3.2 Rede Social, Mídia Social e Gráfico Social

Segundo o autor (TREADAWAY; SMITH, 2010), é importante dividir os termos a seguir: mídia social, rede social e gráfico social.

O termo mídia social refere-se à coleção de tecnologias que captura o conteúdo da comunicação entre o indivíduo e seus amigos ou rede social. Exemplo de sites de redes sociais como *Facebook* e *Twitter*, *blogs* como *TypePad* e *Wordpress*, compartilhamento de foto e vídeo como *Flickr* e *YouTube*. Estas tecnologias ajudam os usuários a criar facilmente conteúdos na Internet e compartilhá-los com os outros.

Rede social são grupos de pessoas ou comunidades, que compartilham de um interesse comum. Estas podem ter o contexto de online, como o *Facebook*, bem como o de *offline*, como um grupo de religião que se reúne todos os finais de semana.

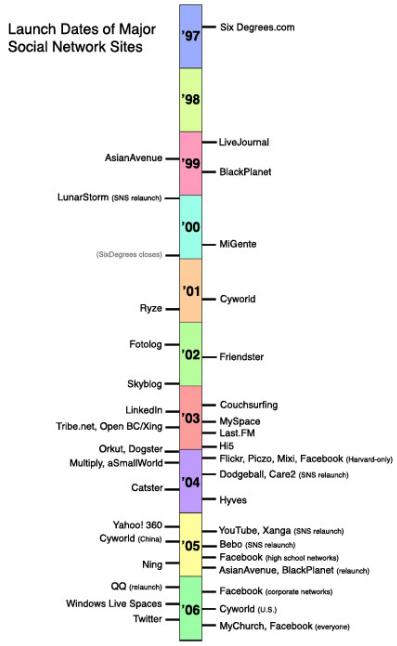


Figura 3.1: Linha do tempo das Redes Sociais

Portanto, exemplos como quem frequenta a Universidade de Caxias do Sul em 2004, pessoas que gostam de surf ou brasileiros na Califórnia, essas redes independente de haver ou não compartilhamento de informação entre os indivíduos da mídia social.

E o gráfico social é um conjunto amplo de pessoas, lugares e interesses que nos tornam indivíduos. Nesse caso o conceito é o que somos, definido pelo que sabemos, as associações que fizemos ao longo dos anos. Antes das mídias sociais, informações sobre nosso gráfico social eram difícies de encontrar. Mídias sociais nos mantém conectados com nossos interesses, com nosso passado e com nossos amigos.

A figura 3.2 ilustra como todos esses conceitos se encaixam.

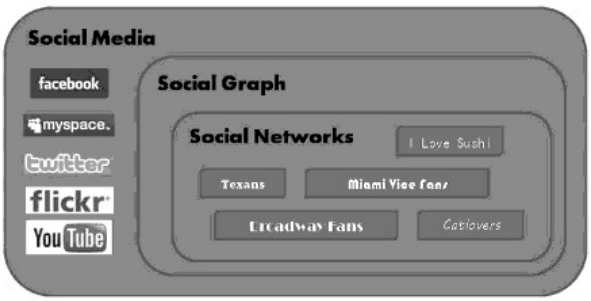


Figura 3.2: Mídia Social, Rede Social e Gráfico Social

3.3 O Facebook

O *Facebook* é atualmente a rede social com maior número de usuários no mundo, através de sua plataforma aberta propicia que desenvolvedores possam acessar suas

API e agregar em seus serviços a visibilidade que a rede social proporciona. Na seção 3.3.1 será explicado o funcionamento de uma API que permite o acesso e a extração dos dados, e também a estrutura da rede social quanto aos dados. Essas informações foram extraídas do site do *Facebook*.

3.3.1 *API Graph*

O API Graph apresenta um visão simples e consistente do gráfico social do *Facebook*, que representam objetos de maneira no gráfico e as conexões entre eles. Cada objeto do gráfico social tem um ID Único, que pode ser acessados através do https://graphs.facebook.com/ID. Todas as respostas retornam um objeto *JavaScript Object Notation* (JSON).

3.4 Redes Sociais x *Marketing*

Atualmente as empresas estão voltadas para seus planos de marketing, onde não podem mais se limitar a atingir seu público-alvo com mídias tradicionais, como rádio e televisão. Percebe-se que a rede de amigos e conexões de nosso círculo social é muito mais honesta que empresas que querem vender seu produto, então há um apelo muito forte em ouvir essas comunidades.

Da mesma forma, essas mesmas tecnologias crescem consideravelmente fazendo com que os planos de marketing estejam sendo reformulados. Segundo (TREADAWAY; SMITH, 2010) segue uma definição de alguns itens conduzirão as mídias sociais e o marketing para os próximos anos.

Necessidade de compartilhar informação. O aumento da utilização de mídia social aumentou devido a necessidade das pessoas compartilharem informações, hoje todos podem transmitir informações positivas ou negativas sobre a compra de um marca ou a experiência com uma empresa. O marketing boca-a-boca tornou-se uma oportunidade e uma ameaça para o marketing moderno.

Imediatismo. Todas as ferramentas de mídia social dão a oportunidade de resposta imediata e compartilhada para seus clientes e sua rede de amigos, portanto da mesma foram que uma experiência positiva ou negativa pode ser imediatamente espalhada na rede. Esses casos podem trabalhar a favor de uma marca ou produto, como também pode trabalhar contra a marca. Algumas empresas já estão buscando formas de se comunicar com esse clientes satisfeitos e insatisfeitos em tempo real através das mídias sociais.

Nível de ruído. Como todos são uma fonte de publicação, faz com que as empresas ao lançarem algum produto necessitem de um esforço maior para manter um nível elevado do mesmo.

Segundo (DRAMBROS; REIS, 2008), o ambiente virtual é um espaço que per-

mite o armazenamento, a busca e a divulgação de conteúdo de forma rápida e barata, o que torna as comunidades um repositório de opiniões, experiências e conhecimentos. Tal característica resulta na criação de capital intelectual e também de informações de marketing, que aumentam o seu valor tanto para seus membros como para as empresas.

Visando os assuntos tratados anteriormente, que as empresas estão voltadas para as mídias sociais com planos de marketing específicos e buscando cada vez mais conhecer o perfil de seus clientes e manter a comunicação direta.

4 ESTUDO DE CASO

Nos dias atuais, a informação e o conhecimento são considerados um bem intangível para as organizações, e para a área de *marketing* o contato com o cliente pode gerar uma base de informações que podem ser utilizadas para auxiliar nas campanhas de *marketing* de relacionamento. Conhecer o cliente e quem frequenta suas mídias sociais é fundamental para a empresa.

Neste capítulo estão detalhadas todas as etapas do estudo de caso, desde a definição da base de dados e as fases da descoberta do conhecimento que são: seleção, limpeza, transformação, mineração e avaliação dos resultados.

4.1 Definição dos Dados para Análise

Os dados utilizados nesse trabalho são informações dos perfis do "Facebook da Procad". Para um melhor entendimento, nesta seção será apresentada a empresa, a forma de extração dos dados e a ferramenta que será utilizada para a mineração dos dados.

