UNIVERSIDADE FEDERAL DO CEARÁ
CAMPUS DE QUIXADÁ
BACHARELADO EM ENGENHARIA DE SOFTWARE

FRANÇOES DA SILVA PEREIRA

DESCOBERTA DE PERFIS PROFISSIONAIS DOS USUÁRIOS DO SITE STACKOVERFLOW

QUIXADÁ

2017

DESCOBERTA DE PERFIS PROFISSIONAIS DOS USUÁRIOS DO SITE STACKOVERFLOW

Trabalho de Conclusão de Curso submetido à Coordenação do Curso Bacharelado em Engenharia de Software da Universidade Federal do Ceará, como requisito parcial à obtenção do grau de bacharel. Área de concentração: Computação.

Orientadora: Profª. Dra. Ticiana Linhares Coelho da Silva.

QUIXADÁ

2017

Dados Internacionais de Catalogação na Publicação
Universidade Federal do Ceará
Biblioteca Universitária
Gerada automaticamente pelo módulo Catalog, mediante os dados fornecidos pelo(a) autor(a)

P491d Pereira, Françoes da Silva.
Descoberta de perfis profissionais dos usuários do site StackOverflow / Françoes da Silva Pereira. – 2017.
47 f. : il. color.

Trabalho de Conclusão de Curso (graduação) – Universidade Federal do Ceará, Campus de Quixadá, Curso de Engenharia de Software, Quixadá, 2017.
Orientação: Profa. Dra. Ticiana Linhares Coelho da Silva.

1. Mineração de dados (Computação). 2. Cluster (Sistema de computador). I. Título.

CDD 005.1

FRANÇOES DA SILVA PEREIRA

DESCOBERTA DE PERFIS PROFISSIONAIS DOS USUÁRIOS DO SITE STACKOVERFLOW

Trabalho de Conclusão de Curso submetido à Coordenação do Curso Bacharelado em Engenharia de Software da Universidade Federal do Ceará, como requisito parcial à obtenção do grau de bacharel. Área de concentração: Computação.

Aprovada em:___/___/_____.

BANCA EXAMINADORA

_____
Profa. Dra. Ticiana Linhares Coelho da Silva (Orientadora)
Universidade Federal do Ceará (UFC)

_____
Profa. Ma. Lívia Almada Cruz Rafael
Universidade Federal do Ceará (UFC)

_____
Prof. Dr. Paulo de Tarso Guerra Oliveira
Universidade Federal do Ceará (UFC)

A Deus por sua graça e providencia em todas as minhas necessidades. A minha mãe, Jeovani da Silva Pereira pelo carinho e apoio.

AGRADECIMENTOS

Agradeço à Deus por ter me agraciado com essa graduação e me guiado e sustentado em todas as minhas necessidades, confortando e suprindo todo o necessário em todo momento para que eu superasse cada desafio, me dando sabedoria e força quando estas me faltaram.

Aos meus pais, Jeovani da Silva Pereira e Francisco Helvio Pereira, pela paciência, carinho e orações que me confortaram e me motivaram a continuar. Aos meus familiares, Maria Cambé de Lima, André Cambé de Lima e Andréa Cambé de Lima, que me acolheram em sua residência no inicio dessa trajetória, sendo pacientes e atenciosos para comigo.

A todos os professores da Universidade Federal do Ceará campus Quixadá, o qual tive o prazer de ser instruído e crescer profissional e pessoalmente. Em especial a minha orientadora, professora Dra. Ticiana Linhares, não apenas prestou orientação para o trabalho acadêmico, mas demonstrou muita dedicação, paciência e excepcional competência em toda produção do trabalho. Ao professor Dr. David Sena Oliveira, meu tutor na bolsa de extensão PREX por dois anos, que contribuiu muito na minha evolução acadêmica.

À CAPES, pelo apoio financeiro com a manutenção da bolsa de auxílio, fundamental para que eu pudesse me manter. Aos meus colegas de faculdade pelos momentos especiais, pelas vitorias frente aos desafios que encaramos juntos e pelo apoio.

“Ó profundidade das riquezas, tanto da sabedoria, como da ciência de Deus! Quão insondáveis são os seus juízos, e quão inescrutáveis os seus caminhos! Porque dele e por ele, e para ele, são todas as coisas; glória, pois, a ele eternamente. Amém” (Romanos 11:33 e 36)

RESUMO

Sites de compartilhamento de conhecimento baseado em perguntas e respostas vem apresentando um crescimento considerável nos últimos anos, tanto em número de usuários, como em conteúdo. Um destes sites é o StackOverflow, que possui mais de 1 milhão de membros, abordando diversos temas na área de tecnologia da informação. Esse grande volume de informações gerados pelos usuários no site têm produzido vários estudos e análises dos dados que identificam padrões, relações e tendências que só são percebidas através de um estudo mais profundo. A partir dessas informações é possível tomar importantes decisões de negócio. Todavia essa analise não é uma tarefa fácil. Este trabalho, analisa os dados dos *surveys* do site StackOverflow a fim de identificar padrões e perfis profissionais dos seus usuários utilizando o processo de descoberta de conhecimento em banco de dados. São aplicadas técnicas de limpeza e pré-processamento dos dados, seguido da clusterização utilizando o algoritmo DBSCAN, validação dos padrões gerados e, por fim, a análise e descrição dos resultados encontrados.

Palavras-chave: Descoberta de informação em banco de dados. Clusterização. StackOverflow.

ABSTRACT

Question and answer-based knowledge sharing sites have been growing considerably in recent years, both in terms of number of users and content. One of these sites is the StackOverflow, that has more than 1 million members, addressing several topics in the area of information technology. This large volume of information generated by the users on the site, has produced several studies and analyzes of the data that identify patterns, relationships and trends that are only perceived through a deeper analysis. From this information you can take important business decisions. However, this analysis is not an easy task. This work analyzes the data from the *surveys* of the StackOverflow site to identify the patterns and professional profiles of its users using the process of Knowledge Discovery in database. Techniques of cleaning and pre-processing of the data are applied, followed by clustering using the DBSCAN algorithm, validation of the generated patterns and, finally, the analysis and description of the results found.

Keywords: knowledge-discovery in databases. Clustering. StackOverflow.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

Sites de compartilhamento de conhecimento baseado em perguntas e respostas vem apresentando um crescimento considerável nos últimos anos, tanto em número de usuários, como em conteúdo. Um destes sites de compartilhamento de conhecimento, StackOverflow[1], conta com uma base de mais de 1 milhão de membros, abordando os mais diversos temas na área de tecnologia da informação (MOVSHOVITZ-ATTIASet al., 2013). O grande volume de informações que são abordados por essas mídias têm atraído a atenção de vários estudiosos de mineração de dados em diferentes estudos. Uma vez que a análise dos dados pode identificar padrões, relações e tendências que só são percebidas através de um estudo mais profundo e, que ajudam a tomar decisões de negócio importantes, às vezes vitais, de forma inteligente e baseada em evidências reais. Esse se torna então um tópico de grande importância.

Em sites de compartilhamento de conhecimento baseado em perguntas e respostas os usuários interagem entre si fazendo perguntas sobre temas específicos, respondem dúvidas de outros usuários, avaliam as perguntas e respostas feitas, e podem ainda responder à *surveys* sobre tecnologias, empregos, produtos, educação, entre outros. O que produz uma massa de dados imensa. Portanto, pode conter informações e padrões importantes que não são óbvios ou não estão explícitos, sendo necessário uma análise mais profunda. Essa análise pode transformar essa massa de dados bruto em informações úteis. Neste trabalho, as respostas do *survey* realizado pelo site StackOverflow[2] referente ao ano de 2016 foi selecionada como base de dados para análise.

Transformar dados brutos em informações úteis é o processo conhecido por descoberta de conhecimento em bancos de dados, abreviada como KDD (*Knowledge Discovery in Database*) (Silberschatz et al. 2012). De acordo com ELMASRI (2013), KDD compreende as fases de seleção dos dados, limpeza e enriquecimento dos dados, transformação, mineração de dados e por fim, o relato expondo as descobertas. O trabalho SILVA (2004) apresenta essas fases de forma bem detalhada, descrevendo os sete passos que compreendem esse processo.

---

[1] www.stackoverflow.com
[2] Disponível em: https://insights.stackoverflow.com/survey

A mineração de dados, que é uma parte central do KDD, engloba diversas técnicas. Essas técnicas permitem analisar grandes depósitos de dados com diversos objetivos. De forma geral é possível entender os objetivos da mineração de dados como classificação de dados, realizar previsão, identificação de padrões e otimização de dados (ELMASRI, 2013).

Uma dentre as diversas técnicas de mineração de dados é a clusterização. Essa técnica agrupa registros com características semelhantes e que se relacionem entre si formando um grupo, que são chamados de *cluster*. Ao mesmo tempo que registros de grupos diferentes são altamente dissimilares. Geralmente, os *clusters* costumam ser disjuntos (ELMASRI, 2013).

O problema abordado neste trabalho é encontrar perfis de usuários do StackOverflow baseado em características, tais como: 1) gênero; 2) idade; 3) salário; 4) experiência; 5) ocupação; e por último 6) satisfação com o emprego.

Para alcançar este objetivo é realizada uma análise das respostas do *survey* do site StackOverflow no ano 2016[3], identificando padrões entre as respostas dos usuários. Para isso, foi utilizada a técnica de clusterização de mineração de dados seguindo o processo de KDD. A clusterização permitiu encontrar perfis frequentes, ou seja, cadeias de respostas similares foram agrupadas, formando um possível perfil.

Os resultados são de grande importância para a equipe do site StackOverflow e sua comunidade em geral, além de entusiastas de tecnologia da informação na área de mineração de dados, uma vez que mostra os perfis de usuários descobertos de acordo com as técnicas utilizadas e os dados analisados.

A principal contribuição deste trabalho é a análise das respostas, identificando os padrões e os perfis no ano de 2016 referentes aos usuários do site StackOverflow.

O capitulo 2 apresentará a fundamentação teórica, explicando os conceitos fundamentais para a compreensão deste trabalho. O capitulo 3 descreve os trabalhos relacionados. O capitulo 4 descreve em detalhes o passo a passo da execução do processo adotado neste trabalho. O capitulo 5 apresenta os resultados obtidos e, o capitulo 6, a conclusão e os trabalhos futuros em ralação a este trabalho.

---

[3] https://insights.stackoverflow.com/survey/2016

2 FUNDAMENTAÇÃO TEÓRICA

A seguir, será apresentada uma revisão bibliográfica listando e definindo os conceitos principais para o desenvolvimento deste trabalho. A seção 2.1 define o que é descoberta de conhecimento em banco de dados. A seção 2.2 descreve em mais detalhes a mineração de dados que é a tarefa central deste trabalho, bem como é um dos passos do processo de descoberta de conhecimento em banco de dados. A seção 2.3 apresenta os passos de pré-processamento e transformação de dados, que é uma atividade fundamental na mineração de dados. A seção 2.4 define a clusterização e os principais algoritmos que podem ser utilizados no processo. A seção 2.5 detalha o DBSCAN que foi o algoritmo de clusterização utilizado neste trabalho.

