UNIVERSIDADE FEDERAL DO CEARÁ

CAMPUS DE RUSSAS

CURSO DE GRADUAÇÃO EM ENGENHARIA DE SOFTWARE

CINTIA LIMA MAIA

PREVENDO A FORÇA DE CONEXÃO POR MEIO DA REDE SOCIAL ONLINE FACEBOOK

RUSSAS

2020

CINTIA LIMA MAIA

PREVENDO A FORÇA DE CONEXÃO POR MEIO DA REDE SOCIAL ONLINE FACEBOOK

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Engenharia de Software do Campus de Russas da Universidade Federal do Ceará, como requisito parcial à obtenção do grau de bacharel em Engenharia de Software.

Orientadora: Prof. Ms. Tatiane Fernandes Figueiredo

RUSSAS

2020

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)

M185p Maia, Cintia Lima.
Prevendo a força de conexão por meio da rede social online Facebook / Cintia Lima Maia. – 2020.
55 f. : il. color.

Trabalho de Conclusão de Curso (graduação) – Universidade Federal do Ceará, Campus de Russas, Curso de Engenharia de Software, Russas, 2020.
Orientação: Prof. Me. Tatiane Fernandes Figueiredo.

1. Redes sociais. 2. Força de conexão. 3. Algoritmo de clusterização. I. Título.

CDD 005.1

CINTIA LIMA MAIA

PREVENDO A FORÇA DE CONEXÃO POR MEIO DA REDE SOCIAL ONLINE FACEBOOK

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Engenharia de Software do Campus de Russas da Universidade Federal do Ceará, como requisito parcial à obtenção do grau de bacharel em Engenharia de Software.

Aprovada em:

BANCA EXAMINADORA

Prof. Ms. Tatiane Fernandes Figueiredo (Orientadora)
Universidade Federal do Ceará (UFC)

Prof. Ms. Eurinardo Rodrigues Costa
Universidade Federal do Ceará (UFC)

Prof. Dr. Pablo Luiz Braga Soares
Universidade Federal do Ceará (UFC)

AGRADECIMENTOS

Deixo aqui meus sinceros agradecimentos a todos que me ajudaram direta e indiretamente na construção deste trabalho. Cito inicialmente a intervenção de Deus, na pessoa de Jesus Cristo, que proporcionou forças e discernimento em toda minha graduação e em especial no desenvolvimento deste trabalho. Agradeço a Santíssima Virgem Maria que me tranquilizou nos momentos de preocupação e inseguranças no decorrer da pesquisa.

Agradeço a toda minha família, em especial minha mãe Francisca das Chagas e minha tia-madrinha Ana Francisca, que investiram em mim toda confiança, apoio emocional e financeiro para que eu pudesse estudar e concluir com êxito a minha pesquisa.

Agradeço a UFC pelo investimento nos meus estudos e por me proporcionar conhecimento para construção desta pesquisa e consequentemente minha graduação. Em especial agradeço a todos integrantes do grupo de pesquisa NEMO, principalmente aos professores pelas experiências e conhecimentos transmitidos aos alunos.

Agradeço também a orientação da professora Tatiane Fernandes que esteve comigo desde do início da pesquisa, me incentivando e auxiliando nas minhas decisões. Com certeza ela tem uma papel especial na construção deste trabalho, pois com seu jeito motivado e seu conhecimento ela me conduziu no desenvolvimento de cada capítulo.

Devo citar a contribuição do meu digníssimo namorado Hygor Lima, que esteve comigo desde do início da pesquisa e sempre me motivou a produzir, mesmo nos dias de preguiça.

Por fim, agradeço a todos meus amigos em especial a Érika Castro, Julyana Rodrigues, Larissa Stefanne, Leilia Ribeiro, Lenita Ribeiro e Lilia Lima, por me proporcionar momentos de descontração e diversão (importantes para conservar minha sanidade mental), além de compreenderem minha ausência em certos momentos para dedicação aos estudos.

