Análise do Comportamento de Rotatividade no LinkedIn utilizando aprendizagem de máquina

Ana Carolina C. de Jesus[1], **Wladmir C. Brandão**[1]

[1]Instituto de Informática e Ciências Exatas e Informática
Caixa Postal 15.064 – 91.501-970 – Belo Horizonte – MG – Brasil

[2]Departamento de Computação – Ciência da Computação

{ana,wladmir}@sga.pucminas.br

Abstract. *RESUMO A SER PRODUZIDO*

Resumo. *RESUMO A SER PRODUZIDO*

1. Introdução

Atualmente observamos um cenário econômico instável e um aumento e intensificação da rotatividade profissional no mercado de trabalho brasileiro. De acordo [WIKIPÉDIA 2015], a rotatividade pessoal, também conhecida como *turnover*, refere-se a relação entre admissões e demissões de trabalhadores, ou ainda a taxa de substituição de trabalhadores antigos por novos em uma organização. Para [HAMMES et al. 2016] o *turnover* é resultado da saída de alguns colaboradores e da entrada de outros para substitui-los nos postos de trabalho. Essa movimentação pode ser motivada pela iniciativa da empresa ou do próprio funcionário, que pode decidir seu desligamento da empresa por motivos pessoais ou por causa da relação de trabalho com o empregador. Do ponto de vista da organização, o desligamento pode ocorrer por fatores como: substituição por um profissional mais adequado ao cargo, seleção inadequada, redução do quadro de funcionários ou mudanças no desenho de cargos. A organização espera do profissional que ele obtenha um desempenho satisfatório em relação as atividades exercidas. A organização tem que proporcionar condições de trabalho dignas e motivação aos seus funcionários. Por sua vez o profissional, mediante essas boas condições e motivação, precisa buscar mecanismos para atender a essas necessidades da organização, para que seu posto de trabalho não seja ameaçado.

Do ponto de vista do profissional, esse movimento de entrada e saída do mercado de trabalho pode ocorrer por diversos fatores. Dentre eles, podemos destacar: ambiente de trabalho hostil; no sentido do profissional se sentir acuado, reprimido ou humilhado no trabalho; baixa remuneração e falta de benefícios, liderança ruim, falta de um plano de carreira, rotina, dentre outros motivos. Além disso, a rotatividade também pode ser influenciada pelo ramo de atividade que o profissional está atuando e questões referentes a insalubridade da atividade exercida. Trabalhos considerados perigosos, onde os profissionais ficam em contato com substâncias químicas, serviços da área da saúde, industria metalúrgica, são ramos em que o profissional não pode ficar por muito tempo exposto as atividades para que não desenvolva problemas de saúde. Segundo [LÚCIO 2015], no Brasil, as empresas têm liberdade para contratar e demitir a qualquer momento o trabalhador, sem precisar apresentar nenhuma justificativa, basta apenas arcar com os custos da rescisão do contrato de trabalho.

A rotatividade profissional pode causar repercussões nocivas a organização como custos tangíveis e intangíveis [SIQUEIRA 2014]. Os custos tangíveis dizem respeito a recrutamento, seleção, benefícios, treinamento, integração e desligamento, dentre outros. Enquanto que os custos intangíveis dizem respeito a perda de *know-how* e conhecimento, quebra do fluxo de trabalho, quebra de vínculo com fornecedores e clientes, etc. Ainda existem custos inesperados, ligados a contratação de advogados e processos na justiça do trabalho, que podem ser movidos por ex-funcionários. Segundo [RHPORTAL 2015], por exemplo, em empresas do ramo metalúrgico, o custo de rotatividade de pessoal pode chegar até o equivalente a 8 salários nominais, por empregado, dependendo do cargo. Isso quer dizer que pelo mesmo valor, o mesmo funcionário poderia ser mantido trabalhando durante 8 meses. Caso esses custos não sejam estimados, podem contribuir para agravar a saúde financeira de uma corporação, vindo a trazer problemas para os gestores. Portanto, a rotatividade profissional é uma das variáveis que influenciam a saúde financeira de uma organização e por esse motivo, sua estimativa e controle são importantes para manter o bom funcionamento de uma corporação.

As redes sociais são ambientes perfeitos para o estudo de temas da ciência da computação, como sistemas distribuídos, padrões de tráfego na Internet, mineração de dados entre outros [BEVENUTO et al. 2011]. Como essas redes permitem que usuários criem conteúdos, elas se tornam um ambiente perfeito para realizar pesquisas relacionadas à organização e tratamento de grandes quantidades de dados, onde aplicamos técnicas de mineração de dados para extrair conhecimento. O LinkedIn é uma das maiores redes profissionais do mundo, e conta atualmente com 300 milhões de usuários em mais de 200 países e territórios no mundo todo [LINKEDIN 2015]. Além disso, Linkedin é uma empresa de capital aberto e sua receita vem de assinaturas de usuários, vendas de publicidade e soluções de talentos (recomendação de perfis profissionais) [LINKEDIN 2015]. Seu principal objetivo é conectar profissionais e torná-los mais produtivos e bem sucedidos. Ao se cadastrar no Linkedin, rapidamente o usuário pode adicionar pessoas em seu perfil, além de preencher informações a respeito de suas habilidades, experiências, desejos e metas para a vida profissional. Portanto, o perfil dos usuários do Linkedin serve como um currículo on-line ou um portifólio contendo informações que podem ser importantes para contratantes que buscam profissionais no mercado de trabalho. Isso faz com que o Linkedin seja uma fonte com potencial muito elevado para o estudo da rotatividade, usando técnicas de mineração de dados.