2.1 Descoberta de conhecimento em bancos de dados

A transformação de dados brutos em informações úteis é o processo de descoberta de conhecimento em bancos de dados, abreviada como KDD (*Knowledge Discovery in Database*) (Silberschatz et al. 2012). Esse é um processo macro que têm a mineração de dados como um de seus vários passos, e apresenta um caráter exploratório, interativo e iterativo com base nas seguintes etapas (SILVA, 2004):

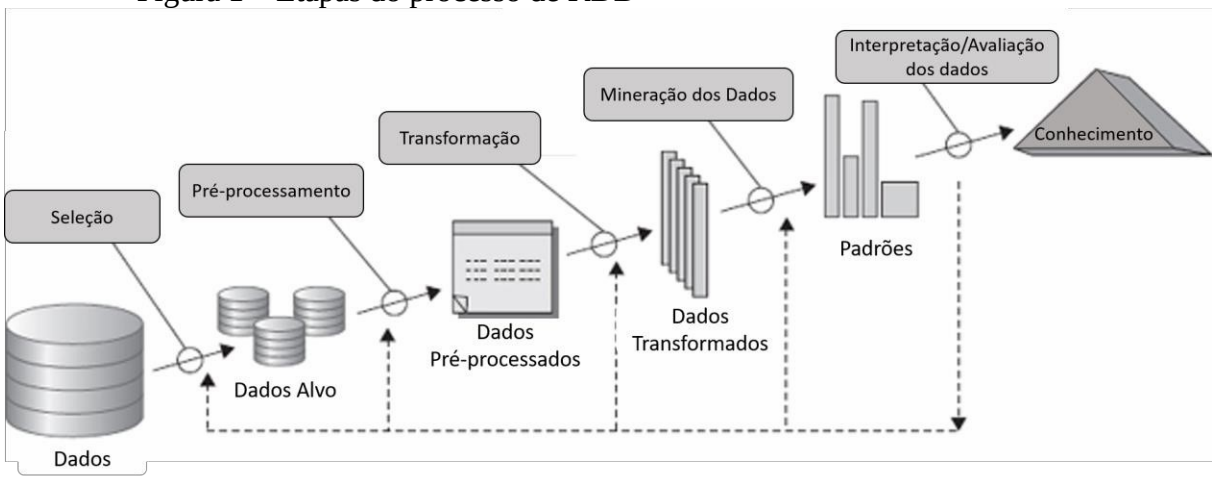
1. Definição do tipo de conhecimento a descobrir, o que pressupõe uma compreensão do domínio da aplicação bem como do tipo de decisão que tal conhecimento pode contribuir para melhorar.
2. Criação de um conjunto de dados alvo (*Selection*): selecionar um conjunto de dados, ou focar num subconjunto, onde a descoberta deve ser realizada.
3. Limpeza de dados e pré-processamento (*Preprocessing*): operações básicas tais como remoção de ruídos quando necessário, coleta da informação necessária para modelar ou estimar ruído, escolha de estratégias para manipular campos de dados ausentes, formatação de dados de forma a adequá-los à análise dos dados.
4. Redução de dados e projeção (*Transformation*): localização de características úteis para representar os dados dependendo do objetivo da tarefa, visando a redução do número de variáveis e/ou instâncias a serem

consideradas para o conjunto de dados, bem como o enriquecimento semântico das informações.

5. Mineração de dados (*Data Mining*): selecionar os métodos a serem utilizados para localizar padrões nos dados, seguida da efetiva busca por padrões de interesse numa forma particular de representação ou conjunto de representações; busca pelo melhor ajuste dos parâmetros do algoritmo para a tarefa em questão.
6. Interpretação dos padrões minerados (*Interpretation/Evaluation*), com um possível retorno aos passos 1-5 para posterior iteração.
7. Implantação do conhecimento descoberto (*Knowledge*): incorporar este conhecimento à performance do sistema, ou documentá-lo e reportá-lo às partes interessadas.

Este trabalho seguiu os passos de descoberta de conhecimento em bancos de dados descritos acima para a identificação dos perfis e relações entre os perfis de usuários nos dados analisados.

Figura 1 – Etapas do processo de KDD



Fonte: Fayyad et al. (1996), adaptado pelo autor.

2.2 Mineração de dados

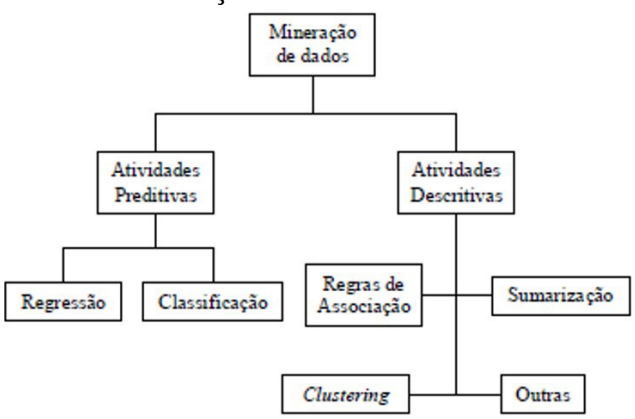
De acordo com ELMASRI (2013), KDD compreende as fases de seleção dos dados, limpeza e enriquecimento dos dados, transformação, mineração de dados e por fim, o relato expondo as descobertas. Todavia, nem todo processo de descoberta de informação em

banco de dados é considerado como mineração de dados. A utilização de um sistema gerenciador de banco de dados para encontrar registros individuais, por exemplo, não é mineração de dados. Esse processo pertence a outra área, a de recuperação de informação (TAN MICHAEL STEINBACH, 2009). A mineração de dados é capaz de melhorar o processo de recuperação de informações pois apresenta um conjunto de técnicas que permitem identificar informações a partir dos dados que não seria possível identificar com os meios tradicionais (FAYYAAD et al., 1996).

Ao longo dos anos os dados coletados passaram por mudanças. O aumento no volume dos dados que passaram atingir escalas cada vez maiores criou um problema de escalabilidade, dificultando a manipulação desses dados pelos algoritmos. A necessidade de superar esse e outros problemas no tratamento dos dados como, aumento nas dimensões, complexidade e heterogeneidade dos dados, foram os principais motivadores para o surgimento da mineração de dados (TAN MICHAEL STEINBACH, 2009). Há ainda um outro ponto importante para o tratamento dos dados que a mineração de dados aborda conhecido como pré-processamento, utilizado neste trabalho e discutido em mais detalhes na seção seguinte.

A mineração de dados possui basicamente duas classes de tarefas, prever e descrever. A tarefa de prever baseia-se em um ou mais atributos a fim de prever o valor de outro atributo. A tarefa de descrever busca padrões, regras, correlações, anomalias, grupos ou tendências a fim de resumir os relacionamentos nos dados (TRONCHONI et al., 2010). A mineração de dados, que é uma parte central em KDD, engloba diversas técnicas. Essas técnicas permitem analisar grandes depósitos de dados com diversos objetivos. De forma geral, é possível entender os objetivos da mineração de dados como classificação de dados, realizar previsão, identificação de padrões e otimização de dados (ELMASRI, 2013).

Figura 2 – Tarefas de mineração de dados



Fonte: Tronchoni et al. (2010)

Neste trabalho, a mineração de dados é realizada como uma atividade descritiva utilizando a técnica de clusterização com o objetivo de identificar padrões de perfis de usuários do site StackOverflow.

2.3 Pré-processamento e transformação de dados

O pré-processamento de dados é uma área vasta em estratégias e técnicas. A seguir, são mencionadas algumas dessas técnicas de pré-processamento e transformações de dados e quais foram utilizadas neste trabalho.

2.3.1 Pré-processamento dos dados

Bases de dados em geral são suscetíveis a conter erros ou ruídos que é um componente aleatório do erro de medição que distorce ou adiciona valores ilegítimos ao dado. Esses registros podem apresentar dados com valores estranhos, incompletos, inconsistentes ou ausentes (FAYYAD et al., 1996). Para melhorar a qualidade e expressão dos dados são realizados o pré-processamento e a transformação dos dados (SILVA, 2004). Além de melhorar a capacidade de análise dos dados, o pré-processamento e a transformação dos dados são necessários para adaptação destes à uma ferramenta ou técnica específica da mineração de dados (TAN MICHAEL STEINBACH, 2009). Neste trabalho, tanto a limpeza como a transformação são necessários para o processo de clusterização.

Para tratar ruídos e valores ausentes diversas ações são viáveis. Por exemplo, a remoção do registro, ou substituição do valor, seja manualmente, através de uma variável, uma constante ou por um valor mais provável. Esta última escolha visa evitar tendências ou viciar os dados, todavia, nem sempre é possível fazer essa substituição (DE MORAES, 2010). A interpolação, agrupamento ou regressão também são formas de tratamento que podem ser utilizadas para eliminar ruídos nos dados (SILVA, 2004).

Neste trabalho, por uma questão de simplicidade, foi adotada a técnica de remoção do registro quando um valor faltante é encontrado ou quando a substituição por um valor mais provável não é possível. A segui, é apresentado como os dados são geralmente transformados após a fase de limpeza, inclusive esta metodologia foi utilizada neste trabalho.

2.3.2 Transformação dos dados

Algoritmos de classificação em geral requerem que os dados tenham seus atributos categorizados. Por isso é necessário um processo de transformação nos dados. Esse processo é chamado de discretização. Algoritmos que identificam padrões de associação por sua vez requerem que os atributos estejam em formato binário. Processo conhecido como binarização. Portanto, independente do algoritmo os dados podem necessitar passar por uma ou outra forma de transformação (CORNELIUS JUNIOR, 2015).

Uma forma simples de binarização é mapear cada valor original do atributo a um inteiro único no intervalo de [0 até (m-1)], onde m é o número de valores distintos que o dado apresenta. Se o atributo for ordinal a ordem deve ser mantida. Em seguida, é necessário converter cada número para um valor binário que irá requerer um total de $[\log_2(m)]$ novas colunas de atributos para representar o valor original (TAN MICHAEL STEINBACH, 2009).

Para discretizar um atributo é necessário considerar o algoritmo a ser utilizado e os atributos a serem transformados. Essa tarefa envolve basicamente duas etapas, decidir quantas categorias e como mapear cada categoria. Uma forma simples de executar esses passos é definir uma largura igual ou proporcional para os valores de atributos e um intervalo para divisão desses valores (HAN et al., 2011).

Outra forma de discretização é a codificação por rótulos, no qual, é atribuído um inteiro único para cada valor único da coluna. Essa abordagem tem a desvantagem de dificultar interpretações que envolvem comparações, ou criar relações que não existem nos dados originais. Podemos ter por exemplo, uma coluna contendo 20 valores diferentes da altura de pessoas. Em seguida fazer a ordenação dos valores e a atribuição de inteiros únicos consecutivos a partir do valor 1. Embora a altura com valor 5 seja maior que a altura com valor 1, não significa que ela seja cinco vezes maior. A binarização também pode ser utilizada como discretização. Neste caso a coluna dos valores originais é substituída pelas colunas resultado da binarização (HAN et al., 2011).

Neste trabalho, foi utilizado o DBSCAN, um dos mais populares algoritmos de clusterização baseado em densidade (HAN et al., 2011). Além disso, foi adotada a abordagem de codificação por rótulos, onde cada valor foi mapeado para um inteiro único. Em seguida, foi aplicada a discretização em intervalos proporcionais aos valores dos registros. Como resultado final cada coluna passou a ter valores numéricos reais em um intervalo de [0.0 até

1.0] com largura proporcional a distância que os valores originais possuem entre si. A seção 5.4 ilustra o resultado dessa transformação dos dados.

2.4 Clusterização

Uma dentre as diversas técnicas de mineração de dados é a clusterização. Em algumas literaturas também referido como agrupamento. Neste trabalho será utilizado o termo clusterização. Em um *cluster* registros com características similares formam um grupo, ao mesmo tempo que registros de grupos diferentes são altamente dissimilares (ELMASRI, 2013). Neste trabalho, a clusterização é um processo fundamental para análise dos dados e descoberta de perfis.