RESUMO

Nos últimos anos a análise das redes sociais ganhou uma atenção significativa. Parte desse sucesso deve-se a popularização das Redes Sociais *Online* e ao grande volume de dados que essas redes têm gerado. O trabalho apresenta um algoritmo para examinar a Força de Conexão entre usuários da Rede Social *Online Facebook*. Além do algoritmo a autora desenvolveu uma Métrica de Similaridade para classificar essa Força de Conexão. A similaridade entre um par de usuários é calculada com base em suas reações a publicações do *Facebook*. A Metodologia do trabalho foi baseada no modelo *Cross-Industry Standard Process of Data Mining* (CRISP-DM), um processo da Mineração de Dados. Inicialmente o trabalho apresenta as técnicas para tratamento da base de dados, em especial as publicações, em seguida o desenvolvimento dos algoritmos para medir a Força de Conexão e Métrica de Similaridade. Usando os resultados gerados por esse algoritmo, o trabalho analisa as conexões existente entre amigos da rede social e mostra que mais de 90% dos perfis estão conectados a outros perfis similares, indicando que as relações construídas no *Facebook* são influenciadas pelos interesses dos usuários. Por fim, o trabalho apresenta dois modelos de clusterização, o Algoritmo Hierárquico destinado a construir grupos com tamanhos variados e o Algoritmo Hierárquico modificado desenvolvido pela autora para construir grupos com a mesma quantidade de perfis. O algoritmo proposto demonstrou bons resultados para clusterizar muitos perfis em poucos grupos.

Palavras-chave: Redes Sociais. Força de Conexão. Algoritmo de Clusterização.

ABSTRACT

In recent years the analysis of social networks has gained significant attention. Part of this success is due to the popularization of Online Social Networks and the large volume of data that these networks have generated. The work presents an algorithm to examine the Tie Strength between users of the online social network Facebook. In addition to the algorithm, the author developed a similarity metric to classify this Tie Strength. The similarity between a pair of users is based on their reactions to Facebook posts. The work methodology was based on the CRISP-DM model, a Data Mining process. Initially the work presents the processing for handling the database, especially publications, then the development of algorithms to measure the Tie Strength and similarity metric.Using the results generated by this algorithm, the work analyzes the existing connections between friends of the social network and shows that more than 90% of the profiles are connected to other similar profiles, indicating that the relationships built on Facebook are influenced by the interests of users. Finally, the work presents two clustering models, the Hierarchical Algorithm designed to build groups of varying sizes and the modified Hierarchical Algorithm developed by the author to build groups with the same number of profiles. The proposed algorithm demonstrated good results for clustering many profiles in a few groups.

Keywords: Social networks. Tie Strength. Clustering Algorithm.

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ALGORITMOS

LISTA DE ABREVIATURAS E SIGLAS

CRISP-DM	*Cross-Industry Standard Process of Data Mining*
JSON	*JavaScript Object Notation*
KDT	*Knowledge Discovery from Text*
OSN	*Online Social Networks*
SSE	*Error Sum of Squares*
URL	*Uniform Resource Locator*

LISTA DE SÍMBOLOS

A	Matriz Amizade
C	Matriz Perfil-Categoria
D	Matriz Distância
m	Número de categorias
n	Número de perfis

SUMÁRIO

1 INTRODUÇÃO

Por uma questão natural e de sobrevivência, a humanidade necessita e deseja unir-se e agrupar-se, sendo de suma importância para seres humanos a formação de grupos para conquistar metas e objetivos. Essa formação é um processo complexo, onde os indivíduos com suas forças internas (crenças, opiniões, objetivos, destreza, medos) podem ou não se transfigurar em grupos ativos, decorrente de objetivos comuns ou semelhantes (XAVIER, 1990).