A mineração de dados é um processo que permite descobrir conhecimento em bancos de dados ou *Knowledge Discovery in Databases*, também conhecida como KDD [BUENO and VIANA 2012]. É responsável por analisar grandes volumes de dados utilizando diferentes perspectivas, com o objetivos de descobrir informações úteis que normalmente não estão visíveis ou não são facilmente encontradas [SILVA and BRANDÃO 2015] [VASCONCELOS and CARVALHO 2004]. Segundo o autor de [FAYYAD et al. 1996], mineração de dados é um passo no processo de KDD que consiste em analise de dados e algoritmos de descobrimento que produzem uma enumeração de padrões particular sobre os dados. Para o autor de [JAIWEI and KAMBER 2006], a descoberta de conhecimento em bancos de dados é dividida em sete grandes etapas: limpeza de dados, integração, seleção, transformação, mineração, avaliação dos modelos encontrados e apresentação do conhecimento adquirido.

As organizações dão uma importância muito grande em se precaver quanto á possíveis custos tangíveis e intangíveis, ligados a rotatividade profissional, por isso, é importante analisar o perfil do profissional que está sendo contratado. O estudo da rotatividade no mercado de trabalho é um tema que está sendo muito discutido atualmente [HAMMES et al. 2016], justamente por causa desses custos. Ainda segundo [HAMMES et al. 2016], a preocupação dos gestores das organizações em controlar a intenção de rotatividade, tem feito com que invistam mais na identificação dessas intenções de rotatividade para atuarem maneira preventiva e rigorosa sobre essas probabilidades. A caracterização da rotatividade profissional de acordo com as variáveis que compõem seu comportamento é uma das maneiras de se tentar obter uma tendência. A motivação desse trabalho está no fato de que a rotatividade profissional é um dos parâmetros que ditam a boa saúde organizacional, que é de interesse direto aos níveis de gestão de uma corporação. Por esse motivo, os objetivos principais desse trabalho são: realizar a análise estatística da base de dados de perfis coletados do LinkedIn, utilizar técnicas de aprendizagem de máquina a fim de encontrar padrões relevante ao que diz respeito a rotatividade profissional, comparar resultados das aplicações de diferentes técnicas de aprendizado de máquina e por fim apresentar os padrões relevantes encontrados bem como a(s) técnica(s) que obtiverem os melhores resultados.

As seções a seguir apresentam referencial teórico, metodologia, discussão dos resultados, conclusões e trabalhos futuros e referências.

2. Referencial Teórico

Esta seção apresenta o referencial teórico de embasamento do trabalho.

2.1. Rotatividade Profissional e LinkedIn

Segundo [Palhoça 2016], o bem mais precioso de uma empresa são seus funcionários e colaboradores, uma vez que, são eles os responsáveis por garantir que as atividades da empresa serão realizadas. Logo, a mão de obra trabalhadora de uma empresa representa um papel importante garantindo de fato que a empresa produza resultados. De acordo com o autor de [FONTAINHAS Mariana 2013], muito mais do que apenas instalações modernas, alta tecnologia, estruturas organizacionais bem definidas e planejamento estratégico; as empresas precisam que os funcionários contribuam adequadamente dentro de cada um dos processos da empresa que lhes são delegados para garantir o bom funcionamento e o crescimento da empresa.

Dessa maneira, para garantir esse bom funcionamento e crescimento da empresa é necessário que existam mecanismos ou órgãos responsáveis pela gestão de pessoas dentro da organização. Esses órgãos são os chamados Recursos Humanos (RHs). A função dos RHs está fortemente vinculada a gestão de pessoas e procura trabalhar os mais variados aspectos ligados a essa tarefa como potencial, interação, compromisso, inovação, força de trabalho e criatividade. Também é papel de um RH selecionar, recrutar e muitas vezes realizar o treinamento dos profissionais que chegam a empresa. Um dos desafios que esses órgãos de RHs enfrentam é tentar reter os colaboradores que agregam valores importantes a empresa, afim de minimizar ao máximo a rotatividade de pessoal que também é conhecida como *turnover*. Segundo o autor de [PINHEIRO and SOUZA 2013] turnover ou rotatividade de pessoal refere-se à relação entre admissões e os desligamentos profissionais ocorridos de forma voluntária ou involuntária, em determinado período. Além disso,

ainda em [PINHEIRO and SOUZA 2013] os autores ressaltam que a rotatividade de pessoal influencia diretamente nos resultados das empresas e que o turnover representa um aspecto importante da dinâmica organizacional.