Os *clusters* podem ser classificados em diversas formas. Quando as distâncias entre objetos de *clusters* distintos é alta, dizemos que o *cluster* é bem separado. *Clusters* baseados em protótipos apresentam suas medidas de similaridade baseadas em pontos centrais, ou representacionais dos objetos, chamados centróides ou medóides. Se os dados estão representados em formato de grafos, temos um *cluster* baseado em grafo. Outro tipo comum de *clusters* são os baseados em densidade, tais *clusters* possuem regiões de alta densidade e baixa densidade de objetos (HAN et al., 2011).

Um dos algoritmos utilizados para clusterização mais conhecidos é o *K-means* (HAN et al., 2011). Este algoritmo geralmente utiliza uma média de pontos para criar um centróide e define um protótipo. O usuário especifica o número de grupos desejados, K, a seguir, cada ponto é atribuído ao centróide mais próximo, formando um *cluster* a partir da coleção de pontos do centróide (TAN MICHAEL STEINBACH, 2009). Os *clusters* são atualizados conforme os pontos são atribuídos, até que ocorra a estabilização dos centróides, que pode ocorrer ao se atingir um número máximo de iterações, ou outra condição estabelecida no algoritmo. Para essa atribuição de pontos, diversos tipos de medidas podem ser adotadas a depender do tipo de dado analisado, como a distância Euclidiana para dados no espaço Euclidiano ou medida de *Jaccard* para documentos (TAN MICHAEL STEINBACH, 2009). A figura 3, a seguir ilustra o algoritmo K-means.

Figura 3 – Algoritmo K-Means

K-means

Entrada: **K** clusters, **n** objetos em um conjunto **D**

Saída: um conjunto contendo **K** clusters, que minimiza o erro quadrático com relação aos centros de gravidade de cada cluster

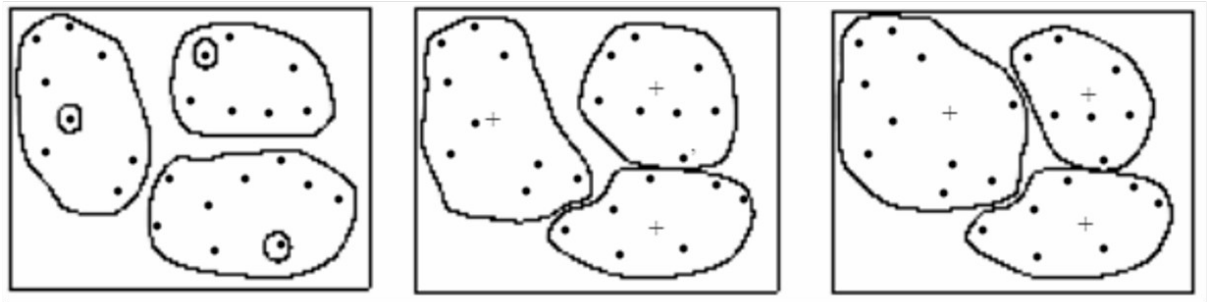
Algoritmo:

1 – Selecionar arbitrariamente **k** {*p1, p2,...pk*} objetos de **D** como centros iniciais dos clusters

2 – Atribuir cada objeto O diferente de *pi*, ao cluster com centro mais próximo

3 – Quando todo os objetos forem agrupados, recalcular o centros de cada cluster

4 – Repetir os passos 2 e 3 até que os cluster se estabilizem, não havendo mudanças de centro ou objetos mudando de cluster



Primeira iteração **Segunda iteração** **Terceira iteração**

Fonte: elaborada pelo autor.

A clusterização hierárquica aglomerativa é outro algoritmo de clusterização importante. Esse algoritmo possui duas abordagens básicas. A primeira, utiliza uma definição de proximidade, e em cada iteração toma os pontos individuais mais próximos e junta esses pontos dois-a-dois. A segunda abordagem, inicia com um grupo, depois define critérios de divisão e então segue realizando divisões sucessivas, até restarem apenas grupos únicos (HAN et al., 2011).

Outro algoritmo de clusterização muito conhecido, o DBSCAN, é baseado em densidade. Ele possui dois parâmetros fundamentais, o primeiro, *eps*, que representa tamanho

da vizinhança. O segundo, *minPts*, determina a quantidade mínima de pontos dentro do raio *eps*. O DBSCAN então usa esses dois valores para classificar os dados em ponto central, pontos de limite ou como pontos de ruídos (TAN MICHAEL STEINBACH, 2009).

Dentre os três algoritmos citados acima, os mais populares são o *K-means* e o DBSCAN (HAN et al., 2011). O *K-means* quando comparado com o DBSCAN, requer que o número de *clusters* seja informado antecipadamente, enquanto que o DBSCAN identifica o número dos *clusters* com base nos parâmetros informados. O *K-means* também é mais sensível aos *outliers* e ruídos que o DBSCAN (TAN MICHAEL STEINBACH, 2009). Ambos os algoritmos são sensíveis aos valores de seus parâmetros. Quanto a performance, o *K-means* é superior ao DBSCAN. Apesar de inferior em desempenho, o DBSCAN, é mais preciso quando os dados apresentam alto nível de ruídos e densidade de *outliers* que o *K-means* (DUDIK et al., 2015). Por esta razão ele foi o algoritmo adotado neste trabalho.

Neste trabalho, o algoritmo DBSCAN foi utilizado para clusterizar os dados dos *survey* do site *StackOverflow* do ano de 2016 objetivando identificar perfis de usuários.

2.5 DBSCAN

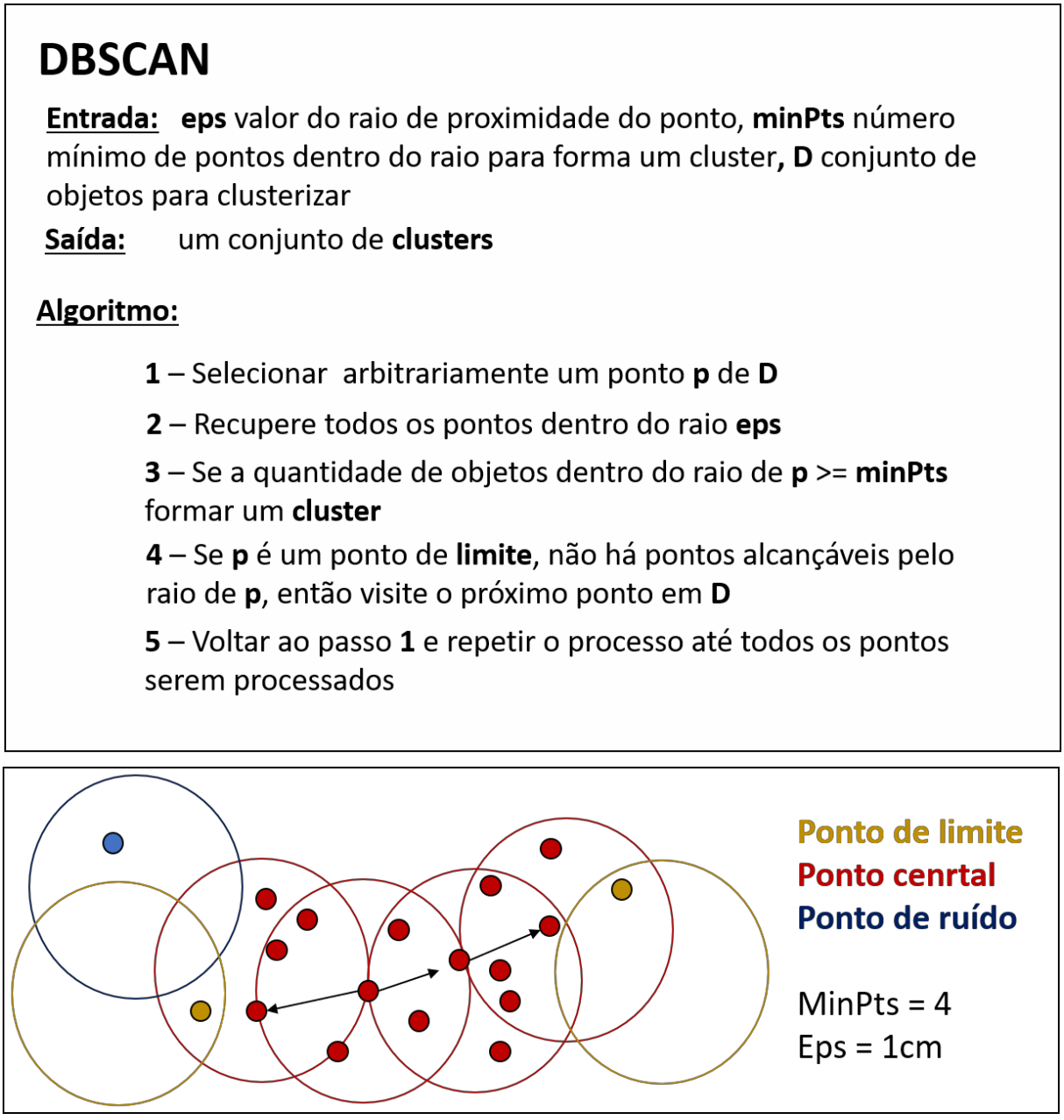
O DBSCAN é um algoritmo baseado em densidade. Essa densidade é aferida em função do número de pontos dentro de um determinado raio referente a um ponto. Ou seja, são contabilizados todos os pontos dentro do raio de um determinado ponto central, incluindo o próprio ponto. Esse raio, denominado *eps*, é um dos parâmetros do DBSCAN (TAN MICHAEL STEINBACH, 2009). Se o *eps* for suficientemente grande todos os pontos estarão dentro do raio. No caso oposto, nenhum ponto ficará dentro do raio de proximidade, exceto ele próprio, indicando que não há similaridade entre os demais pontos. Dessa forma torna-se essencial a atribuição de um valor apropriado para o *eps* a fim de uma correta clusterização dos dados.

O segundo parâmetro do DBSCAN, *minPts*, determina a quantidade mínima de pontos dentro do raio *eps*, ou seja, ele determina quantos pontos, no mínimo, devem estar dentro do raio *eps* para que um *cluster* seja criado. O DBSCAN então usa esses dois valores para clusterizar os dados.

Quando um ponto possui, dentro do raio de seu *eps*, uma quantidade de pontos maior ou igual ao *minPts* ele é classificado como central. Um ponto que esteja dentro do *eps*

de um ponto central é classificado como ponto de limite. Caso nenhuma das condições anteriores se apliquem ao ponto ele é classificado como ponto de ruído, um *outlier*, não pertencendo a nenhum *cluster* (TAN MICHAEL STEINBACH, 2009). A figura 4, a seguir ilustra o algoritmo DBSCAN.

Figura 4 – Algoritmo DBSCAN



Fonte: elaborada pelo autor.

Uma forma de determinar valores apropriados para o *eps* e *minPts* do DBSCAN, é através da análise das distâncias dos pontos até seus vizinhos. Essa distância sofrerá uma variação considerável quando houver uma distribuição aleatória dos pontos ou variação na densidade, caso contrário esse valor sofrerá pouca variação representando assim um valor

apropriado para o *eps*. Essa abordagem depende do valor de *minPts*. Ao atribuir valores crescentes para *minPts*, *clusters* pequenos tendem a ser rotulados como ruídos. Caso esse valor venha diminuindo, os ruídos serão clusterizados (TAN MICHAEL STEINBACH, 2009). A descoberta de valores para *eps* e *minPts* não é fácil, em muitos casos envolve a realização de vários testes com diferentes valores para ambos os atributos e apresenta uma forte dependência do conjunto de dados e seu contexto, bem como do objetivo da clusterização.

3 TRABALHOS RELACIONADOS

Esta seção apresenta os principais trabalhos relacionados ao contexto deste trabalho.