A tentativa de estudar grupos e seus relacionamentos tem sido o objetivo de cientistas há muito tempo, sendo a Sociometria uma das primeiras técnicas apresentadas na literatura para explorar a formação de grupos e a interação entre pessoas. Esta técnica, proposta por Moreno (1992), busca analisar, mapear e medir os possíveis vínculos construídos em Redes Sociais. Após a apresentação da Sociometria na literatura, o número de pesquisas que buscam estudar e analisar Redes Sociais tem crescido, principalmente nos últimos anos, devido a popularização das Redes Sociais *Online* (em inglês, *Online Social Networks* (OSN)). A possibilidade de coletar importantes informações por meios tecnológicos ocasionou uma mudança significativa no estudo das OSN, gerando o surgimento de uma nova ciência, denominada como Ciência Social Computacional (LAZER *et al.*, 2009).

A Interação Social é um conceito muito importante na área de Ciência Social Computacional. Em uma OSN, as interações são manifestadas através da troca de conteúdos midiáticos (fotos, textos, *links*, músicas, vídeos, entre outros), onde cada usuário pode se conectar a esses conteúdos por meio de ações como compartilhar, curtir e comentar (GOMES, 2013). Desta forma, os indivíduos e organizações são afetados diretamente por essas interações, podendo estes relacionamentos serem medidos através da força de suas conexões.

Segundo Xiang *et al.* (2010) existem dois tipos de Força de Conexão: Conexão Fraca e Conexão Forte. A relação entre colegas, conhecidos ou desconhecidos normalmente é classificada como uma conexão fraca. Nesse caso, indivíduos com esse tipo de conexão podem influenciar outros indivíduos a conhecerem lugares novos, gerar ideias e indicarem empregos. Já entre amigos e familiares normalmente a conexão é forte, nesse caso os indivíduos podem afetar emocionalmente o outro, e podem influenciar a tomar uma decisão difícil.

Este trabalho tem como foco examinar a Força de Conexão entre um grupo de usuários da Rede Social *Online Facebook*. Para tal, esta monografia apresenta o desenvolvimento de um algoritmo que analisa interações entre contas de usuários da rede social mencionada, inferindo a Força de Conexão entre os perfis analisados. A Força de Conexão é definida baseada

nos interesse de cada perfil, também pode ser chamada de Grau de Similaridade. Buscando alcançar o objetivo almejado, inicialmente, foi realizado um estudo literário, com o intuito de listar quais relações e interações poderiam influenciar a compatibilidade/afinidade dos usuários de uma rede social. Em seguida, foi criado um perfil de interesses para cada usuário estudado, utilizando como base as suas publicações e reações à publicações de terceiros. A partir da comparação entre esses interesses foi criado um algoritmo capaz de mensurar a similaridade entre usuários envolvidos.

A partir do algoritmo similaridade criado, o presente trabalho define uma nova métrica para classificação de Força de Conexão, podendo esta métrica ser utilizada para mensurar a afinidade entre qualquer grupo de usuários da rede social *Online Facebook*, assim como responder questionamentos relacionados a este grupo. Desta forma, esta monografia busca responder o seguinte questionamento: "Interesses semelhantes realmente influenciam a amizade entre pessoas na rede social *Online Facebook*?". Por fim, é apresentada uma técnica para clusterização de usuários, buscando maximizar suas afinidades.

A estrutura deste trabalho encontra-se da seguinte forma: no Capítulo 2 é apresentado o objetivo geral e os específicos; no Capítulo 3 são apresentados os tópicos chaves da pesquisa; no Capítulo 4 são apresentados trabalhos encontrados na literatura que são similares a este; no Capítulo 5 são apresentados os procedimentos utilizados para realizar a pesquisa e desenvolvimento da aplicação; no Capítulo 6 são mostrados os resultados obtidos com a pesquisa; e por fim, no Capítulo 7 tem a conclusão do trabalho e a descrição dos trabalhos futuros.

2 OBJETIVOS

2.1 Objetivo geral

Inferir a Força de Conexão entre usuários da Rede Social *Online Facebook*.