4.1.1 Procad Softwares Ltda.

A Procad Softwares Ltda foi fundada no dia primeiro de julho de 1994 pelos sócios os Srs. André Pivoto, Edson Witt e Vanderlei Buffon, com o foco no setor moveleiro, é líder absoluta no mercado brasileiro no ramo de software para projetos de ambientes e vem consolidando sua marca no mercado mundial com mais de 60.000 licenças distribuídas por mais de 30 países. Atualmente possui uma rede de filiais e revendas nas principais capitais brasileiras e no exterior.

Atuando como principal fornecedor de tecnologia do setor moveleiro, a Procad é uma das empresas que investe na melhoria de meios e processos com intuito de auxiliar o setor na continuidade de sua expansão.

A Procad possui um perfil no Facebook desde 2009, e no momento da extração dos dados em Outubro de 2011 possuía 1314 perfis conectados.

4.1.2 Extração dos Dados do Facebook

O Facebook possui uma API para leitura e escrita de dados em seu núcleo. Ela é chamada de *API Graph* . Com ela é possivel, de maneira simples, requisitar informações como o relacionamento entre usuários, suas fotos, gostos, eventos, páginas entre outras coisas. Para que uma aplicação consiga acessar as funcionalidades da *API Graph* é necessário passar por um processo de autorização. A aplicação só poderá consultar informações de um usuário.

Todos os retornos da API são feitos no formato JSON. Um formato para representação textual de dados de maneira leve e simplificada, baseado na sintaxe do JavaScript. Para facilitar o uso da API, o Facebook dispõe de alguns Kits de Desenvolvimento (SDKs). Existem kits para as linguagens de programação JavaScript, PHP e Python e para os sistemas operacionais,de smartphones, iOS e Android.

Para esse estudo de caso os dados foram extraídos através do SDK PHP disponibilizado pelo Facebook com o formato do arquivo em XML.

4.1.3 Ferramenta de Mineração de Dados - WEKA

O WEKA (*Waikato Environment for Knowlegde Analysis*) é um software desenvolvido pela Universidade de Waikato na Nova Zelândia, site www.cs.waikato.ac.nz/weka. Atualmente o *Weka* encontra-se licenciado ao abrigo da *General Public License* sendo portanto possível estudar e alterar o seu código fonte. O objetivo da ferramenta é agregar algoritmos provenientes de diferentes abordagens na subárea da inteligência artificial dedicada ao estudo da aprendizagem de máquinas (mineração) e de comparação de resultados gerados pela aplicação de diferentes técnicas. (WAIKATO, 2011).

O WEKA possui quatro diferentes implementações de *interface*, onde seus algoritmos podem ser chamados diretamente via código Java. As *interfaces* são:

- *Explorer:* ambiente gráfico para a execução dos algoritmos, é a interface de utilização mais comum, e enquadra separadamente as etapas de pré-processamento (filtros), mineração de dados(associação, clusterização e classificação) e pós-processamento (apresentação dos resultados).

- *Experimenter:* ambiente de experimentação, em que testes estatísticos podem ser conduzidos a fim de avaliar o desempenho de diferentes algoritmos de aprendizado.

- *Knowledge Flow:* mostra de forma gráfica o funcionamento interno do programa.

- *Simple CLI:* interface por linha de comando.

O Weka possui implementados diversos métodos, como o de classificação, clusterização e associação. Pode ser realizada de forma simples e rápida a inclusão ou remoção de novos métodos, tornando a ferramenta customizável.

Para cada método são implementados diversos algoritmos, conforme demonstrado nas figuras 4.1 e 4.2. Para o método de associação o principal algoritmo implementado é a Apriori.

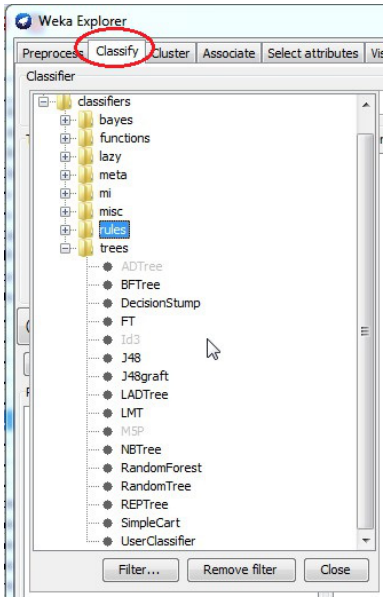


Figura 4.1: Algoritmos de classificação.

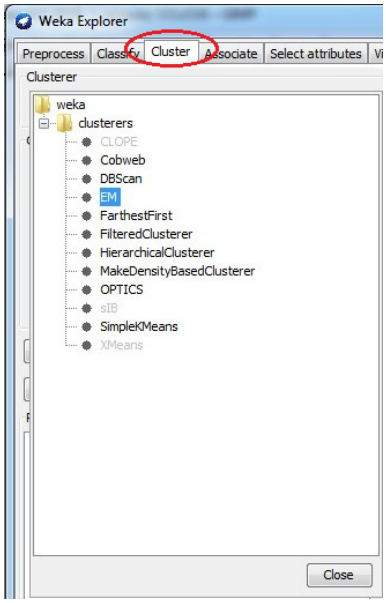


Figura 4.2: Algoritmos de clusterização.

O WEKA suporta arquivos com extensões no formato de ARFF que é o formato padrão, e também os formatos CSV e C45. Ele permite a visualização dos dados em forma de histogramas, e a apresentação de resultados através de árvore de decisão ou regras, entre outros.

A figura 4.3, demonstra um resumo das principais características do WEKA. (GOLDSCHMIDT; PASSOS, 2007).

Na seção seguinte serão mostrados as etapas do processo de transformação do arquivo original dos dados para o formato *.csv que pode ser utilizado pelo Weka.

4.1.4 A Base de Dados

Os dados retirados do *Facebook* através da *API Graph* estão em formato XML, onde cada instância é um item de um *array*. Como o perfil da Procad possuía no momento da extração mais de 1300 perfis, o processo executado gerou um arquivo XML de 7.379 Kb, e neste arquivo único continham os arquivos XML encadeados cada um com um *array* de 0 a 99. Após fazer o processo de separação resultou em 13 arquivos XML.

A figura 4.4 apresenta parte do arquivo XML extraído do Facebook.

Características		Valores
Acesso a Fontes de Dados Heterogêneas		Sim
Integração de Conjuntos de Dados		Não
Facilidade para Inclusão de Novas Operações		Sim
Facilidade para a Inclusão de Novos Métodos		Sim
Recursos para Planejamento de Ações		Sim
Processamento Paralelo/Distribuído		Não
Operações/Métodos Disponíveis	Visualização de Dados	Distribuição de Freqüências; Medidas de Dispersão; Histogramas
	Redução de Dados	Amostragem
	Limpeza de Dados	Substituição
	Codificação dos Dados	Discretização automática e manual
	Classificação	Árvores de Decisão, Bayes, Redes Neurais...
	Clusterização	Simple-KMenas, Cobweb, FarthesFirst...
	Simplificação de Resultados	N/D
	Organização dos Resultados	Agrupamentos de Padrões; Ordenamento de Padrões
	Apresentação dos Resultados	Conjunto de Regras; Árvores de Decisão
Estruturas para Armazenamento de Modelos de Conhecimento		Sim
Estruturas para Acompanhamento de Históricos de Ações		Sim

Figura 4.3: Características do WEKA. (GOLDSCHMIDT; PASSOS, 2007)