A cada ano desde de 2011 quando teve início os questionários sobre tecnologias, hábitos de programação, aprendizado e crescimento profissional no site StackOverflow[4], o número de participantes que responderam ao questionário vem aumentando. Em 2017 atingiu o valor de 64 mil desenvolvedores respondendo ao questionário e informando suas preferências tecnológicas, hábitos e evolução profissional. StackOverflow (2017), apresenta apenas os dados de desenvolvedores de acordo com região, gênero, idade, tecnologia, salário, experiência, entre outros, referentes ao ano de 2017.

Este trabalho se assemelha ao de StackOverflow, (2017), por analisar a mesma base de dados, as respostas dos questionários, identificando os perfis dos usuários. Mas difere quanto a análise. StackOverflow (2017) não realiza uma análise profunda dos dados. Limitando-se a expor os resultados de forma quantitativa. Por essa razão seu trabalho está melhor inserido na área de recuperação da informação, uma vez que apenas as informações óbvias dos dados são exibidas. Por outro lado, neste trabalho é realizada a mineração nessa base de dados. Dessa forma extraindo informações que não são visíveis sem uma análise mais profunda nos dados. Além da descrição dos padrões identificados.

Em (MOVSHOVITZ-ATTIAS et al., 2013), o sistema de reputação de usuários do site StackOverflow é analisado com base na interação dos usuários nos primeiros meses de atividade no site desde seu cadastro. O foco se dirige para usuários *expert*, com grande nível de contribuição e domínio sobre determinado assunto e usuários que não são *expert*, e representam as fontes primárias de perguntas com pouca contribuição em relação às respostas de perguntas feitas por outros usuários.

MOVSHOVITZ-ATTIAS et al. (2013) identificaram que existe uma forte relação entre as atividades nos primeiros três meses de atividade dos usuários e seu futuro nível de reputação no site. Em sua análise eles mensuraram a interação dos usuários ao criar ou responder perguntas e sua interação com outros usuários. A análise mostra que usuários *expert* possuem um padrão de interação no site muito diferente de usuários que não são *experts*,

[4] https://insights.stackoverflow.com/survey/

desde os primeiros meses, e que é possível prever, com base nos primeiros meses de uso do site, quais usuários se tornarão ou não *experts*.

Este trabalho se assemelha ao de MOVSHOVITZ-ATTIAS et al. (2013), uma vez que faz uma análise nos dados de usuários do site StackOverflow a fim de encontrar padrões e relações nos dados utilizando mineração de dados que identifiquem perfis de usuários. Este trabalho difere, no entanto, porque analisa dados referentes aos perfis profissionais dos usuários e não a interação entre os usuários. A base de dados utilizada neste trabalho também é diferente. Este trabalho adotado o *survey* de 2016. Esse *survey* contém dados referentes ao perfil pessoal do usuário, enquanto que a base de dados utilizada por MOVSHOVITZ-ATTIAS et al. (2013) contém as respostas e perguntas feitas pelos usuários, que é produto da interação entre eles.

Em seu trabalho, DE MORAES (2010) cria uma ferramenta utilizando a linguagem PHP[5] para realizar a extração dos dados de currículos dos professores da Escola Politécnica de Pernambuco da plataforma Lattes[6]. Utilizando essa ferramenta ele realiza a extração dos dados e a criação de uma base de dados para análise. Seguindo os passos do processo de KDD, DE MORAES (2010) faz a mineração dos dados na base criada.

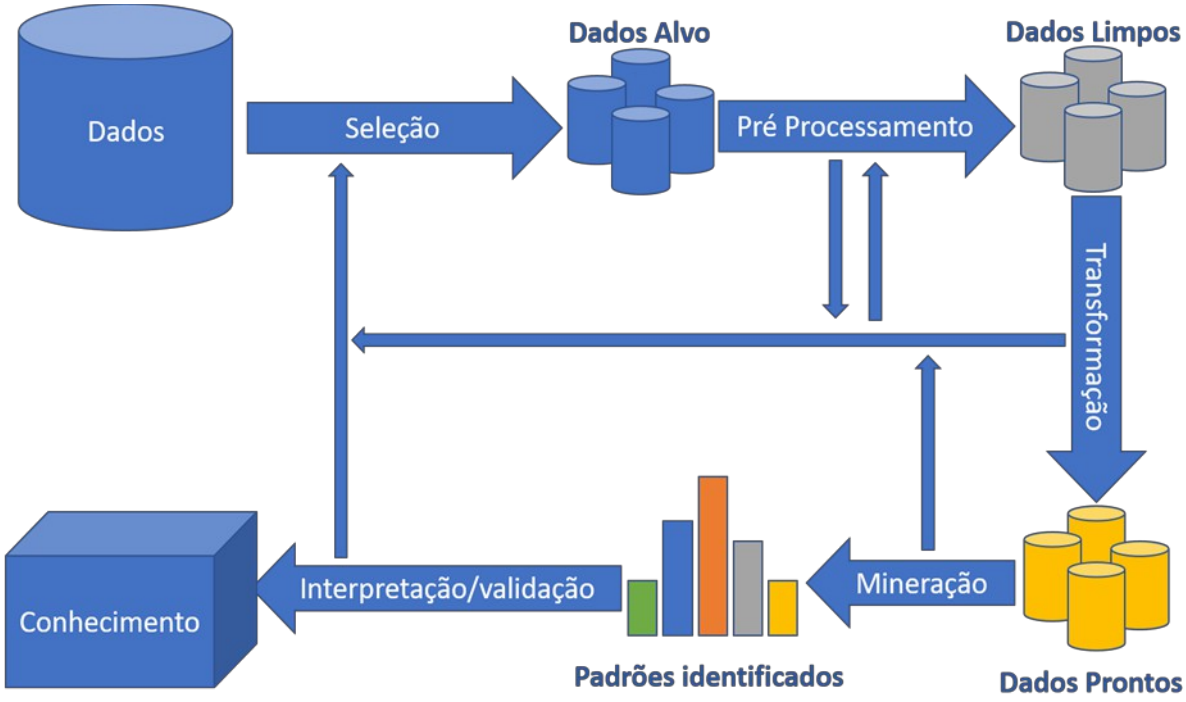
Semelhantemente a DE MORAES (2010), este trabalho segue os passos de KDD para a realização da mineração de dados em busca de perfis profissionais. Difere, no entanto, da fonte e base de dados utilizada, e ainda no perfil analisado. DE MORAES (2010), busca identificar perfis de professores, enquanto este trabalho analisa perfis profissionais com foco em tecnologia da informação. Outra diferença é que este trabalho não desenvolve nenhuma ferramenta auxiliar para extração dos dados, para tal fim foi utilizado o RapidMiner[7], enquanto que DE MORAES (2010), utiliza o WEKA[8] e a ferramenta criada para a análise.

[5] https://secure.php.net/
[6] http://lattes.cnpq.br/
[7] https://rapidminer.com/
[8] https://www.cs.waikato.ac.nz/ml/

4 METODOLOGIA

Este capitulo apresenta os procedimentos metodológicos adotados neste trabalho. A ferramenta RapidMiner foi utilizada para a execução do processo de ETL, e ainda, para a geração dos *clusters* e exibição dos gráficos e análises. A Figura 5 ilustra o processo de KDD adotado neste trabalho, e a seguir são descritos os passos de cada fase do KDD de forma mais detalhada.

Figura 5 – Fluxo do processo de Descoberta de Conhecimento em Banco de Dados



Fonte: elaborada pelo autor.

4.1 Definição do tipo de conhecimento a descobrir

Nesta fase foi definido o contexto do projeto. Foi proposto um estudo para identificação de perfis dos usuários do site StackOverflow, com foco em: 1) gênero; 2) idade; 3) salário; 4) experiência; 5) ocupação; 6) satisfação com emprego, para gerar informações para o público alvo deste trabalho.

4.2 Seleção dos dados

Nesta fase foram selecionadas os conjuntos de dados contendo as informações dos perfis de usuários para a análise neste trabalho. Essa tarefa foi realizada tomando como critério a existência, nos dados selecionados, de dados que representam os atributos em foco listados na subseção 4.1.

4.3 Pré-processamento dos dados

Nesta fase os datasets selecionados passaram por um pré-processamento para limpeza dos dados. Três critérios foram estabelecidos:

a) Registros com valores faltando; foi considerado registro com valor faltando o registro que possuía em um de seus atributos dado com valor em branco ou nulo. Todos os registros com valor faltando foram removidos.

b) Registros com ruídos; registros que continham dados com erros, como idade negativa ou maior que 150 mostram claramente um erro na coleta desse dado e, portanto, os registros contendo esse tipo de ruído foram removidos.

c) Registros com valores não representativos; foram considerados registros desse tipo, àqueles cujos atributos continham como valor de resposta algum dos seguintes valores 1) "*Other*" 2) "*Prefer not to disclose*", 3) "*Rather not say*", 4) "*Unemployed*", 5) "*I don't have a job*". O primeiro refere-se a um valor que não pode ser conhecido. O segundo e terceiro representam semanticamente a ausência do valor. Os dois últimos, identificam um indivíduo que não possui uma ocupação e, portanto, dificultando relacionar a salário e satisfação com o emprego.

4.4 Transformação dos dados

Nesta fase os dados foram transformados de textos para numéricos. Esse passo é necessário para melhor comparação das similaridades pelo algoritmo de clusterização utilizado com o RapidMiner. O capitulo 5 ilustra em detalhes essa transformação e seus resultados.

4.5 Mineração dos dados

Nesta fase é realizada a mineração dos dados. A técnica aplicada para realizar essa tarefa foi a clusterização utilizando o algoritmo DBSCAN, que permite agrupar dados a partir de similaridades, criando então dessa forma, os possíveis perfis dos usuários.

4.5.1 Validação e interpretação dos padrões minerados

Nesta fase os resultados são interpretados e validados. A fim de caracterizar uma melhor definição dos perfis, a clusterização foi realizada múltiplas vezes, com diferentes valores de *eps* e *minPts*.

O RapidMiner foi utilizado para construir os gráficos utilizados para interpretar os resultados, descrever os perfis e os próprios *clusters*. Em seguida, os *clusters* são analisados a fim de identificar os padrões que respondam às questões levantadas neste trabalho.

5 RESULTADOS

Esta Seção apresenta os resultados da mineração de dados realizada nos *surveys* do site StackOverflow selecionados neste trabalho.

5.1 Definição do tipo de conhecimento a descobrir

O contexto deste trabalho foi definido como uma tarefa de mineração de dados nos *surveys* do site StackOverflow, a fim de identificar perfis profissionais nas respostas dadas pelos usuários do site.

A mineração focou a identificação de perfis profissionais que contenham como atributos informações de: 1) gênero; 2) idade; 3) salário; 4) experiência; 5) ocupação; 6) satisfação com emprego.

Algumas questões foram definidas a fim de tangenciar a análise dos perfis. A seguir, são listadas as questões que este trabalho busca responder a partir da análise realizada.

- quais os perfis de satisfação no mercado de TI?
- quais os perfis de experiência no mercado de TI?
- quais os perfis de salário no mercado de TI?