A rotatividade de pessoal pode refletir na produção, no clima organizacional e no relacionamento interpessoal tanto negativamente quanto positivamente, isso depende se o RH da empresa faz observação, análise e controle da rotatividade no âmbito empresarial. Como impactos positivos, uma rotatividade alta pode representar a renovação de postos de trabalho que não rendem desempenho esperado, chance de encontrar novos talentos e idéias. Em contrapartida, uma rotatividade baixa pode indicar que a organização anda estagnada e não renova ou modifica sua mão de obra, o que pode impedir que novos talentos entrem na organização. É importante para a saúde organizacional, que os RHs monitorem a rotatividade, afim de garantir um equilíbrio de rotatividade de seus empregados. De acordo com [Nery 2015] alta rotatividade nos postos de trabalho faz com que os empregados e empregadores fiquem desestimulados a investir nas relações de trabalho, principalmente a qualificação, e esse é um dos fatores que podemos utilizar para explicar a baixa produtividade da economia do pais.

Em [Portal-Brasil 2014] o autor ressalta a importância de estudar esse tema uma vez que o pais está vivendo um período em que as altas taxas de rotatividade verificadas no mercado estão cada vez mais evidentes e em crescimento. Os motivos da alta rotatividade podem depender de diversos fatores, área, experiência, flexibilidade contratual, idade, dentre outros motivos. De acordo com [DIEESE 2014] cerca de 30% de todos os contratos desligados possuem seus vínculos interrompidos com menos de três meses de vigência. Esses vínculos empregatícios são contratos de experiência, onde os empregados são isentos de pagar a multa rescisória do Fundo de Garantia do Tempo de Serviço (FGTS) e o aviso prévio. Ainda segundo [DIEESE 2014] cerca de 45% de todos os contratos desligados por ano são interrompidos com seis meses ou menos de vigência. Isso porque o critério para recebimento do seguro desemprego era de no mínimo seis meses de permanência no emprego. Em [G1 2014] é possível verificar que o governo aumentou o tempo mínimo para um ano e meio de permanência no emprego para garantir o direito ao seguro desemprego. Essa é uma medida clara para tentar amenizar essa rotatividade no mercado apenas para garantir a obtenção do direito ao seguro desemprego.

O *Linkedin* é uma das maiores redes profissionais do mundo, e conta atualmente com 300 milhões de usuários em mais de 200 países e territórios no mundo todo [LINKEDIN 2015]. Além disso, *Linkedin* é uma empresa de capital aberto e sua receita vem de assinaturas de usuários, vendas de publicidade e soluções de talentos (recomendação de perfis profissionais) [LINKEDIN 2015]. Seu principal objetivo é conectar profissionais e torná-los mais produtivos e bem sucedidos. Ao se cadastrar no *Linkedin*, rapidamente o usuário pode adicionar pessoas em seu perfil, além de preencher informações a respeito de suas habilidades, experiências, desejos e metas para a vida profissional. Portanto, o perfil dos usuários do Linkedin serve como um currículo on-line ou um portifólio contendo informações que podem ser importantes para contratantes que buscam profissionais no mercado de trabalho. Isso faz com que o *Linkedin* seja uma fonte com potencial muito elevado para o estudo da rotatividade, usando técnicas de mineração de dados. O *Linkedin* se torna então um ambiente propício para o estudo da rotatividade dos profissionais no mercado de trabalho, uma vez que possui informações como ex-

periências, competências, formação, área, cargo, dentre outras coisas a respeito dos mais diversos profissionais presentes no mercado de trabalho. O perfil de um profissional no Linkedin é um currículo online que pode ser utilizado por empresas para analisar o seu histórico profissional e tomar a decisão em favor da contratação ou não do profissional.

2.2. Coleta de Dados em Redes Sociais

As redes sociais são ambientes que tem atraído milhões e bilhões de usuários por todo o mundo. Tal atração, se deve ao fato de que é um meio permite seus usuários criar e compartilhar informações, opiniões e sentimentos. Dessa maneira, as redes sociais constituem um lugar propício para se realizar estudos utilizado esses dados. Esse tipo de estudo, utilizando dados de redes sociais, tem se tornado cada vez mais comuns, justamente pelo potencial de aplicação que o conhecimento adquirido nessas analises podem representar. Por exemplo, a análise de sentimentos dos usuários de determinado produto em redes sociais, baseado em comentários que eles produzem a respeito do produto em seus perfis, pode ser aplicado para melhorar o produto em questão e também ajudar no desenvolvimento de estratégias de marketing melhores, a partir dos problemas e observações relevantes identificadas no estudo.

Existem variadas formas de obtenção dos dados de redes sociais de maneira metódica e automática. Basicamente, existem dois tipos de coleta, a fechada e a aberta. A coleta fechada utiliza as APIs ou *Aplicattion Programming Interface*, e está sujeita a limitações e restrições impostas pelo desenvolvedor da API. Enquanto que a coleta aberta não utiliza as APIs fornecidas pelas redes sociais, e sim outros tipos de mecanismos próprios na obtenção de dados que são compartilhados pelos usuários dessas redes em perfis públicos. A maioria das redes sociais oferecem uma API que permite quem utiliza, coletar dados e desenvolver aplicações utilizando essas interfaces [XAVIER and de CARVALHO 2011]. Mas devido as restrições impostas pelos desenvolvedores das APIs, a coleta aberta acaba sendo necessária, como primeira opção ou mesmo complementação das informações. Independente do estilo da coleta para se obter de metódica e automaticamente os dados é necessário desenvolver um *crawler* ou coletor ou rastreador. O *crawler* é um programa de computador capaz de varrer a *web* recuperando de maneira eficiente todo e qualquer documento ou informação que possa ser relevante.