2.2 Objetivo específicos

- Listar as relações encontradas na literatura que possam influenciar a Força de Conexão entre pares de indivíduos pertencentes a uma rede social;
- Normalizar a base dados gerada através da Rede Social *Online Facebook*;
- Criar uma base de interesse para cada usuário baseado nas reações à publicações de outros usuários;
- Definir um algoritmo para inferir a Força de Conexão em número os entre perfis;
- Criar uma métrica para classificar a Força de Conexão entre perfis;
- Analisar os perfis da base de dados na tentativa de responder o seguinte questionamento: "Interesses semelhantes realmente influenciam a amizade entre usuários da Rede Social *Online Facebook*?";
- Apresentar um algoritmo de Clusterização que forme grupos maximizando suas afinidades.

3 FUNDAMENTAÇÃO TEÓRICA

Para uma compreensão clara do trabalho, neste capítulo serão apresentados os conceitos chaves da pesquisa. Na Seção 3.1, são expostas as definições relacionadas às Redes Sociais, assim como os termos fundamentais no campo de aplicação desta pesquisa. Na Seção 3.2, é introduzido o conceito de Força de Conexão, uma das definições mais importantes deste capítulo. Na Seção 3.3, será apresentado processo para Mineração de Texto. Por fim, na Seção 3.4 são tratados os principais algoritmos de Clusterização presente na literatura.

3.1 Rede Social

Segundo Barbosa *et al.* (2000), uma rede é um conjunto de nós conectados, em que os nós podem representar pessoas, grupos ou outras unidades, e suas conexões representam relacionamentos. Mais especificamente, uma Rede Social é o conjunto de indivíduos agrupados relacionados através de conexões.

Cientificamente, uma Rede Social pode ser considerada uma estrutura complexa, constituída por pessoas com ideias, valores e objetivos comuns ou divergentes, interligadas ou não. Como a Rede Social tratada neste trabalho faz referência a espaços virtuais, esta é definida como Rede Social *Online* (OSN), onde grupos de pessoas se relacionam, trocando mensagens e compartilhando conteúdo através da internet.

3.1.1 Rede Social Online

Redes Sociais *Online* (OSN) são sistemas onde cada pessoa, conhecida como usuário, têm um perfil público ou semipúblico. Estes usuários podem se ligar a outros através de "amizades", seguindo seus perfis e interagindo através do compartilhamento de conteúdos (MISLOVE *et al.*, 2007). Por meio das OSN, os usuários podem fazer novas amizades e intensificar as relações sociais existentes na vida real. Atualmente, existem diversos tipos de Redes Sociais *Online*, cada uma com objetivos e públicos diferentes, como os citados abaixo:

- *Facebook*: Interação entre contatos;
- *YouTube*: Compartilhar vídeos;
- *WhatsApp*: Envio de mensagens;
- *Instagram*: Compartilhar fotos e vídeos.

O presente trabalho é baseado especificamente na Rede Social *Online Facebook*.

O *Facebook* foi criado no ano de 2003 por *Mark Zuckerberg* na Universidade de *Harward*. Atualmente, é a Rede Social com mais usuários do mundo e está disponível em mais de 70 idiomas. O *Facebook* é uma Rede Social *Online* gratuita, e os usuários administram o seu espaço através de uma linha do tempo. Está linha permite o gerenciamento de várias funções como: adicionar pessoas a lista de amigos; *post* de informações em um mural (textos, vídeos, fotos); parabenizar amigos; comentar publicações; curtir páginas, participar de grupos; enviar mensagens privadas; organizar eventos; entre outros.

3.1.2 Análise de Redes Sociais

Redes Sociais *Online* são usadas para investigar as relações dos indivíduos pertencentes a um grupo. De acordo com Barbosa *et al.* (2000, p. 41) "A análise de redes sociais baseia-se no pressuposto da importância das relações entre unidades que interagem, isto é, relações definidas como ligações entre unidades constituídas, componente fundamental das teorias de redes. As regularidades ou padrões de interação dão origem às estruturas."