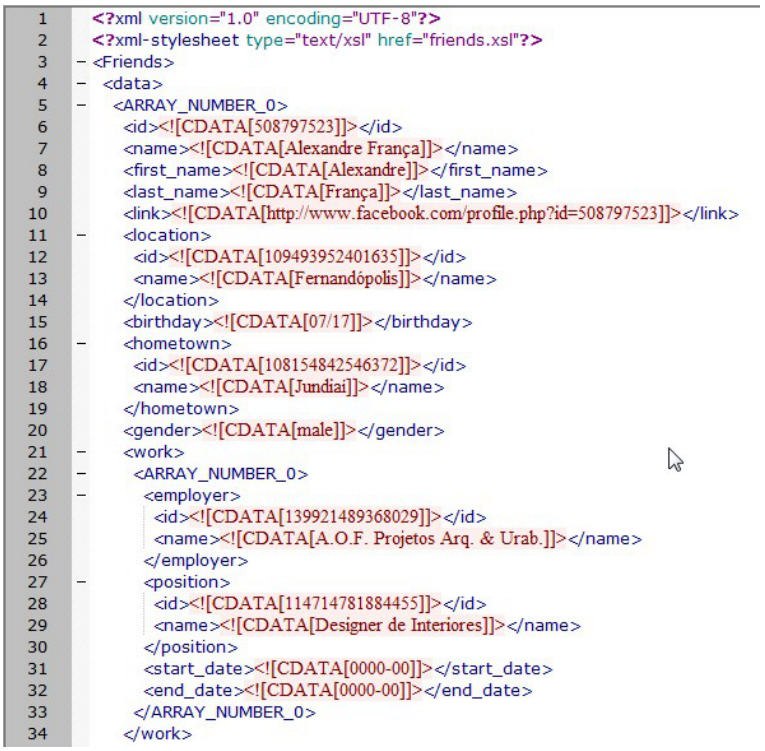
O arquivo XML extraído, contém em média 15 atributos ou *tags* por perfil, no processo de formatação foram selecionados apenas alguns atributos relevantes para serem minerados. Para a execução do processo de separação e formatação foi utilizado um código XSL, que teve a função de selecionar os atributos escolhidos e separar por ';', conforme mostra a figura 4.5 para formar a base de dados *.csv.

4.2 Pré-Processamento dos Dados e Seleção dos Dados

A etapa de pré-processamento que tem funções de seleção, limpeza e a transformação dos dados, tem o objetivo de determinar uma representação adequada para os algoritmos de mineração a partir da base de dados. Os dados não tratados e mal organizados geram uma baixa taxa de aproveitamento pelo algoritmo de classificação, clusterização e associação que serão abordados para mostrar as relações dos dados na ferramenta. Essa etapa é considerada de grande importância para a etapa de mineração de dados.

4.2.1 Seleção dos Dados e Limpeza dos Dados

Por ser uma base extraída de uma rede social, muitos atributos estão sem resposta ou com dados não estruturados, por exemplo, no campo *'location'* foi respondido como "Caxias do Sul, RS, Brazil","Caxias, RS, Brazil","Caxias". Ou seja, formas diferentes para o mesmo valor. Como esse atributo é importante, não foi



```
<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet type="text/xsl" href="friends.xsl"?>
<Friends>
 <data>
  <ARRAY_NUMBER_0>
   <id><![CDATA[508797523]]></id>
   <name><![CDATA[Alexandre França]]></name>
   <first_name><![CDATA[Alexandre]]></first_name>
   <last_name><![CDATA[França]]></last_name>
   <link><![CDATA[http://www.facebook.com/profile.php?id=508797523]]></link>
   <location>
    <id><![CDATA[109493952401635]]></id>
    <name><![CDATA[Fernandópolis]]></name>
   </location>
   <birthday><![CDATA[07/17]]></birthday>
   <hometown>
    <id><![CDATA[108154842546372]]></id>
    <name><![CDATA[Jundiai]]></name>
   </hometown>
   <gender><![CDATA[male]]></gender>
   <work>
    <ARRAY_NUMBER_0>
     <employer>
      <id><![CDATA[139921489368029]]></id>
      <name><![CDATA[A.O.F. Projetos Arq. & Urab.]]></name>
     </employer>
     <position>
      <id><![CDATA[114714781884455]]></id>
      <name><![CDATA[Designer de Interiores]]></name>
     </position>
     <start_date><![CDATA[0000-00]]></start_date>
     <end_date><![CDATA[0000-00]]></end_date>
    </ARRAY_NUMBER_0>
   </work>
```

Figura 4.4: Arquivo XML dos dados do perfil da Procad.

retirado da seleção, apenas foi normalizado. E ao normalizar foram criados outros dois atributos que são **'cidade' e 'país'**.

Os atributos *'school','employer' e 'position'* também possuem diversos valores diferentes, podendo comprometer a qualidade dos modelos de conhecimento extraídos, os valores inconsistentes foram normalizados e os que estão em branco foram preenchidos com o valor "não informado"e traduzidos para português, ficando **'escola','empregador','profissão'**.

Outro atributo de grande importância é a idade, porque dessa forma podemos determinar a faixa etária predominando dos perfis conectados, portanto foi extraído a data de nascimento (*'birthday'*), que no geral alguns perfis não respondem o ano de nascimento, gerando informação incompleta. Para esse caso, foi calculado através do Excel, as idades dos perfis que estavam com esse campo completo e preenchido com '0' quando o campo estava em branco ou sem o ano de nascimento.

A figura 4.6, apresenta algumas das alterações executadas na base de dados. No primeiro conjunto de dados têm-se os dados originais, no segundo conjunto os dados foram modificados.

As entidades que estavam com mais de 5 atributos em branco foram excluídas da base de dados.

```
<xsl:stylesheet
 xmlns:xsl="http://www.w3.org/1999/XSL/Transform"
 version="1.0">

 <xsl:output method="text" indent="yes"/>

 <xsl:template match="Friends">
     <xsl:for-each select="data/*">
         <xsl:value-of select="birthday"/>
         <xsl:value-of select="string(';')"/>
         <xsl:value-of select="gender"/>
         <xsl:value-of select="string(';')"/>
         <xsl:value-of select="location/name"/>
         <xsl:value-of select="string(';')"/>
         <xsl:value-of select="education/*/school/name"/>
         <xsl:value-of select="string(';')"/>
         <xsl:value-of select="education/*/type"/>
         <xsl:value-of select="string(';')"/>
         <xsl:value-of select="work/*/employer/name"/>
         <xsl:value-of select="string(';')"/>
         <xsl:value-of select="work/*/position/name"/>
         <xsl:value-of select="string(';')"/>
         <xsl:value-of select="interests/*/*/name"/>
         <xsl:call-template name="QuebrarLinha"/>
     </xsl:for-each>
 </xsl:template>

<xsl:template name="QuebrarLinha">
     <xsl:text>&#13;&#10;</xsl:text>
 </xsl:template>

 </xsl:stylesheet>
```

Figura 4.5: Código XSL para separar os atributos.

4.2.2 Transformação dos Dados

Outra função dessa etapa é a transformação dos dados, onde é necessário alimentar a base de dados selecionada, transformando valores reais em categorias ou intervalos, para os dados ficarem numa forma que possam ser usados como entrada dos algoritmos de mineração de dados, como por exemplo, para a classificação e clusterização. Para a aplicação do algoritmo de associação a base utilizada não deverá conter atributo 'idade', por ele ser numérico.