Quando se trata de coleta na web, existem uma série de problemas e desafios. Alguns principais que podem ser citados são: rastreamento de *urls* redundantes, limitações quanto ao servidor em que os arquivos são coletados, sincronismo da coleta entre *crawlers*; para o caso de arquiteturas distribuídas, como saber se o coletor está recolhendo mesmo *urls* que sejam relevantes, dentre outras coisas. Segundo [Park and Lee 2014], o estudo e implementação de coletores abrangem várias metodologias. Algumas dessas metodologias podem impactar diretamente no desempenho do *crawler*, executar múltiplos coletores em um mesmo servidor pode não resolver o problema de escalabilidade em relação ao rápido crescimento de conteúdo que se deseja coletar, pois uma rede social é um ambiente que possui milhões ou bilhões de usuários. Por esse motivo, pode surgir a necessidade de implementar uma arquitetura além de paralela, também distribuída, que funcione de maneira efetiva na coleta. Além disso, é importante analisar questões ligadas as restrições dos servidores de onde se pretende realizar a coleta e também pensar em uma maneira de balancear a carga, uma vez que muitas requisições em intervalos de

tempo menores podem contribuir para sobrecarregar servidores e até mesmo derruba-los. As políticas de atendimento e monitoração são bastante rigorosas, para evitar esses tipos de problema. Caso essas políticas não sejam respeitadas, o crawler pode acabar indo para a lista negra no servidor ou mesmo podem bloquear o IP da máquina que faz as requisições.

As duas principais arquiteturas encontradas na literatura para *crawlers* distribuídos são os modelos cliente-servidor e o *peer-to-peer* e cada uma dessas arquiteturas tem suas vantagens e desafios. O modelo cliente-servidor constitui-se de um servidor que contém uma lista de *urls* que ainda não foram visitadas. De tempos em tempos, uma *url* é repassadas a um cliente que solicita, que por sua vez, busca os conteúdos requisitados e envia o que foi coletado ao servidor. O servidor quando recebe o que foi coletado pelo cliente, marca em sua lista que tal *url* foi coletada ou visitada, extrai as *urls* presentes no conteúdo coletado e adiciona em sua lista como não visitados e repete o processo enviando um novo *url* a um cliente. Em relação a arquitetura *peer-to-peer*, basicamente ocorre a divisão da lista de *urls* entre as demais máquinas. Cada servidor fica responsável por rastrear um subconjunto de *url* da lista inicial. É importante chamar a atenção que essa arquitetura não é escalável, uma vez que, no caso de aumentarmos o número de máquinas dentro dessa rede, é necessário recalcular a porção de *urls* e redistribuir os subconjuntos. Apesar disso, a arquitetura *peer-to-peer* tem a vantagem de ser mais fácil de se implementar que a cliente-servidor, onde é preciso modelar o comportamento de dois módulos distintos.

2.3. Descoberta de Conhecimento em Banco de Dados

A Internet é a maior fonte presente na atualidade quando falamos em obtenção de dados e informação. De acordo com o autor de [BUENO and VIANA 2012] a última década tem sido marcada por avanços significativos na computação e na tecnologia em geral, de tal modo, isso tem impulsionado cada vez mais a produção de dados, que só tende a crescer. Esse cenário se deve, principalmente ao surgimento de redes sociais, fóruns e páginas com os mais diversos tipos de conteúdo, que são disseminados na web. Isso se justifica, uma vez que esses ambientes permitem que seus usuários criem e compartilhem rapidamente vários tipos de conteúdo. Para o autor de [SILVA and BRANDÃO 2015] todos esses meios de interação virtuais respondem ou suprem a necessidade de informação de diferentes grupos de utilizadores que por sua vez possuem múltiplos interesses, tais como relacionamento, emprego e negócios. Ainda segundo [BUENO and VIANA 2012], o valor desses dados está diretamente ligado á capacidade de extrair informações em mais alto nível. Isso significa que para o cenário presente apenas armazenar e recuperar grandes quantidades de dados de maneira eficiente não engloba todas as vantagens possíveis quando se trata da manipulação dos dados.

Podemos facilmente comprovar que o valor dos dados está justamente na capacidade de extrair conhecimento deles, quando pensamos nas potenciais aplicações do conhecimento. Em conformidade com o autor [da Costa Côrtes et al. 2002], o conhecimento pode ser aplicado em vários contextos. Por exemplo, no marketing o comportamento do consumidor pode ser analisado, com base em nas transações de compra, e assim, os padrões descobertos, podem ser utilizados para promover melhorias na tomada de decisão quanto a propagandas, localização de produtos e lojas, dentre outras aplicações. No setor de finanças, a aplicação do conhecimento adquirido pode servir para avaliar a concessão de crédito a clientes; analisar o desempenho de investimentos como ações, bônus, fundos;

avaliar opções financeiras em alta e detectar fraudes. No setor das industrias, o conhecimento pode ser aplicado na otimização de recursos como força de trabalho e matéria prima, além do projeto de processos de produção. Na área da saúde, o conhecimento extraído de grandes bases medicas pode ser usado para, analisar a eficácia de um tratamento, otimização de processos dentro do hospital, análise de efeitos colaterais de drogas, entre outros. São inúmeras aplicações, que podem ser importantes para prover melhoras significativas em processos dentro de qualquer organização que realize um certo controle dos dados.