5.2 Seleção dos dados

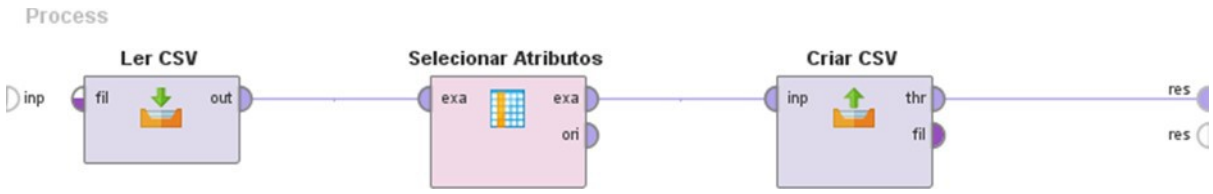
Nesta fase, foram selecionados os conjuntos de dados e os atributos que foram analisados neste trabalho. Para identificar os conjuntos de dados, uma pesquisa no site StackOverflow foi realizada buscando fontes de dados de perfis dos usuários. Como resultado da pesquisa foi identificado que a partir de 2011, a cada ano, os usuários do site têm respondido a pesquisas com questões referentes a tecnologias e informações pessoais propostas pelo site. O resultado das pesquisas feitas pelo site levou a criação de *surveys* a partir das respostas dadas pelos usuários. Estes *surveys* foram escolhidos como a base de dados inicial. A lista de *surveys* e os próprios *surveys* estão disponíveis ao público para download no próprio site do StackOverflow[9].

[9] https://insights.stackoverflow.com/survey/

O critério para seleção de quais *surveys* seriam analisados, foi a identificação de respostas que representassem os seis atributos em foco: 1) gênero; 2) idade; 3) salário; 4) experiência; 5) ocupação; 6) satisfação com emprego. Ou seja, o *survey* que não apresentou dados que permitissem identificar um ou mais atributos em foco, não foi incluído na análise. Sendo candidatos a análise apenas os *surveys* que continha todos os seis atributos.

A verificação dos atributos foi realizada através da leitura dos *surveys*. A Figura 6 ilustra o processo de seleção dos atributos para análise. A Tabela 1 resume essa verificação e o resultado final pode ser visto na Tabela 2. Como resultado dessa verificação, dois *surveys*, 2015 e 2016, atenderam aos critérios estabelecidos, sendo escolhido o mais recente entre eles. A partir do survey selecionado, os seis atributos foram projetados em um novo conjunto de dados, sendo essa a base de dados alvo da análise neste trabalho.

Figura 6 – Processo de seleção dos atributos



Fonte: elaborada pelo autor.

Tabela 1 – Identificação dos atributos selecionados nos surveys

Contém o atributo	**Ano do Survey**						
	2011	2012	2013	2014	2015	2016	2017
Gênero	Não	Não	Não	Sim	Sim	Sim	Sim
Idade	Sim	Sim	Sim	Sim	Sim	Sim	Não
Salário	Sim	Sim	Sim	Sim	Sim	Sim	Sim*
Experiência	Sim	Sim	Sim	Sim	Sim	Sim	Sim
Ocupação	Sim	Sim	Sim	Sim	Sim	Sim	Sim*
Satisfação com emprego	Sim	Sim	Sim	Não	Sim	Sim	Não

Fonte: elaborada pelo autor.

Tabela 2 – Seleção dos surveys

Survey	Critério
2011	Não selecionado por falta do atributo Gênero
2012	Não selecionado por falta do atributo Gênero
2013	Não selecionado por falta do atributo Gênero
2014	Não selecionado por falta do atributo Satisfação com emprego
2015	Atende aos critérios, mas não selecionado
2016	Selecionado por atender aos critérios
2017	Não selecionado por falta de 2 atributos, (Satisfação, Idade)

Fonte: elaborada pelo autor.

5.3 Pré-processamento dos dados

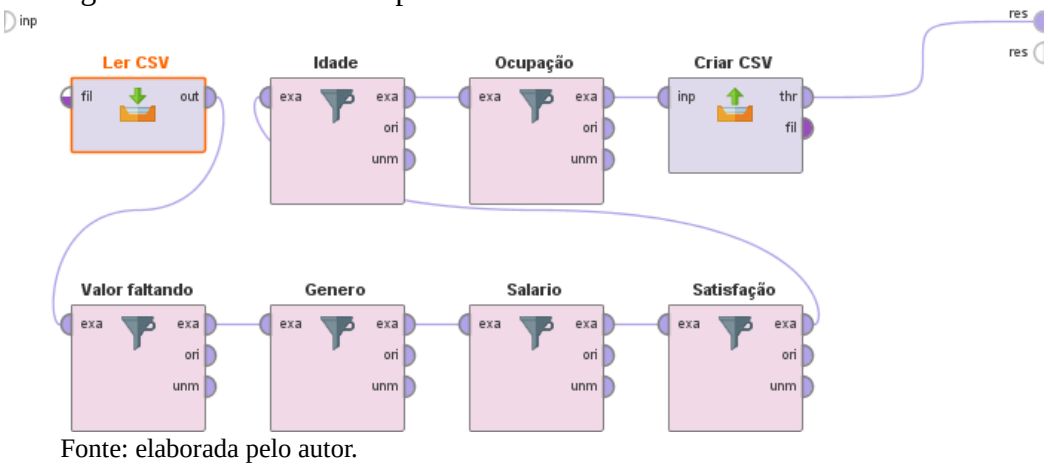
Nesta fase foi realizada a limpeza na base de dados alvo. Para realizar esse pré-processamento foi utilizada a ferramenta RapidMiner e aplicado os critérios de limpeza definidos na seção 4.3. A Tabela 3 apresenta o resultado dessa etapa, e a Figura 7 ilustra os passos realizados nessa tarefa. Após a limpeza foi gerado um novo conjunto de dados a partir do conjunto de dados contendo aproximadamente 58,5% dos registros da base de dados alvo. Uma redução de mais de 40% no volume dos dados, o que ressalta a importância da realização dessa etapa.

Tabela 3 – Resultado da limpeza dos dados

Etapa	#Registros
Antes da limpeza	56030
Após remover valores faltando	37512
Após remover ruídos/valores não representativos	32775

Fonte: elaborada pelo autor.

Figura 7 – Processo de limpeza dos dados



Fonte: elaborada pelo autor.

5.4 Transformação dos dados

Nesta fase, o conjunto de dados gerado na seção 5.3 contém apenas os seis atributos selecionados, além disso, todos os registros com valores faltando ou ruídos foram removidos. Porém alguns dados apresentavam valores não numéricos. Mesmo os valores de salário, idade e experiência por exemplo, estavam em um formato que não permitia a manipulação direta como número, sendo necessário uma transformação dos dados. Essa transformação foi realizada em duas etapas: 1) mapeamento dos valores originais (texto) para numéricos; 2) normalização dos dados para valores entre 0 (zero) e 1 (um) distribuídos em uma faixa de valores dentro desse intervalo como base na distância entre os valores.

O atributo ocupação conta com vinte e seis (26) valores distintos de respostas, esses valores foram discretizados em uma faixa de intervalos fixos entre si com valores de 0.0 a 1.0. A distância, ou seja, o intervalo entre os valores foi de 0.040. Ou seja, o primeiro valor é 0.0, o segundo 0.040, o terceiro 0.080, e assim sucessivamente até o último valor, 1.0. As Tabelas 4 e 5 mostram o resultado dessa tarefa em relação a cada atributo. Por uma questão de legibilidade o atributo ocupação foi colocado em uma tabela separada.

Tabela 4 – Resultado da discretização dos atributos gênero, idade, satisfação com emprego, experiência e salário

Atributo	Valor original	Transformação	Valor original	Transformação
Gênero	Masculino	0.0	Feminino	1.0
Idade	< 20 anos	0.1	20-29 anos	0.2
	30-39 anos	0.3	40-49 anos	0.4
	50-59 anos	0.5	> 60 anos	0.6
Satisfação com emprego	Baixa	0.0	Média	0.5
	Alta	1.0		
Experiência	<= 2 anos	0.0	2-5 anos	0.5
	> 6 anos	1.0		
Salário	<= 50000	0.2	<= 100000	0.4
	<= 150000	0.6	<= 200000	0.8
	> 200000	1.0		

Fonte: elaborada pelo autor.

Tabela 5 – Resultado da discretização do atributo ocupação

Valor original	Transformação	Valor original	Transformação
Mobile developer - iOS	0.000	Back-end web developer	0.040
Full-stack web developer	0.080	Desktop developer	0.120
Data scientist	0.160	Student	0.200
Engineering manager	0.240	Enterprise level services developer	0.280
Product manager	0.320	Designer	0.360
DevOps	0.400	Embedded application developer	0.440
Front-end web developer	0.480	Executive (VP of Eng., CTO, CIO, etc.)	0.520
Analyst	0.560	Mobile developer	0.600
Growth hacker	0.640	Developer with a statistics or mathematics background	0.680
System administrator	0.720	Database administrator	0.760
Graphics programmer	0.800	Business intelligence or data warehousing expert	0.840
Quality Assurance	0.880	Mobile developer - Android	0.920
Machine learning developer	0.960	Mobile developer - Windows Phone	1.000

Fonte: elaborada pelo autor.

5.5 Mineração dos dados

Nesta fase a ferramenta RapidMiner é utilizada para clusterizar os dados utilizando o algoritmo DBSCAN com medida de distância Euclidiana. A fim de caracterizar

uma melhor definição dos perfis, a clusterização foi realizada múltiplas vezes com diferentes valores de *eps* e *minPts*.

5.6 Validação dos *clusters* gerados

Nesta fase o RapidMiner foi utilizado para construir os gráficos utilizados na análise e validar os *clusters* gerados. Na seção 5.7 os *clusters* são analisados a fim de identificar os padrões que respondam às questões levantadas neste trabalho.

Após testes com diferentes valores de *eps* e *minPts*, os valores que apresentaram mudanças significativas nos *clusters* gerados foram 0.1, 0.2 e 0.3 para *eps* e os valores 50, 100 e 150 para *minPts*. Neste trabalho esses foram os valores finais adotados na clusterização.

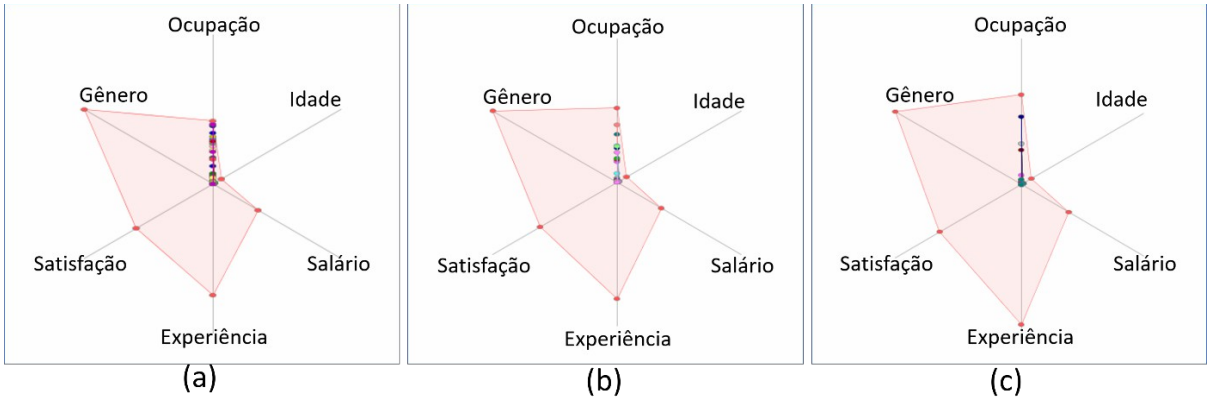
Para validar os *clusters* e os valores de *eps* e *minPts* dos *clusters* gerados foram construídos gráficos de teias a partir da clusterização com as variações de *eps* e *minPts* definidas. O gráfico de teia foi construído a partir de um grafo com os seis atributos selecionados como arestas a partir do centro do grafo. Cada *cluster* é representado por um ponto de uma cor e, registra no gráfico seu valor de variação para cada atributo. Quanto mais distante do centro um ponto de uma cor estiver, maior a variação, ou seja, maior dissimilaridade para os valores daquele atributo dentro do *cluster.* Quanto mais próximo ao centro o ponto estiver maior à similaridade do atributo no *cluster.* Este é, portanto, o critério de validação adotado neste trabalho. Os valores de *eps* e *minPts* que apresentarem menor dissimilaridade no gráfico de teia serão valores ideais para a clusterização. A clusterização que apresentar maior similaridade em seus atributos será considerada uma boa clusterização a ser analisada.