Após ter calculado a idade de cada entidade na etapa de limpeza dos dados, para ter uma melhor visualização com a ferramenta transformou-se os valores reais em intervalos, conforme faixas mostradas na tabela 4.1.

IDADE	INTERVALO
0	sem valor
1	18 a 25
2	26 a 30
3	31 a 35
4	36 a 40
5	Acima de 40

Tabela 4.1: Definição dos intervalos das idades.

Este processo visa facilitar o agrupamento e classificação na mineração dos dados. Assim, cada faixa de dados representada por um valore entre 0 a 5 é apresentado

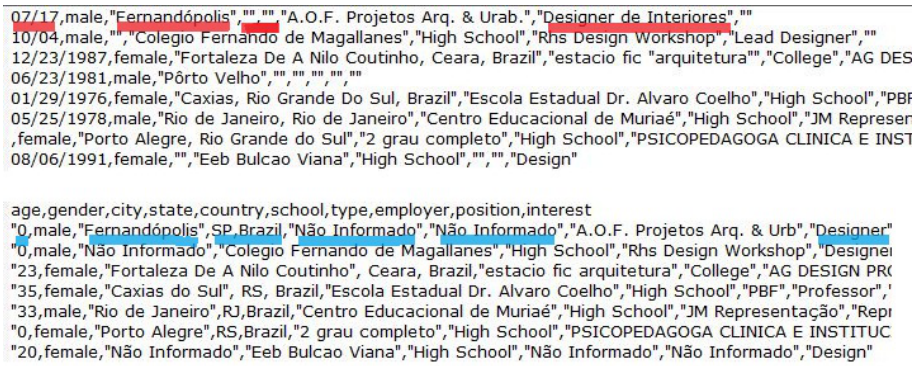


Figura 4.6: Correção do arquivo csv.

na figura 4.7, como parte do arquivo *.csv, com o atributo *'age'* destacado.

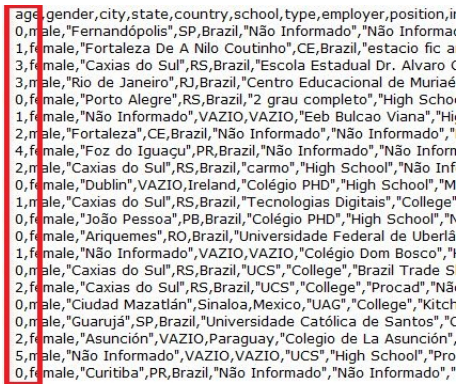


Figura 4.7: Arquivo com o atributo alterado.

Os dados similares, mas que estavam com valores diferentes foram agrupados para que ao minerar as análises não apresentem muitos valores diferentes, fazendo dessa forma que a mineração não tenha um resultado satisfatório. Portanto dados como **'profissão'** com valores de arquiteto, arquiteta foram modificados para "arquiteto(a)", valores como sócio, proprietário, empresário foram modificados para "empresário/proprietário/sócio", valores como designer de interiores e designer foram modicados para "designer".

O mesmo foi executados nos demais atributos onde verificou-se a necessidade por motivos de agrupamento ou grafia incorreta.

4.3 Mineração dos Dados

A mineração dos dados é a principal etapa do processo da DCBD, onde são definidas as técnicas e os algoritmos a serem utilizados. Nessa etapa é feita a busca por conhecimento desconhecido e relevante a partir dos dados fornecidos (base de dados). A técnica aplicada depende basicamente do tipo de tarefa a ser realizada. Cada algoritmo tem uma tarefa em que melhor se adapta e obtém melhor resultado,

sendo necessário em geral a utilização de vários algoritmos até encontrar aquele que apresenta os melhores resultados. (GOLDSCHMIDT; PASSOS, 2007)

As regras a seguir foram utilizadas nesse estudo de caso, a descrição das regras aplicadas estão descritas no capítulo 2, seção 2.2.

Após o processo de tratamento dos dados, formou-se uma base de dados em formato CSV, que será carregado no WEKA para a análise dos dados. A versão do Weka que será utilizada será a 3.6.5, conforme mostra na figura 4.8.

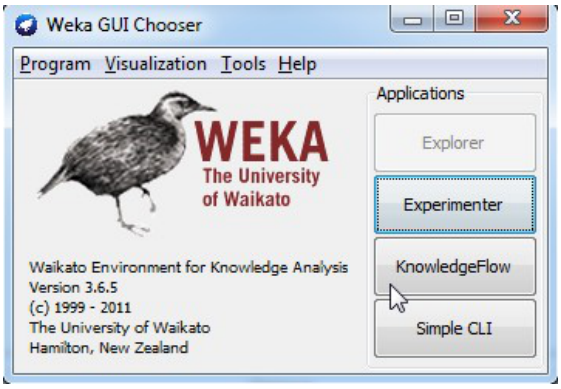


Figura 4.8: Versão do WEKA utilizada.

No modo Explorer serão tratadas as etapas de pré-processamento (seleção, limpeza e transformação dos dados), mineração de dados (classificação e associação) e o pós-processamento (apresentação dos resultados).

A figura 4.9 mostra os passos do módulo Explorer, e a guia de pré-processamento.

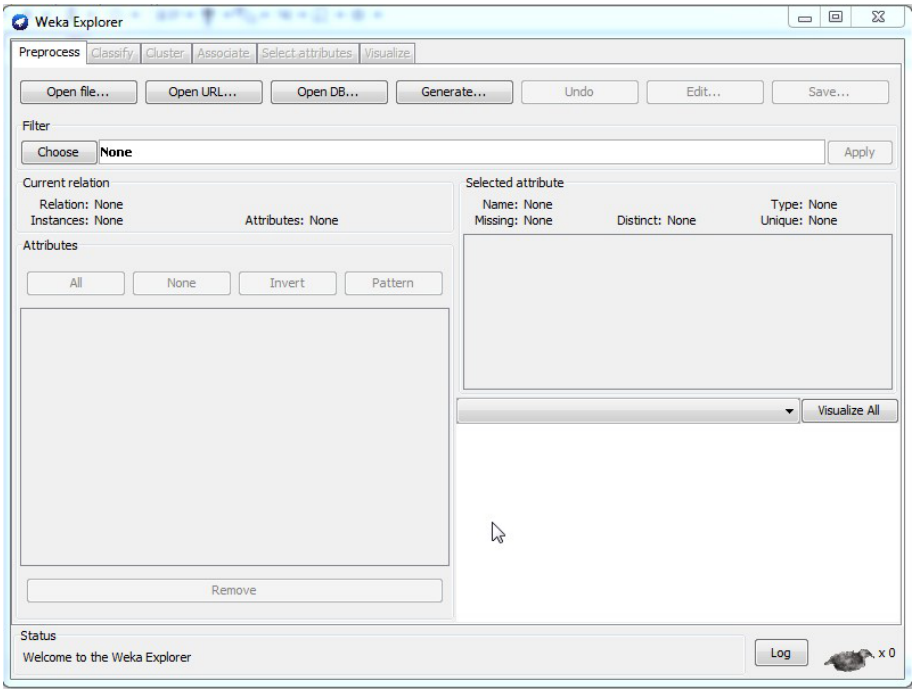


Figura 4.9: Tela do Módulo Explorer

Nesta etapa é selecionada a base de dados **Facebook.csv**, e na opção *"open file"* carregada a base de dados na ferramenta, a seguir, selecionam-se os atributos necessários para gerar as regras que irão gerar o perfil.