É nesse panorama que surge a Busca de Conhecimento em Banco de Dados ou *Knowledge Discovery in Database* ou KDD. Conforme [Ruas 2010], o KDD é um processo não trivial que identifica padrões válidos, novos e potencialmente úteis que estão escondidos nos dados. De acordo com [BUENO and VIANA 2012] o processo de KDD é dividido em cinco principais etapas: seleção, pré-processamento, transformação, mineração de dados e interpretação dos resultados. Em contrapartida, de acordo com [Ruas 2010], o processo engloba ainda mais duas etapas:a compreensão do domínio de aplicação antes da seleção e a consolidação do conhecimento, após a interpretação dos resultados.

A etapa de compreensão do domínio de aplicação em acordo com [Ruas 2010] é a fase em que realizamos a definição dos objetivos e definimos quais os requisitos da extração, identificamos as prováveis fontes de dados, a distribuição do conhecimento extraído e realizamos o estudo de viabilidade e custos da aplicação do processo. Esta é a fase de se entender completamente o âmbito de aplicação do conhecimento que se deseja estudar, pesquisar sobre as fontes de obtenção dos dados, analisar o custo e viabilidade do processo para que não tenhamos surpresas. Existem custos temporais e econômicos nos processo de estudo e dependendo do caso devem ser avaliado com cautela para evitar transtornos. Após essa fase de compreensão os dados são preparados para a fase seleção de dados.

De acordo com [BUENO and VIANA 2012] e [Ruas 2010], a seleção de dados é a etapa em que fazemos a separação do conjunto de dados que serão relevantes para o estudo,ou seja, somente as informações que são importantes para o domínio da aplicação, são usadas.

A fase de pré-processamento, segundo [BUENO and VIANA 2012], é aquela responsável pela limpeza dos dados e seleção de atributos importantes para realizar a montagem da base de dados. Nessa fase, dados ausentes, informações incorretas ou discrepantes são corrigidas na base de dados para não comprometer a qualidade dos modelos que serão extraídos ao final do processo de KDD. Essa fase é importante, pois pode prejudicar a inferência e tornar os dados tendenciosos, caso não seja tratada da maneira adequada. Ao final dessa etapa iniciamos a fase de transformação dos dados.

A etapa de transformação nada mais é do que a fase onde os dados são modificados e reorganizados para depois ser aplicada a mineração dos dados. Os dados são reorganizados para que possam ser interpretados por algum software próprio para investigações da mineração de dados.

Na etapa da mineração de dados, é onde de fato ocorre a extração do potencial conhecimento presente na base de dados. Segundo [Ruas 2010], a mineração de dados

é uma sequência de etapas dentro do processo de KDD, que pode envolver repetidas iterações da aplicação de um determinado algoritmo, com o propósito de extrair padrões dos dados. As técnicas que são utilizadas para criar modelos utilizando os dados vem de áreas como: Aprendizado de Máquina, Reconhecimento de Padrões e Estatística. De acordo com [da Costa Côrtes et al. 2002] a mineração de dados é a exploração e análise de uma grande quantidade de dados, de maneira automática ou semi-automática, com o intuito de descobrir regras ou padrões que sejam interessantes. Posteriormente, temos a fase de interpretação e avaliação dos padrões descobertos, onde podemos descobrir novos fatos, relacionamentos que podem ser utilizados na pesquisa e otimização do problema que está sendo abordado.

Após todas essas etapas, temos a fase de consolidação do conhecimento que foi obtido. Nessa fase, utilizamos técnicas de visualização e representação do conhecimento que foi adquirido. Todos os passos anteriormente citados, constituem o processo busca de conhecimento em base de dados ou KDD.

2.3.1. Mineração de Dados: Tarefas x Técnicas

A mineração de dados é uma das etapas do processo de KDD. Para [DEL-FIACO 2012], é importante definir o quais objetivos da pesquisa queremos alcançar para escolher melhor a tarefa mineração de dados, bem como a técnica que será utilizada para alcançar tal objetivo. Para [BUENO and VIANA 2012], a tarefa nada mais é do que a especificação do que queremos buscar nos dados, ou seja, categoria de padrões temos interesse em encontrar. Enquanto a tarefa de mineração diz respeito à especificação do método que vai nos ajudar a descobrir os padrões que nos interessam, a técnica se refere ao(s) algoritmo(s) que serão utilizados para explorar os dados e descobrir os padrões relevantes de acordo com o problema abordado. Segundo [BUENO and VIANA 2012], de maneira geral, podemos analisar os dados utilizando basicamente três tipos de análise: amostragem, descritiva e prognóstico.

A análise ou tarefa de amostragem, basicamente tem como objetivo descobrir padrões que fogem muito do comportamento geral dos dados, com isso a confiabilidade da amostragem e dos resultados aumenta. Existem dois subtipos ligados a essa tarefa, são eles: detecção de desvios e análise de desvios. O objetivo principal da detecção de desvios é encontrar informações que destoam do padrão de comportamento geral do modelo de dados. No caso da análise de desvios, semelhante a detecção, o objetivo é determinar se um dado foge ou não ao comportamento, que é previamente estabelecido antes de se iniciar o estudo.