O cluster destacado em vermelho em todas as imagens agrupa todos os *outliers*. Esse *cluster* aparece com a maior variação entre seus atributos e, por não apresentar informação relevante para este trabalho não é considerado durante a análise, na identificação dos padrões. A seguir, é apresentado o resultado da similaridade nos *clusters* gerados a partir da variação de *eps* e *minPts* de acordo com os valores acima citados.

A clusterização com *eps* igual a 0.1 e *minPts* igual a 50, apresentou maior quantidade de *clusters* e com maior similaridade em relação às variações de *minPts* para esse mesmo *eps*. Ao elevar o valor de *minPts* observou-se que número de *clusters* diminuía e eles passavam a se aproximar mais do *cluster* contendo os *outliers*, ou seja, passavam a apresentar

atributos com maior dissimilaridade. A Figura 8 resume o resultado em gráfico de teia para *eps* igual a 0.1, com variação de *minPts*.

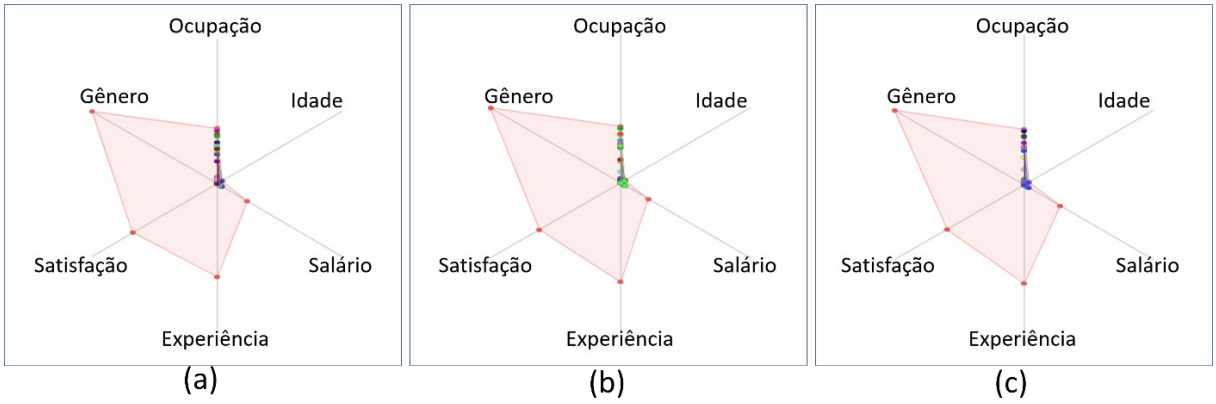
Figura 8 – Clusterização com eps fixo em 0.1 e minPts 50 em (a), 100 em (b) e 150 em (c)



Fonte: elaborada pelo autor.

A clusterização com *eps* igual a 0.2 mostra uma variação no número de *clusters* gerados significativa quando utilizando *minPts* 50 em relação aos outros dois valores de *minPts*. No entanto não apresentou variação significativa de similaridade. Os *clusters* gerados mostraram ainda dissimilaridades muito próximas ao *cluster* que contém os *outliers*. Indicando que os *clusters* gerados possuem alta dissimilaridade. O resultado pode ser visto na Figura 9 que exibe o resultado da variação de *minPts* iguais a 50, 100 e 150 em (a), (b) e (c) respectivamente com *eps* fixo em 0.2.

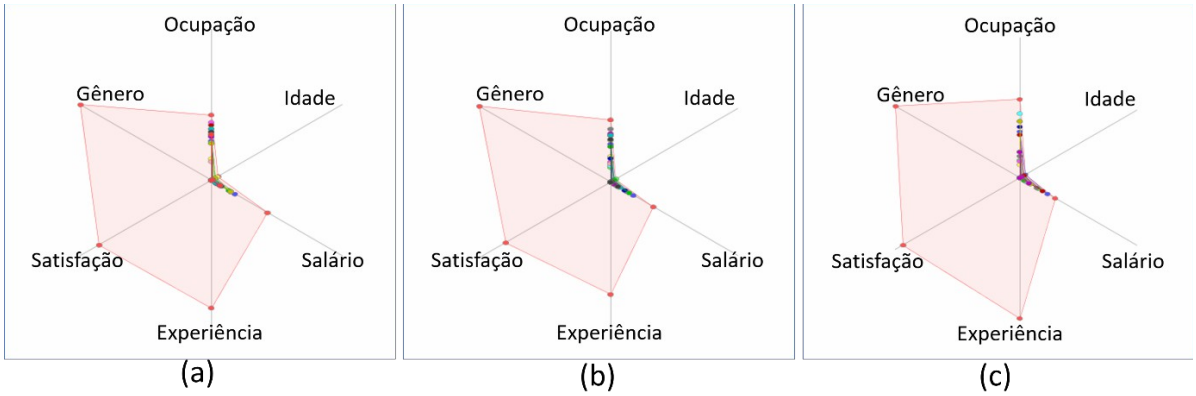
Figura 9 – Clusterização com eps fixo em 0.2 e minPts 50 em (a), 100 em (b) e 150 em (c)



Fonte: elaborada pelo autor.

A clusterização com *eps* igual a 0.3 apresentou *clusters* com maior similaridade utilizando *minPts* igual a 100. Todavia, tal qual a clusterização com *eps* igual a 0.2, os *clusters* gerados apresentam dissimilaridades muito próximas do *cluster* que contém os *outliers*. A Figura 10 ilustra o resultado da variação de *minPts* com *eps* fixo em 0.3 e variação de *minPts* em 50, 100 e 150 em (a), (b) e (c) respectivamente.

Figura 10 – Clusterização com eps fixo em 0.3 e minPts 50 em (a), 100 em (b) e 150 em (c)



Fonte: elaborada pelo autor.

Analisando os resultados da variação de *eps* e *minPts* na clusterização, observou-se que a clusterização com *eps* igual a 0.1 e *minPts* igual a 50 apresentou *clusters* com maior similaridade em relação às demais variações de *eps* e *minPts*. Além disso, apresenta variação apenas em um de seus atributos, ocupação, o que simplifica a explicação dos padrões. Por essa razão a clusterização com *eps* igual e *minPts* igual a 0.1 e 50 respectivamente, foi selecionada para análise de perfis realizada neste trabalho.

O *cluster* em vermelho, que contém os *outliers*, conforme já mencionado anteriormente, não apresenta informações relevantes para este trabalho e por isso será desconsiderado na apresentação dos padrões identificados.

A Tabela 6 resume os *clusters* gerados a partir da variação de *eps* e *minPts* utilizando os valores adotados neste trabalho. Os números de *clusters* gerados inclui o *cluster* com os *outliers*.

Tabela 6 – Resultado da clusterização com variação de eps e minPts

eps	minPts	Clusters Gerados
0.1	50	47
	100	33
	150	28
0.2	50	26
	100	18
	150	17
0.3	50	16
	100	15
	150	13

Fonte: elaborada pelo autor.

5.7 Interpretação dos padrões minerados

Os *clusters* gerados a partir dos valores de *eps* 0.1 e *minPts* 50, foram analisados a fim de identificar os padrões que respondam às questões levantadas neste trabalho. O resultado da análise é apresentado a seguir e cada questão será abordada separadamente na descrição dos padrões identificados.

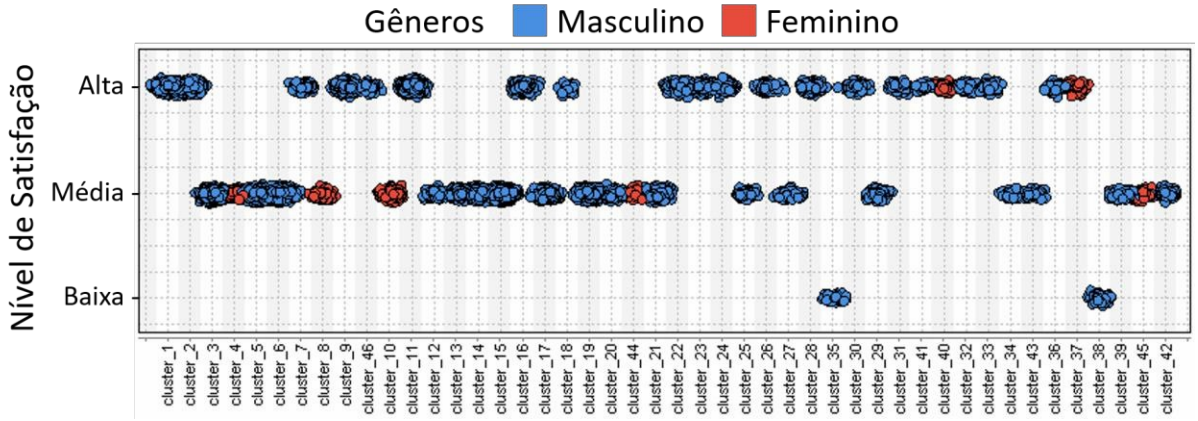
5.7.1 Quais os perfis de satisfação no mercado de TI?

A Figura 11 mostra que a satisfação na área de TI, de forma geral, é bem elevada. Dois *clusters*, 35 e 38, se destacam porque apresentam insatisfação. Esses dois *clusters* são compostos de seis áreas de ocupação; 1) *Mobile developer – iOS*, 2) *Desktop developer*, 3) *Data scientist*, 4) *Student*, 5) *Back-end web develope*r e 6) *Full-stack web developer*, sendo que, as duas últimas áreas juntas apresentam frequência de mais de 90% nos dois *clusters*. Essas seis áreas mencionadas foram as únicas que registraram insatisfação como pode ser visto na Figura 13. Entretanto, as mesmas áreas apresentam elevado nível de satisfação em outros *cluster*s.

Este nível de insatisfação identificado nos dois *clusters* em questão, 35 e 38, pode ser entendido como o resultado de dois outros fatores. O primeiro fator é a faixa salarial presente nos dois *clusters* que é a menor possível, 0.2, que corresponde à salários de até 50.000 dólares por ano. Além deste fator os *clusters* não apresentam registros de experiência baixa, apresentando níveis acima da média em ambos. Temos então, dois *clusters* que são

formados por indivíduos com alta experiência, e baixos salários. Estes dois fatores juntos, portanto, podem explicar o nível de insatisfação apresentado. Uma vez que em outras condições essas mesmas áreas de atuação presentes nos dois *clusters* apresentam elevados níveis de satisfação, a combinação desses dois fatores mencionados e não as áreas de atuação em si, podem explicar melhor os dados.