4.4 Análise dos Resultados: Relação Sexo x Idade x Estado

4.4.1 Seleção de dados

Para essa análise foram separados os dados apenas com os atributos idade, sexo e estado em um arquivo chamado **FacebookAgeGenderState.csv**, onde os demais atributos foram retirados dessa relação. Na figura 4.10 é selecionada a base de dados que demonstra a remoção dos atributos e tela do pré-processamento dos dados selecionados.

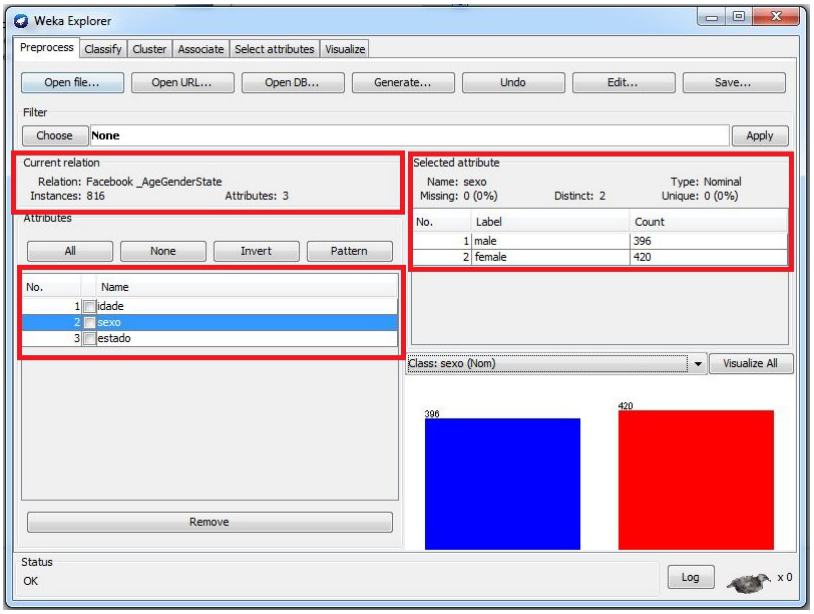


Figura 4.10: Filtro de atributos e seleção da base de dados.

As características dos dados dessa análise são as seguintes:

- Número de instâncias: 816
- Atributos selecionados: 3

Na tabela 4.2, estão listados os atributos 'sexo' e 'idade' com a quantidade que cada campo de atributo possiu.

Na seleção dos atributos, pode-se visualizar no estilo de barras os atributos filtrados, conforme figura 4.11.

4.4.2 Classificação Utilizando o Algoritmo J48

O algoritmo de classificação pode apresentar o resultado sob a forma de árvore de decisão, como é o caso do algoritmo J48 utilizado na apresentação das análises,

SEXO	Female		420
	Male		396
IDADE	0	sem valor	303
	1	18 a 25	207
	2	26 a 30	130
	3	31 a 35	83
	4	36 a 40	34
	5	Acima de 40	59

Tabela 4.2: Tabela com as características da Base de Dados

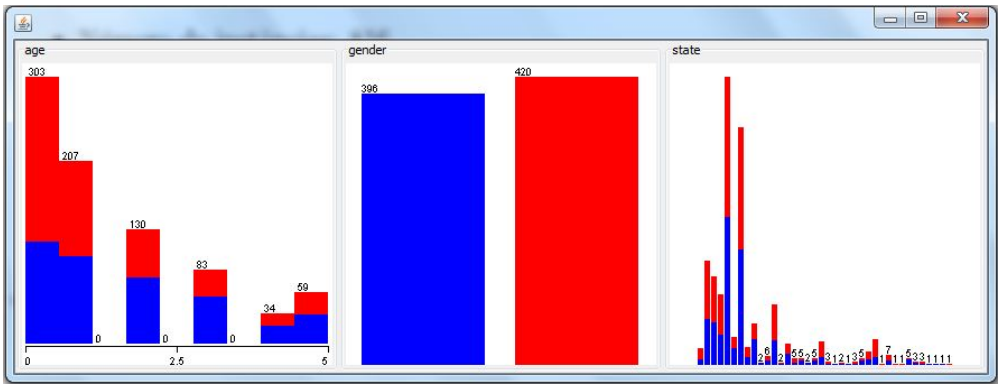


Figura 4.11: Visualização dos atributos da análise.

baseado num conjunto de dados de treinamento, utiliza esse modelo para classificar com exatidão do classificador num conjunto de teste. A exatidão do classificador é analisada através da estatística de Kappa, que são apresentados no resultado na mineração e seguem uma faixa para classificá-los em fraco, regular, moderado, bom e excelente, conforme tabela 4.3.

PROPRABILIDADE DE KAPPA	EXATIDÃO DO CLASSIFICADOR
menor que 0.20	Fraco
0.21 - 0.40	Regular
0.41 - 0.60	Moderado
0.61 - 0.80	Bom
maior que 0.81	Excelente

Tabela 4.3: Probabilidade de Kappa

Na guia de classificação o algoritmo J48 implementado na ferramenta WEKA exibe uma árvore de decisão com as classificações através do atributo 'sexo', mostrando todos os atributos filtrados, classificados por 'sexo'. Este algoritmo foi o que apresentou um melhor percentual de classificação (56.25%) sendo que i índice Kappa está como fraco. As figuras 4.12 e 4.13 apresentam os valores gerados e a árvore de decisão gerada.

Executando a regra de classificação com o algoritmo J48, verifica-se que a maioria dos perfis cadastrados com idade menor ou igual a 1 chegam a 520, sendo que destes

```
J48 pruned tree
------------------

idade <= 1: female (510.0/215.0)
idade > 1: male (306.0/125.0)

Number of Leaves  :     2

Size of the tree :      3


Time taken to build model: 0.05seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         459               56.25   %
Incorrectly Classified Instances       357               43.75   %
Kappa statistic                          0.1234
Mean absolute error                      0.4893
Root mean squared error                  0.4962
Relative absolute error                 97.9364 %
Root relative squared error             99.2849 %
Total Number of Instances              816
```

Figura 4.12: Valores gerados - J48.

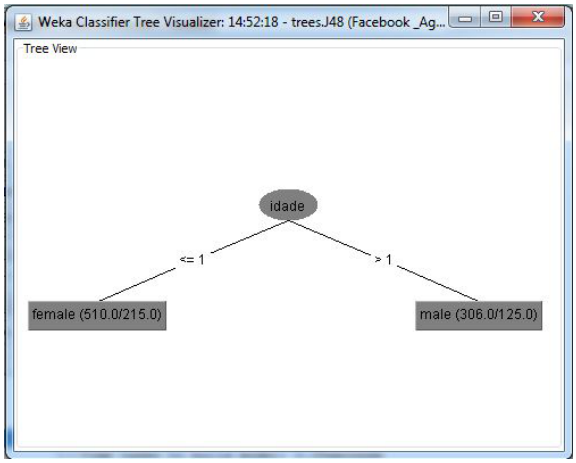


Figura 4.13: Gráfico da Árvore de Decisão - J48

215, são do sexo feminino. E a maioria dos perfis cadastros com idade superior a 1 chegam a 306, sendo que destes, 125 são do sexo masculino.