A análise ou tarefa descritiva tem como principais objetivos estabelecer relações, associações, descrevendo e caracterizando o modelo para encontrar informações relevantes escondidas em meio os dados. Essa tarefa, possui como subtipos de tarefas a classificação, a associação, o agrupamento, a descrição, a detecção de sequências e a segmentação. A classificação basicamente, procura categorizar os dados em classes previamente definidas de acordo com a similaridade de alguma característica dos dados. A associação procura identificar grupos de fatos que ocorrem em conjunto ou de forma condicionada, ou seja, este tipo de tarefa procura associações e relacionamento entre os dados que normalmente são expressos por regras de associação. Uma regra de

associação é expressa da forma X-¿Y (se X então Y), isso quer dizer que em uma base de dados se uma dada transação X ocorrer então Y também tende a ocorrer. O agrupamento procura formar grupos de acordo com a similaridade dos atributos que foram definidos para realizar a categorização. Ao contrario da classificação, no agrupamento não é necessário definir as classes previamente, mas sim os atributos que serão usados para realizar categorização. A descrição procura descrever as características ou particularidades que ocorrem frequentemente nos dados. A detecção de sequências, procura estabelecer relacionamentos temporais entre os fatos. A segmentação nada mais é do que a divisão dos dados em conjuntos menores através da formação de grupos levando em consideração alguma distinção. Geralmente a segmentação vem seguida da aplicação de uma nova tarefa sobre os dados segmentados.

A análise de prognóstico busca inferir um valor, algum comportamento futuro ou estimar valores desconhecidos utilizando as informações obtidas na análise descritiva. A estimação tem como objetivo estimar um valor desconhecido a partir de valores conhecidos. Enquanto que a predição é o processo de predizer determinado valor em um instante futuro baseado em valores conhecidos.

Na figura abaixo é possível visualizar algumas das principais tarefas existentes na mineração de dados, bem como seus objetivos, as principais técnicas e aplicações.A análise de prognóstico procura inferir um valor ou comportamento futuro ou estimar valores desconhecidos utilizando as informações obtidas na análise descritiva. A estimação tem como objetivo estimar um valor desconhecido a partir de valores conhecidos. Enquanto que a predição é o processo de predizer determinado valor em um instante futuro baseado em valores conhecidos.

Na figura abaixo, é possível visualizar algumas das principais tarefas existentes na mineração de dados, bem como seus objetivos, as principais técnicas e aplicações.

Tarefa	Objetivo(s)	Técnica(s)	Aplicação(ões)
Classificação	Categorizar um conjunto de dados	• Árvores de Decisão; • Classificação bayseana; • Redes Neurais; • Regras de Associação;	• Classificar pedidos de créditos; • Classificar perfis de clientes de uma loja; • Identificar qual a melhor forma de tratamento de um paciente;
Estimativa (ou Regressão)	Definir um valor para alguma variável contínua desconhecida	• Regressão Linear; • Regressão Múltipla; • Regressão não linear; • Regressão Logística;	• Estimar o nº de filhos ou a renda total de uma família; • Estimar o valor em tempo de vida de um cliente;
Associação	Determinar quais itens tendem a co-ocorrerem em uma mesma transação	• Regras de Associação	• Determinar quais produtos costumam ser colocados juntos em um carrinho de supermercado;
Segmentação	Processo de partição de uma população heterogênea em vários grupos mais homogêneos	• Indução por Árvores de Decisão	• Agrupar clientes por região do país; • Agrupar cliente com comportamento de compra similar;
Sumarização	Encontrar uma descrição compacta para um subconjunto de dados	• Agregações e gráficos diversos	• Tabular o significado e desvios padrão para todos os itens de dados; • Derivar regras de síntese;

Figure 1. Adaptado a partir de [BUENO and VIANA 2012] e [da Costa Côrtes et al. 2002]

2.4. Aprendizado de Máquina e Algumas Técnicas

Segundo [MACHADO et al. 2012], aprendizado de máquina pode ser descrito como o desenvolvimento de técnicas computacionais, sobre o aprendizado, bem como a construção de sistemas que são capazes de adquirir conhecimento automaticamente. Ainda para [MACHADO et al. 2012], existem três tipos de *maquine learning* ou aprendizagem máquina:

- **Aprendizado Supervisionado**: são utilizadas entradas para treinar o algoritmo e reconhecer o padrão;
- **Aprendizado Não-Supervisionado**: não utiliza entradas para treinamento do algoritmo. Os padrões são encontrados através do agrupamento de dados semelhantes;
- **Aprendizado por Reforço**: o comportamento da rede é avaliado com base em algum critério numérico, fornecido em instantes espaçados de tempo;