Figura 11 – Satisfação por gênero nos clusters

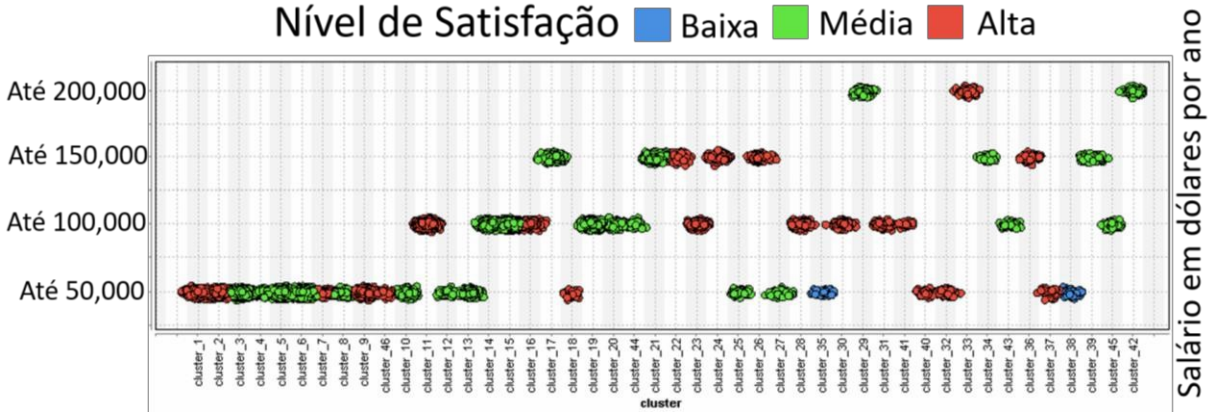


Fonte: elaborada pelo autor.

Em termos de gênero, não há grandes diferenças. De forma geral o nível de satisfação de homens e mulheres é alto em todas as áreas de ocupação.

A satisfação em relação à salário apresentou nível baixo apenas para salários de até cinquenta mil (50,000) dólares por ano. Faixas de salário acima desse patamar mostram que a satisfação tende a ficar entre média e alta como mostra a Figura 12.

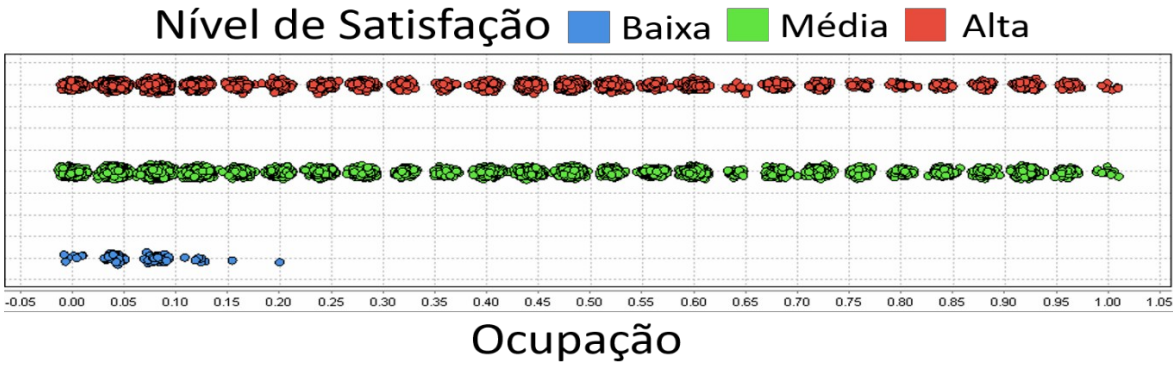
Figura 12 – Satisfação por salários nos clusters



Fonte: elaborada pelo autor.

A satisfação em relação à ocupação apresentou nível baixo em seis áreas de ocupação; 1) *Mobile developer – iOS*, 2) *Desktop developer*, 3) *Data scientist*, 4) *Student*, 5) *Back-end web developer* e 6) *Full-stack web developer*. Essas mesmas áreas também apresentam níveis médio e alto de satisfação. A Figura 13 ilustra essa situação.

Figura 13 – Satisfação por ocupação nos clusters



Fonte: elaborada pelo autor.

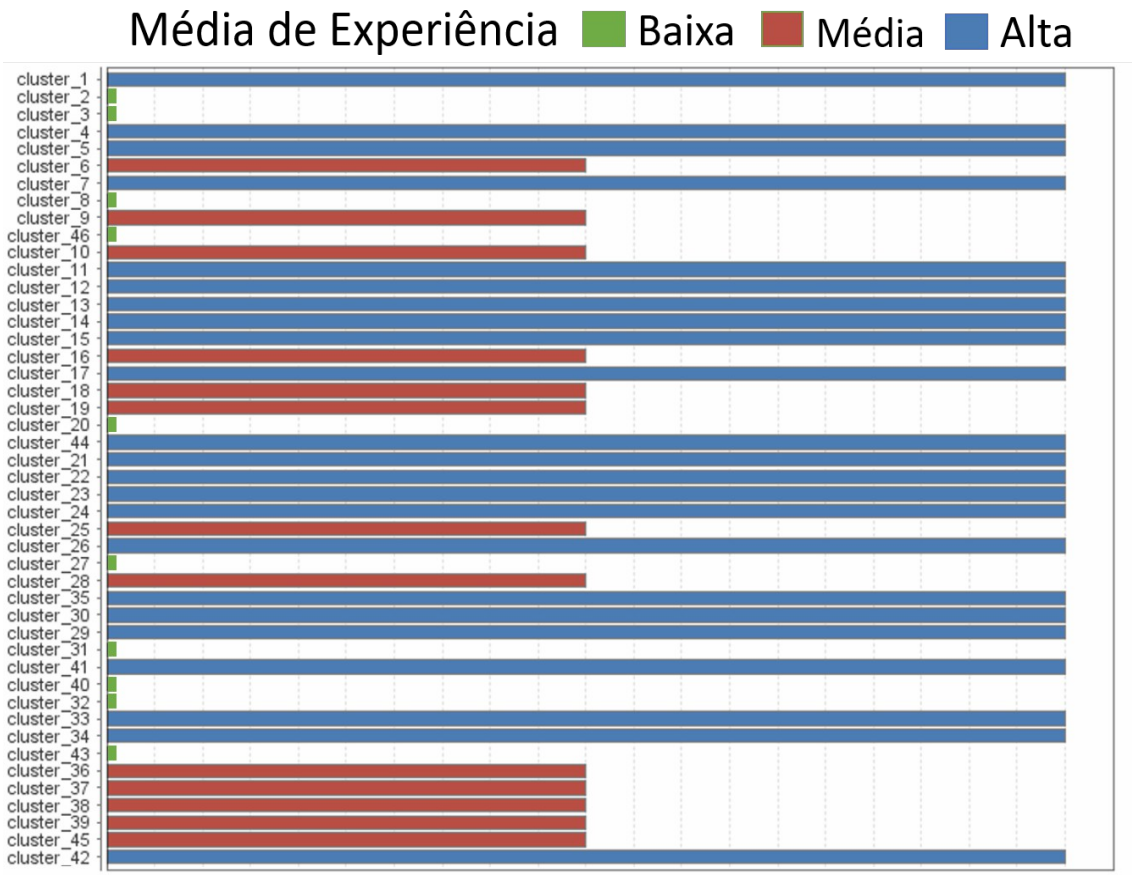
5.7.2 Quais os perfis de experiência no mercado de TI?

A Figura 14 ilustra a experiência média nos *clusters*, ela mostra que os *clusters* 2, 3, 8, 20, 27, 31, 32, 43 e 46, apresentaram uma média baixa no nível de experiência. Enquanto que os demais *clusters* apresentam nível médio a alto neste atributo. Dessa forma 37 dos 46 *clusters* analisados apresentaram um alto nível de experiência profissional.

A análise dos *clusters* 4, 8, 10, 37, 40, 44 e 45, em relação à experiência profissional por gênero ilustrada na Figura 15, mostra que não há uma diferença significativa de experiência entre os gêneros. Considerando que os dados mostram uma proporção de aproximadamente 16 homens para cada mulher, ainda assim, o nível de experiência é similar entre os dois gêneros.

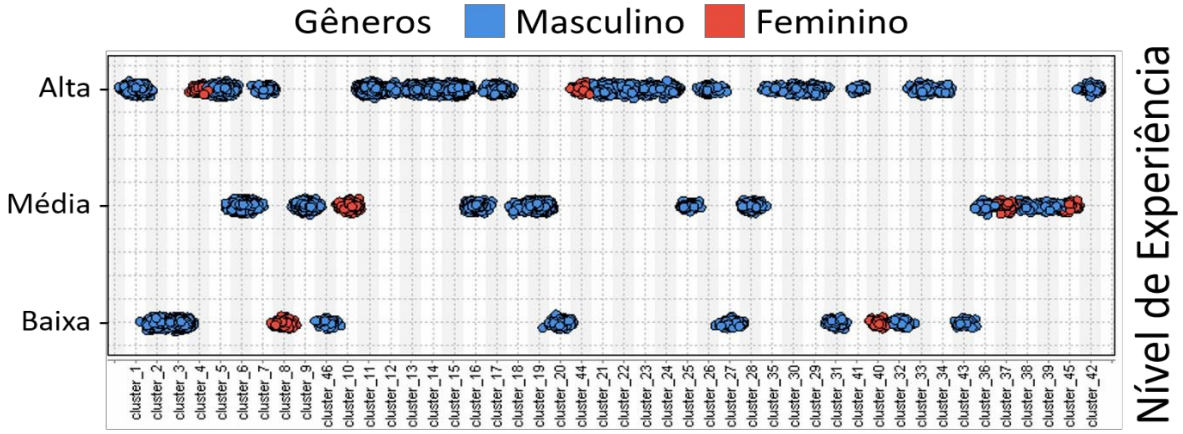
É possível ver na Figura 16 que conforme a experiência aumenta, os salários tendem a aumentar. No entanto, os dados mostram que experiência alta não implica necessariamente em salários altos exibindo salários de até 50,000 dólares mesmo em indivíduos com alta experiência. Os dados apontam ainda para um teto salarial para experiência baixa em 100,000 dólares por ano. Salários maiores que isso aparecem apenas em indivíduos com um maior nível de experiência.

Figura 14 – Experiência média nos clusters



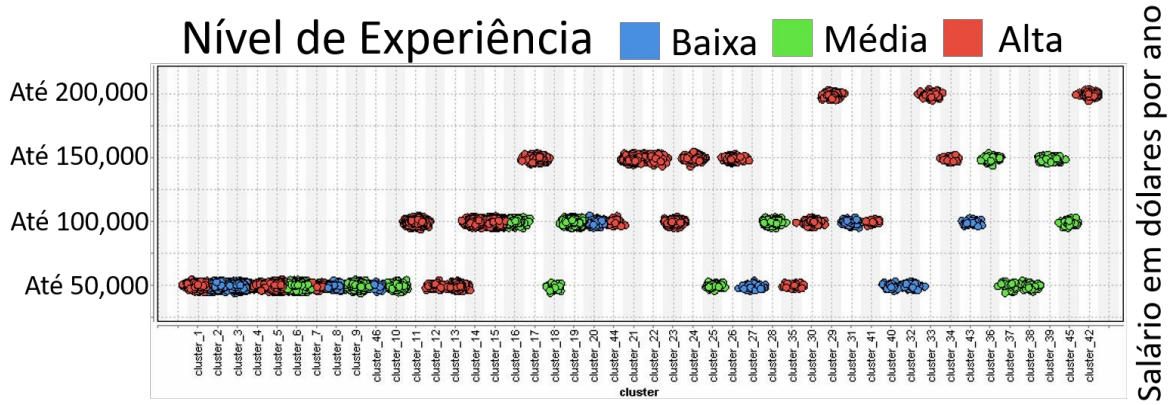
Fonte: elaborada pelo autor.

Figura 15 – Experiência por gênero nos clusters



Fonte: elaborada pelo autor.

Figura 16 – Experiência por salários nos clusters



Fonte: elaborada pelo autor.