Quando levamos essa análise para a área de *marketing*, pode-se concluir que conforme a maioria dos perfis com o índice de idade 0 são do sexo feminino, o que demonstra que as mulheres não costumam informar a idade nas redes sociais. Outro ponto relevante é que conforme aumenta a idade, aumentam os perfis do sexo masculino. Neste caso, pode-se concluir que é devido a haver uma comunidade masculina maior ou que os homens são mais propícios a informar sua idade.

4.4.3 Clusterização Utilizando Algoritmo EM

Na guia *cluster* no botão *choose* escolhe-se qual o algoritmo de clusterização a ser utilizado, e nesse caso será utilizado o algoritmo EM, para isso é necessário configurar alguns parâmetros que são:

- *Debug:* normalmente usa-se *false*, pois o *debug* não se faz necessário;
- *MaxIterations:* número máximos de interações que o algoritmo irá realizar;
- *MinStdDev:* desvio padrão aceitável;
- *NumClusters:* configuração do número de *cluster* que o algoritmo deve gerar, no caso do valor -1, o algoritmo irá determinar o melhor valor;
- *Seed:* número de centroides iniciais para o EM. Depois de configurado, deve-se clicar no botão *start* e aguardar o resultado da execução.

Utilizando a mesma base de dados do algoritmo de classificação, o algoritmo de clusterização apresenta os seguintes resultados, conforme tabela 4.4.

CLUSTER	QUANTIDADE DE INSTÂNCIAS	PROBABILIDADE
0	126	15%
1	185	23%
2	505	62%

Tabela 4.4: Resultado dos *clusters* gerados.

Nota-se que o *cluster* mais populoso e com maior probabilidade de instância é o *cluster* 2. Para um melhor entendimento, separou o resultado dos atributos em tabelas e figuras para correlacionar os valores dos atributos nos *clusters*.

A tabela 4.5 apresenta a média de idade de cada *cluster* gerado, bem como o desvio padrão. Pode-se observar que o *cluster* 2 tem a média mais baixa com a probabilidade mais alta, devido a grande quantidade de instância com valor 0 ser mais significativa.

CLUSTER	MÉDIA IDADE	DESVIO PADRÃO
0	2.6145	1.43
1	2.5289	1.39
2	0.338	0.4778

Tabela 4.5: Resultado do atributo Idade e *cluster* gerados.

No gráfico demonstrado pela figura 4.14, pode-se observar as classificações citadas na tabela acima. O *cluster* 0 está representado pela cor azul, o *cluster* 1 está representado pela cor vermelho e o *cluster* 2 está representado pela cor verde.

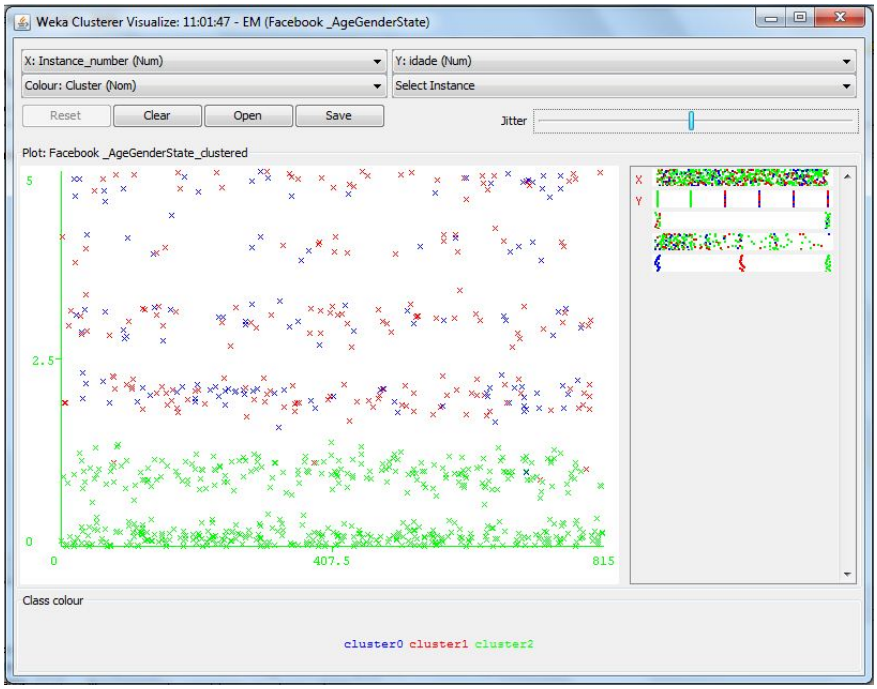


Figura 4.14: Gráfico com relação idade x *cluster*.

Na tabela 4.6, pode-se verificar os valores dos *clusters* com relação ao atributo sexo demonstram que o *cluster* 0 possui um maior número de pessoas do sexo feminino (*female*), enquanto que o *cluster* 1 possui um maior número pessoas do sexo masculino (*male*), e o *cluster* 2 que é o mais populoso esse atributo tem maior relevância nos atributos 'idade' e 'estado', portanto ele está bem balanceado em se tratando do atributo 'sexo'.

	Female	Male
CLUSTER 0	146.2765	10.5994
CLUSTER 1	226.4671	12.2506
CLUSTER 2	264.4728	161.9336

Tabela 4.6: Resultado do atributo Sexo e *cluster* gerados.

No gráfico demonstrado pela figura 4.15, pode-se observar as classificações citadas na tabela.

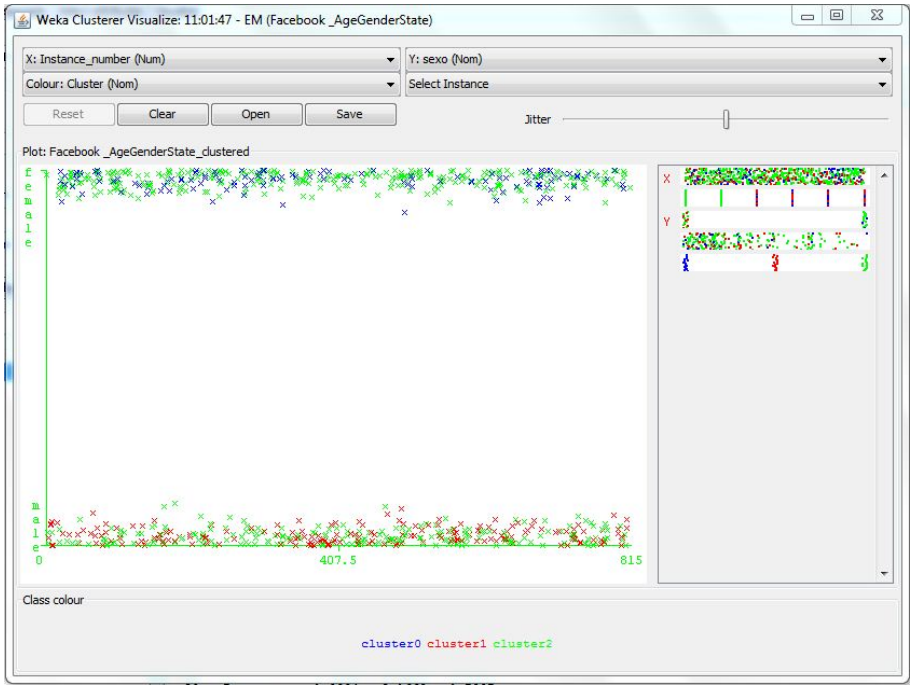


Figura 4.15: Gráfico com relação sexo x *cluster*.