2.4.1. Aprendizado Supervisionado: Arvores de Decisão e Backpropagation

2.4.2. Arvore de Decisão

2.4.3. Backpropagation

2.4.4. Aprendizado Não-Supervisionado: K-Means e SOM

2.4.5. K-Means

2.4.6. SOM

3. Metodologia

A metodologia do trabalho esta fundamentada no processo de descoberta de conhecimento em banco de dados. O projeto do experimento se iniciou com o desenvolvimento de um *crawler* do tipo cliente-servidor e arquitetura paralela e distribuída, utilizando a linguagem de programação Java. O módulo de servidor era responsável por enviar *urls* das páginas aos clientes que solicitavam e receber as páginas. A lista de *urls* era atualizada pelos clientes. O cliente recebia uma *url*, marcava ela como visitada, fazendo uma comunicação com o servidor e então efetuava o download desse perfil. Além disso, o cliente extraia os *urls* presentes no arquivo e repetia o processo até que o servidor do *LinkedIn* recusava o atendimento as requisições, devido as restrições do servidor. A lógica que foi implementada por trás do cliente, provém do algoritmo de busca em largura, para que pudéssemos alcançar a maior amplitude possível no grafo do *LinkedIn* para o campo "As pessoas também viram", disponível nos perfis públicos das páginas do *LinkedIn*, uma vez que, não permitem acesso aos dados de contato no perfil público. Além disso, o servidor contava com um banco de dados com uma tabela para armazenar as *urls* novas que ainda não haviam sido visitadas, um identificador numérico, IP da máquina que coletou a *url*, país declarado pelo profissional, e um campo para marcar a página como alocada e outro para indicar que ela foi visitada.

A coleta teve inicio em janeiro de 2016 e foi até julho de 2016, onde foi coletado 120.000 perfis e desses 87.602 foram de brasileiros. Posteriormente, desenvolvemos uma ferramenta para extrair as informações importantes ao estudo, para um banco de dados, utilizando também Java e o banco de dados utilizado foi o PostgreSQL 9.3. Também foi utilizada a biblioteca *Jsoup*, que é um parser desenvolvido em Java, para se trabalhar com arquivos *".html?*. A extração das informações, só ocorreu mediante a separação de todos os perfis brasileiros. Essa separação ocorreu com base na coluna país, na tabela de coleta, campo esse que armazenou o país declarado pelo profissional em seu perfil. Abaixo, podemos ver o modelo relacional do banco de dados:

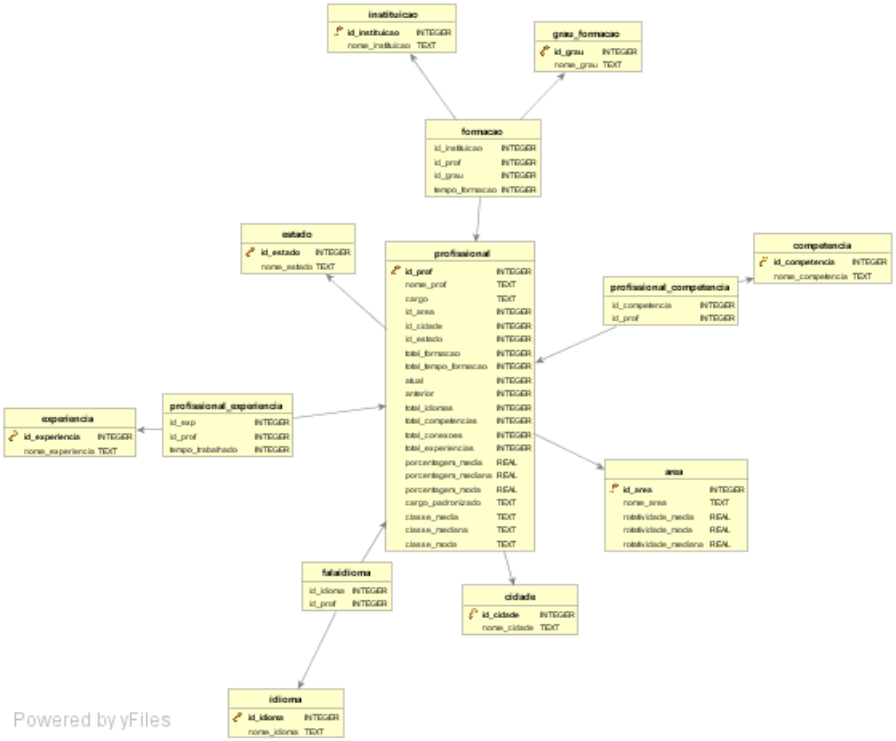

Figure 2. Próprio Autor

Após a coleta e extração das informações, foi necessário realizar o tratamento das inconsistências presentes na base de dados. Alguns problemas encontrados que podem ser citados são: erros de codificação de algumas páginas, inconsistências, poluição, atributos duplicados, campos com valores em branco e desbalanceamento de classes. Os erros de codificação tiveram que ser corrigidos manualmente, nos atributos nominais, área, cidade, estado e cargo. Os tipos de inconsistência mais comuns encontrados na base foram erros de ortografia, sinônimos e dados incorretos ou inválidos. Todos esses problemas foram devidamente tratados. O tipo de poluição encontrado na base, diz respeito a páginas de : spam ou perfis falsos, empresas, instituições de ensino, ongs, comércios, ou seja, toda e qualquer página que não seja de um perfil profissional. Atributos e dados duplicados foram removidos. E valores em branco foram preenchidos como sendo desconhecidos. As classes desbalanceadas foram balanceadas quando necessário, dependendo do experimento que foi realizado.