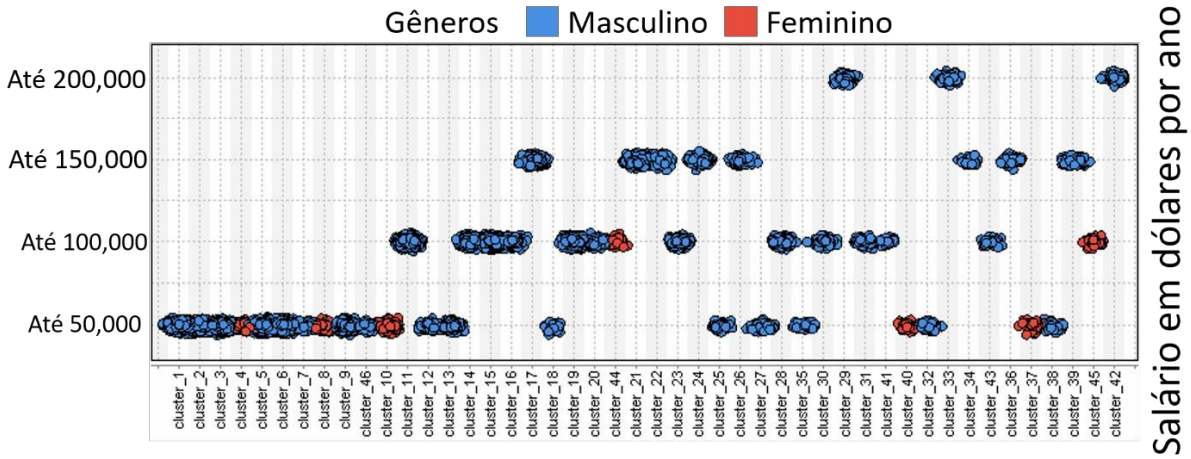
5.7.3 Quais os perfis de salário no mercado de TI?

A Figura 15 ilustra um padrão salarial para mulheres em vermelho e homens em azul. Os *clusters* 4, 8, 10, 37, 40, 44 e 45, de indivíduos do gênero feminino, mostram uma faixa entre 0.2 e 0.4, esses valores correspondem a salários de até 50.000 e até 100.000 dólares por ano respectivamente. É possível perceber que os salários do gênero feminino apresentam um teto bem inferior aos do gênero masculino, que possui um teto de 0.8, correspondendo a uma faixa salarial de até 200.000 dólares por ano.

Quando olhamos para os mesmos *clusters* em relação a experiência o padrão não se mantém. A Figura 15 da seção 5.7.2 mostra que não há a mesma diferença de experiência entre os gêneros que existe entre os salários. Pelo contrário, salvo as devidas proporções o nível de experiência é similar entre os dois gêneros. Essa igualdade, no entanto, não se reflete na remuneração, mostrando que há uma maior remuneração para o gênero masculino. É possível ver que o teto salarial do gênero feminino corresponde à metade do teto salarial masculino.

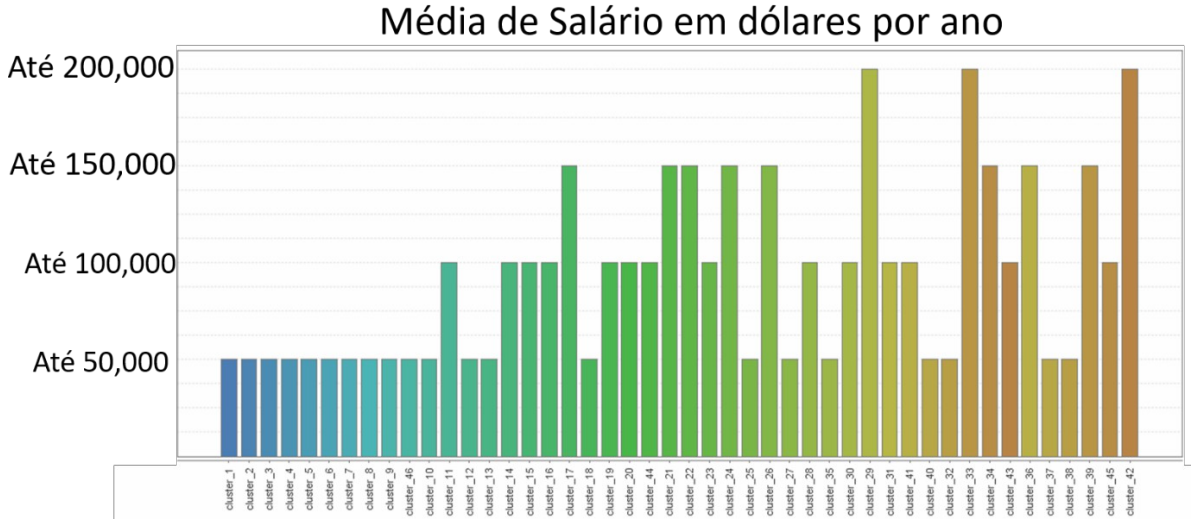
Ao analisar os *clusters* em função da média salarial 21 apresentaram média salarial em dólares por ano de até 50,000, 14 mostraram salários de até 100,000, 8 possuem salários de até 150,000 e apenas 3 *cluster* possuem média salarial de 200,000 dólares por ano. Temos então 35 *clusters* com salários de até 100,000 mil dólares, e 11 *clusters* com salários acima de 100,000. A Figura 18 destaca essas médias. É possível ver que a maioria dos *clusters* apresentam média salarial de até 100,000.

Figura 17 – Salários por gênero nos *clusters*



Fonte: elaborada pelo autor.

Figura 18 – Média salarial por clusters



Fonte: elaborada pelo autor.

6 CONSCLUSÃO E TRABALHOS FUTUROS

Este trabalho teve como objetivo utilizar a mineração de dados seguindo os passos do processo de descoberta de conhecimento em bancos de dados a fim de identificar padrões de perfis de usuários do site StackOverflow. A mineração foi realizada tomando como base de dados o *survey* dos usuários do ano de 2016 através da técnica de clusterização com DBSCAN. A clusterização foi realizada após uma fase de seleção de um conjunto de atributos como base para direcionamento da análise e permitir responder as perguntas levantadas nesse trabalho. A etapa seguinte foi o pré-processamento desses dados. Após essa fase de pré-processamento foi realizada a clusterização e validação dos dados para uma posterior análise e interpretação dos resultados.

O *survey* selecionado como base de dados apresentava uma grande quantidade de atributos e variações de respostas para esses atributos. Para um melhor direcionamento da análise foram selecionados um conjunto de atributos e, um conjunto de questões foram levantadas a fim de serem respondidas a partir dos resultados deste trabalho. Todavia os dados ainda não apresentavam uma estrutura apropriada par realizar a clusterização, sendo necessário realizar um pré-processamento dos dados. O pré-processamento foi uma etapa vital neste trabalho, em virtude da diversidade nos formatos dos dados. Nessa fase foram realizadas a remoção de registros com atributos faltando ou que não foi possível inferir um valor adequando ao processo de clusterização sem influenciar os resultados.

A clusterização foi realizada utilizando a ferramenta RapidMiner com o algoritmo DBSCAN e a medida de distância Euclidiana. Diversos valores de *eps* e *minPts* foram testados a fim de encontrar o mais adequado a clusterização. Ao final foi adotado o valor 0.1 e 50 para *eps* e *minPts* respectivamente.

Ao final do trabalho foi possível identificar alguns perfis de usuários a partir da clusterização realizada. Esses resultados podem ajudar na tomada de decisões das organizações que interagem com a equipe do site StackOverflow, bem como do próprio site na parte de marketing, divulgação de vagas de emprego, recrutamento, entre outras. Além de seus usuários que podem ter uma visão geral de como está o perfil na área de sua atuação e direcionar melhor seus esforços para alcançar seus objetivos profissionais.

Como trabalhos futuros podem ser realizadas a mineração em todos os *surveys* disponibilizados pelo site StackOverflow e comparando como os perfis se comportaram ao longo do tempo dessa forma seria realizada uma análise temporal o que permitiria entender

melhor os padrões identificados. Ao mesmo tempo ampliar o número de atributos contemplados na análise permitiria uma análise mais abrangente. Este trabalho focou em atributos de perfis profissionais, mas outros atributos podem apresentar padrões de tecnologias, educação, áreas de pesquisa, etc. A utilização de outras técnicas de mineração como a análise de associação também permitiria identificar possíveis relações entre os dados, fortalecendo os resultados encontrados.

REFERÊNCIAS

CORNELIUS JUNIOR, Romeu. **Uso da mineração de dados na identificação de alunos com perfil de evasão do ensino superior**. [S.l: S.n], 2015.

DE MORAES, B. C. S. **Extração de conhecimento da Plataforma Lattes utilizando técnicas de Mineração de Dados**: estudo de caso POLI/UPE. Trabalho de Conclusao de Curso (Engenharia de Computação)–Universidade de Pernambuco, 2010.

DIGIAMPIETRI, L. et al. Minerando e caracterizando dados de currıculos lattes. In: **Brazilian Workshop on Social Network Analysis and Mining (BraSNAM)**. [S.l: S.n], 2012.

DUDIK, Joshua M. et al. A comparative analysis of DBSCAN, K-means, and quadratic variation algorithms for automatic identification of swallows from swallowing accelerometry signals. **Computers in biology and medicine**, v. 59, p. 10-18, 2015.

ELMASRI, Ramez; NAVATHE, Shamkant B.; DE OLIVEIRA MORAIS, Rinaldo. **Sistemas de banco de dados**. 6.ed. São Paulo: Pearson 2011. 788p.

FAYYAD, Usama M.; PIATETSKY-SHAPIRO, Gregory; SMYTH, Padhraic. **Advances in Knowledge Discovery and Data Mining**. Menlo Park, EUA: AAAI Press, 1996. 611 p.

HAN, Jiawei; PEI, Jian; KAMBER, Micheline. **Data mining:** concepts and techniques. [S.l]: Elsevier, 2011.

MOVSHOVITZ-ATTIAS, Dana et al. Analysis of the reputation system and user contributions on a question answering website: Stackoverflow. In: **Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining**. ACM, 2013. p. 886-893.

RAPIDMINER. **RapidMiner**, 2017. Disponível em: < https://rapidminer.com/>. Acesso em: 10 out. 2017.

RODRIGUES, Priscila Rocha Ferreira; COELHO DA SILVA, Ticiana L.. **Dinâmica de Temas Abordados no Twitter Via Evolução de Clusters**. 2016. 57 p. TCC (Graduação em Engenharia de Software) - **Universidade Federal do Ceará**, Quixadá 2016.

StackOverflow. Developer Survey Results 2017. **stackoverflow.com**, 2017. Disponível em: < https://insights.stackoverflow.com/survey/2017>. Acesso em: 10 out. 2017.

SILBERSCHATZ, Abraham; KORTH, Henry F.; SUDARSHAN, S. **Sistema de bancos de dados**. 6.ed. Rio de Janeiro: Campus, 2012. 861 p.

SILVA, Tércio Jorge da; COELHO DA SILVA, Ticiana L.. **Extração de conhecimento nos dados da Universidade Federal do Ceará via Mineração de Dados**: Descoberta e análise dos perfis dos alunos. 2014. 66 p. TCC (Graduação em Sistemas de Informação) - Universidade Federal do Ceará, Quixadá 2014.

SILVA, Marcelino P. Santos**. Mineração de Dados-Conceitos, Aplicações e Experimentos com Weka**. In: Artigo. Instituto Nacional de Pesquisas Espaciais (INEP). São José dos Campos-SP. 2004.

STANLEY, Clayton; BYRNE, Michael D. Predicting tags for stackoverflow posts. In: **Proceedings of ICCM**. [S.l: S.n], 2013.

TAN, Pang-Ning; STEINBACH, Michael; KUMAR, Vipin. **Introdução ao datamining:** mineração de dados. Rio de Janeiro: Ciência Moderna, 2009. 928 p.

TRONCHONI, Alex B. et al. Descoberta de conhecimento em base de dados de eventos de desligamentos de empresas de distribuição. **Revista Brasileira de Automática**, v. 21, 2010.