Na figura 4.16, pode-se observar a visão do 'sexo' por 'idade'. Pode-se observar a questão da concentração de valores com idade 0 e também como foram divididos os perfis do sexo feminino e masculino.

Para uma melhor visão quanto a distribuição dos 'estados', a figura 4.17 demonstra a distribuição dos *clusters* por estado, resultando uma maior concentração nos estados do Sul (RS, SC e PR), Sudeste (SP, RJ e MG) e Nordeste (BA e PB).

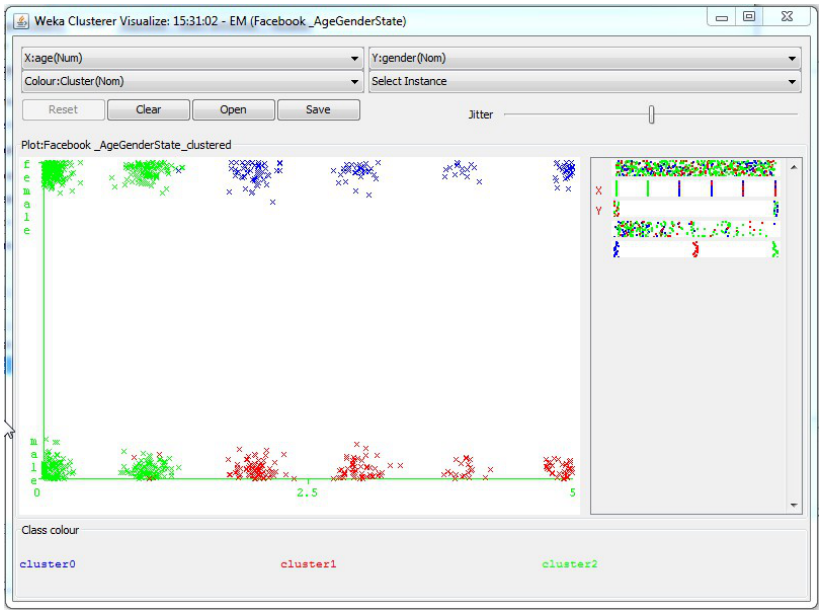


Figura 4.16: Gráfico com relação sexo x idade separados nas cores dos *clusters*.

4.5 Análise dos Resultados: Relação Sexo x Profissão x Interesses

4.5.1 Associação Utilizando Algoritmo Apriori

A descoberta da associação abrange a busca por conjuntos de itens que frequentemente ocorrem de forma simultânea em transações de banco de dados, assim gerando uma grande quantidade de regras. (GOLDSCHMIDT; PASSOS, 2007)

A regra de associação como é denominada, identifica as relações existentes entre os eventos ocorridos em determinada ocasião, sendo considerado um método descritivo, ou seja, usado para identificar padrões em dados históricos.

Para a utilização da tarefa de associação foi necessário um novo arquivo *.cvs com os atributos nominais, como sexo, empregador, profissão e interesses, o nome desse arquivo é **FacebookAssociate.csv**. Na guia *Associate* da ferramenta WEKA o algoritmo Apriori foi selecionado e configurado os parâmetros conforme figura 4.18.

Devido ao grande número de atributos com o valor "Não Informado", ao aplicar o algoritmo Apriori os resultados não são conclusivos, gerando regras apenas com o resultado do atributo "Não Informado". A figura 4.19 demonstra o resultado da aplicação do algoritmo de associação.

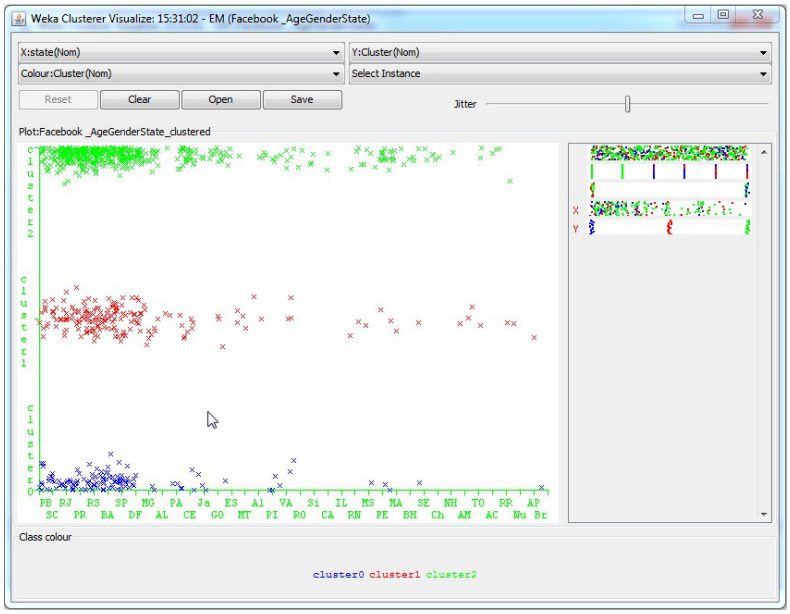


Figura 4.17: Gráfico com relação estado x *clusters*.

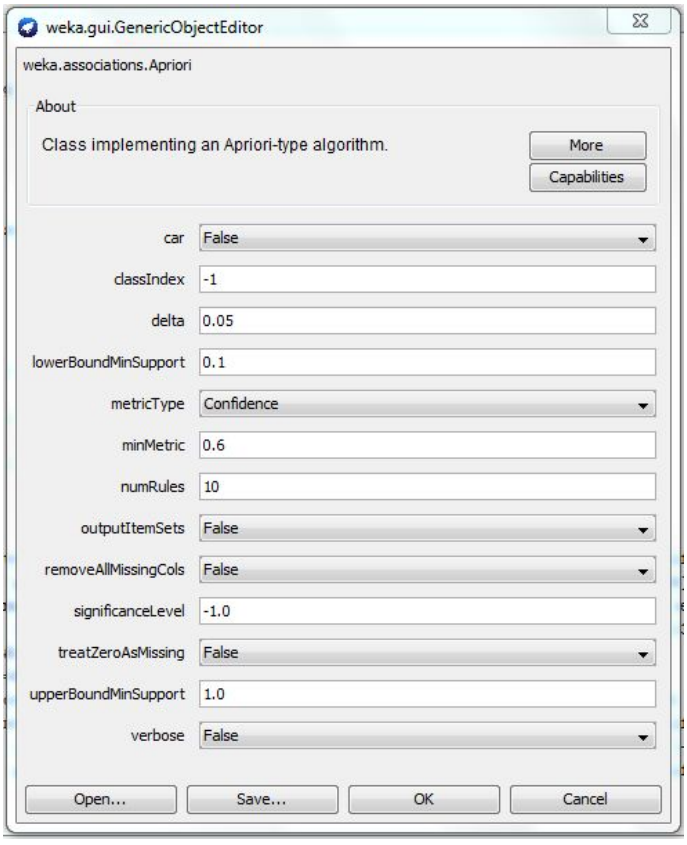


Figura 4.18: Configuração do algoritmo do Apriori.