Após esse tratamento de pré-processamento, com o intuito de caracterizar a base de dados realizamos uma análise estatística dos principais atributos presentes na base de dados para que possamos embasar o experimento de mineração de dados. Foi dada uma ênfase no atributo tempo trabalhado na experiência. Esse atributo diz respeito ao tempo que um dado profissional ficou em uma experiência. As experiências foram todas agrupadas por área e então calculado os valores de média, mediana e moda, para cada área. Além disso, criamos três classificações distintas baseadas nesses valores de média, mediana e moda. Basicamente pegamos experiência por experiência de cada profissional e de acordo com o valor encontrado de média, mediana e moda, por área, calculamos a porcentagem de experiências que estavam menores ou iguais a esses valores.

Depois da preparação da base de dados, escolhemos as tarefa(s) de mineração, bem como da(s) técnica(s) que foram utilizadas. O estudo foi dividido na utilização de aprendizado supervisionado e não supervisionado. As tarefas escolhidas foram a classificação e a clusterização. Enquanto que os algoritmos escolhidos para realizar a experimento foram: Árvore de Decisão, Backpropagation, K-Means e SOM. Para realizar esse estudo utilizamos a ferramenta Weka, que conta com uma coleção de algoritmos implementados para a finalidade de mineração de dados. Por fim, realizamos a análise dos resultados e levantamento, validação e consolidação dos padrões encontrados na base de dados.

4. Discussão dos Resultados

5. Conclusão e Trabalhos Futuros

6. Referências

References

BEVENUTO, Fabrício, Almeida, J., and Silva, A. (2011). Coleta e análise de grandes bases de dados de redes sociais online. *Jornadas de Atualização em Informática (JAI)*, pages 11–57.

BUENO, M. F. and VIANA, M. R. (2012). Mineração de dados aplicações, eficiência e usabilidade. *INCITEL*.

da Costa Côrtes, S., Porcaro, R. M., and Lifschitz, S. (2002). Mineração de dados - técnicas, funcionalidades e abordagens.

DEL-FIACO, R. D. C. (2012). Aplicção da mineração de dados na descoberta de padrões do perfil de alunos do curso de si-unucet-ueg.

DIEESE (2014). Rotatividade setorial dados e diretrizes para a ação sindical.

FAYYAD, U., Piatetsky-Shapiro, G., and Smyth, P. (1996). The kdd process for extracting useful knowledge from volumes of data. *Communications of the ACM*, 39(11):27–34.

FONTAINHAS Mariana, M. C. (2013). O colaborador é uma parte de extrema importância nas engrenagens de uma empresa. [Online; accessed 8-março-2016].

G1, J. N. (2014). Governo aumenta tempo mÃnimo para solicitar seguro-desemprego. [Online; accessed 8-março-2016].

HAMMES, C. C. F., dos SANTOS, A. J., and MELIM, J. M. (2016). Os impactos do turnover para as organizações. *Revista ESPACIOS— Vol. 37 (Nº 03) Año 2016*.

JAIWEI, H. and KAMBER, M. (2006). Data mining: concepts and techniques.

LINKEDIN (2015). Linkedin. [Online; accessed 8-março-2016].

LÚCIO, C. G. (2015). A rotatividade no mercado de trabalho no brasil. [Online; accessed 8-março-2016].

MACHADO, V. P., de LIMA, B. V., and Araújo, S. W. (2012). Classificação automática de usuários de uma rede social utilizando algoritmos não-supervisionados.

Nery, P. F. (2015). O que é rotatividade (e por que é um problema)? [Online; accessed 8-março-2016].

Palhoça, S. (2016). Saiba tudo sobre gestÃ£o de recursos humanos. [Online; accessed 8-março-2016].

Park, S. and Lee, Y. (2014). Implementation of a distributed web community crawler. In *Network Operations and Management Symposium (APNOMS), 2014 16th Asia-Pacific*, pages 1–6. IEEE.

PINHEIRO, A. P. and SOUZA, D. A. D. (2013). "causas e efeitos da rotatividade de pessoal / turnover: Estudo de caso de uma microempresa do setor de educação". Simpósio de Excelência em Gestão de Tecnologia.

Portal-Brasil (2014). Mte discute a rotatividade no mercado de trabalho brasileiro. [Online; accessed 8-março-2016].

RHPORTAL (2015). Rotatividade de pessoal (turn-over) e absenteísmo. [Online; accessed 8-março-2016].

Ruas, T. L. (2010). Mineração de dados em redes complexas: Um estudo sobre a dinâmica do conteúdo social.

SILVA, P. R. C. and BRANDÃO, W. C. (2015). Arppa: Mining professional profiles from linkedin using association rules. *eKNOW*.

SIQUEIRA, M. (2014). *Novas Medidas do Comportamento Organizacional: Ferramentas de Diagnóstico e de Gestão*. Artmed Editora.

VASCONCELOS, L. d. and CARVALHO, C. d. (2004). Aplicação de regras de associação para mineração de dados na web. *Relatório Técnico*.

WIKIPÉDIA (2015). Rotatividade de pessoal — wikipédia, a enciclopédia livre. [Online; accessed 8-março-2016].

XAVIER, O. C. and de CARVALHO, C. L. (2011). Desenvolvimento de aplicações sociais a partir de apis em redes sociais online. *Relatório Técnico, UFG, Goiânia*.