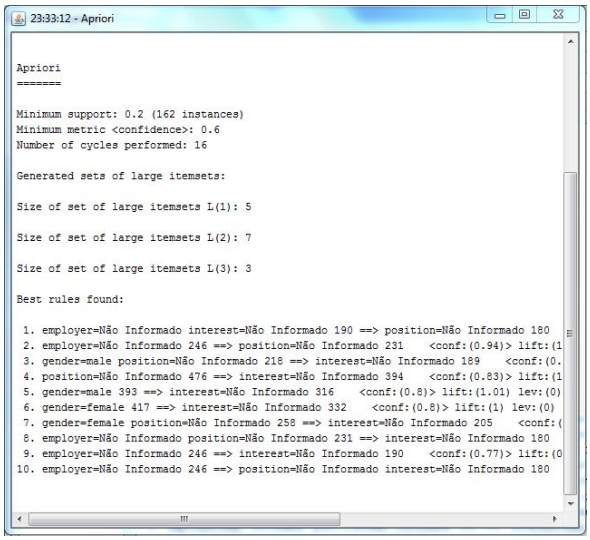


Figura 4.19: Resultado do algoritmo do Apriori.

5 CONSIDERAÇÕES FINAIS

Com o término do Trabalho de Conclusão onde foi realizada primeiramente foi conceituado a Descoberta do Conhecimento em Banco de Dados e as Redes Sociais e após a aplicação desses conceitos no estudo de caso dos Dados do Facebook da empresa Procad Softwares Ltda foi possível extrair as seguintes conclusões.

Uma delas e referente a extração dos dados e como são apresentados esses dados para a manipulação desses arquivos. O problema enfrentado neste tipo de situação é que os dados por ser originados de uma rede social, muitos atributos estão incompletos ou preenchidos de maneiras diferentes, o que ocasionou um trabalho de preparação manual e trabalhoso. Além desse problema, devido ao grande número de valores "não informados"e valores diferentes comprometerem a aplicação dos algoritmos de mineração de dados.

Também vale ressaltar que a cada dia entram novos perfis, o que ocasionaria um trabalho de limpeza de dados e transformação a cada vez que tenha necessidade de extrair os dados.

Como sugestão para trabalhos futuros, seria implementar um aplicativo em forma de pesquisa com campos pré-definidos e que o usuário responda, com esses dados preenchidos é possível criar uma base de dados consistente e de fácil extração para aplicar os métodos de mineração. Como o perfil da Procad tem um grande número de frequentadores seria um canal de contato e a possibilidade de troca de informações.

E como conclusão final tem o fato que embora se tenha dificuldades ao realizar um processo de Descoberta de Conhecimento, este tipo de trabalho pode ser incorporado para o suporte à tomada de decisões. Apesar de que o resultado esperado pela área de *marketing* será parcial, gerando algum resultado apenas nos atributos como idade, cidade e sexo dos perfis cadastrados.

Em se tratando dos atributos 'interesses', 'profissões' e 'empregador' os resultados gerados não foram satisfatórios através da mineração de dados, a busca por padrões analisados na base de dados é comprometida pela grande quantidade de valores zerados. Para a busca desses dados uma consulta e geração de relatórios através de uma banco de dados, poderá definir de maneira mais clara a quantidade de pessoas

em determinada profissão e quais os seus assuntos de interesse.

REFERÊNCIAS

AMO, S. de. Técnicas de mineração de dados. , www.deamo.prof.ufu.br/arquivos/JAI-cap5.pdf, 2011.

BEZERRA, E.; GOLDSCHMIDT, R. A tarefa de classificação em text mining. **Revista de Sistemas de Informação da FSMA**, v.5, p.42–62, 2006.

BOYED, D. M.; ELLISON, N. B. **Social network sites**: definition, history, and scholarship. Disponível em: http://jcmc.indiana.edu/vol13/issue1/boyd.ellison.html. Acesso em: junho/11.

BRESOLIM, A. Desenvolvimento de um sistema de recomendação fazendo uso de técnicas de mineração de dados. , ¡http://unochapeco.edu.br/static/files/trabalhos-anais/Pesquisa/¿, 2011.

BRUSSO, M. J. O uso de mineração de dados na descoberta do comportamento do usuário da web. , 2010. Disponível em: ¡http://www.revistainvestmais.com.br/arquivos/testes/apriori.pdf¿ Acesso em Maio/2011.

DRAMBROS, J.; REIS, C. A marca das redes sociais virtuais: uma proposta de gestão colaborativa. **Intercom - Sociedade Brasileira de Estudos Interdisciplinares da Comunicação**, 2008. Disponível em: http://www.intercom.org.br/papers/nacionais/2008/resumos/R3-0519-1.pdf.

FAYYAD, U. M.; PIATETSKY-SHAPIRO, G.; SMYTH, P.; UTHURUSAMY, R. (Ed.). **Advances in knowledge discovery and data mining**. [S.l.]: AAAI/MIT Press, 1996.

FREITAS, C.; NEDEL, L. P. Extração do conhecimento e análise visual de redes sociais. **Anais do XXVIII Congresso da SBC**, p.106–120, 2008. Disponível em: ¡http://http://www.prodepa.gov.br/sbc2008/anais/pdf/arq0091.pdf¿ Acesso em Julho/2011.

GOLDSCHMIDT, R.; PASSOS, E. **Data mining - um guia prático**. [S.l.]: CAMPUS, 2007.

JESUS, A. P. de; MOSER, E. M.; OGLIARI, P. J. Data mining aplicado na identificação do perfil dos usuários de bibliotecas para a personalização de sistema web de recuperação e disseminação de informações. , http://inf.unisul.br/ ines/workcomp/cd/pdfs/2371.pdf, 2011.

JUNIOR, A. J. S.; PEREZ, A. L. F. Análise de perfil do usuário de serviços de telefonia utilizando técnicas de mineração de dados. **Revista Eletrônica de Sistemas de Informação**, 2006.

KILDER, S. C. **Estudo sobre a gestão do conhecimento aplicada à indústria - uma abordagem apoiada em técnicas de ia**. 2007. Trabalho de Conclusão de Curso de Ciência da Computação — Universidade de Caxias do Sul, Caxias do Sul, RS.

KIRKPATRICK, D. (Ed.). **O efeito facebook**. [S.l.]: Intrínsica, 2011.

MACEDO, D. C. de; MATOS, S. N. Extração do conhecimento através da mineração de dados. **Revista de Engenharia e Tecnologia**, v.2, n.3, 2010.

MENESES, M. P. R. **Redes sociais pessoais**: conceitos, práticas e metodologia. 2007. Tese de Doutorado em Psicologia — Pontíficia Universidade Católica do Rio Grande do Sul, Porto Alegre, RS.

TREADAWAY, C.; SMITH, M. **Facebook marketing**: an hour a day. [S.l.]: John Wiley & Sons, 2010.

WAIKATO, U. of. Disponível em: http://www.cs.waikato.ac.nz/ml/weka/ - Acesso em: Agosto/2011.