3.2 Força de Conexão

O termo *Strength Tie* (em tradução livre para o português, Força de Conexão) foi introduzido em 1973 por Granovetter (1977) em seu trabalho *"The Strength of Weak Ties"*. A Força de Conexão é uma combinação (provavelmente linear) da quantidade de tempo, a intensidade emocional, a intimidade, e os serviços recíprocos que caracterizam o relacionamento. Normalmente a Força de Conexão pode ser representada por um número ou categoria por exemplo, existem dois tipos de conexão:

1. Conexão forte: são pessoas de confiança, que dividem os mesmo grupos sociais, ou seja, que possuem amigos em comum;
2. Conexão fraca: são pessoas que nunca se viram ou apenas conhecidos, sem nenhuma ligação entre elas.

Granovetter (1977) apontou inicialmente quatro dimensões que medem a Força de Conexão: Quantidade de Tempo, Intimidade, Intensidade e Serviços Recíprocos. Porém, pesquisas posteriores aumentaram essa lista. Como exemplo, podemos citar Lin *et al.* (1981) que mostrou que a distância social (Raça, Gênero, Escolaridade, Religião entre outros) influencia na Força de Conexão. Enquanto os autores Wellman e Wortley (1990) mostraram que a abertura

para conselhos apresenta uma forte conexão entre pessoas, indicando que o apoio emocional fosse adicionado as dimensões. Por fim, a topologia da rede e os círculos sociais foram pontos defendidos por Burt (2009), que os resumiu como fatores estruturais, e também passaram a ser considerados.

Atualmente, a literatura (GILBERT; KARAHALIOS, 2009) sugere sete dimensões para medir a Força de Conexão: Intensidade, Intimidade, Duração, Serviços Recíprocas, Fatores Estruturais, Apoio Emocional e Distância Social. Como o objetivo deste trabalho é construir a personalidade de cada perfil baseado nas reações em cada publicação, será levado em consideração apenas quatro destas dimensões: Intimidade, Serviços Recíprocos, Fatores Estruturais e Distância Social. Cada uma destas dimensões será caracterizada e exemplificada utilizando os padrões da Rede Social *Online Facebook* a seguir:

- Intimidade: Palavras de intimidade trocadas no mural como: família, trabalho, lazer, amigos e similares;
- Serviços Recíprocos: Quantidade de *Uniform Resource Locator* (URL) e aplicativos trocados entre amigos através do mural;
- Fatores Estruturais: Demonstração de interesses pessoais por determinados hobbies, culturas e entretenimento. Representado pelos grupos em comuns, páginas compartilhadas entre outros;
- Distância Social: Diferença educacional, política, religiosa e social.

Neste trabalho o termo Força de Conexão pode ser apresentado como similaridade ou afinidade.

3.3 Mineração de Texto

As definições a seguir foram baseadas no trabalho de Moura (2004), onde se define que a Mineração de Texto está relacionada com a *Knowledge Discovery from Text* (KDT), tradução livre, Descoberta de Conhecimento a partir de Texto. A KDT é uma técnica de extração de informações a partir de escritos que podem ser textos, frases ou palavras. Seu principal objetivo é inferir conhecimento de conjuntos de escritos usando algoritmos computacionais. Esses algoritmos se fazem necessários devido ao grande volume de dados que devem ser analisados, sendo seu foco deixar o processo de descoberta de conhecimento mais eficaz.

O processo de Mineração de Texto é composto das seguinte etapas (não necessariamente todas obrigatórias): Seleção dos documentos; Definição de abordagem dos dados (podendo

esta ser análise semântica ou análise estatística); Preparação dos dados; Indexação e normalização; Cálculo da relevância dos termos; Seleção dos termos e por fim, Pós-processamento. A seguir é descrito cada uma das etapas usadas neste trabalho.

3.3.1 Seleção dos documentos

A seleção de documentos é o processo para buscar os dados que serão usados para inferir uma informação. A escolha desses documentos é uma decisão muito importante para o decorrer das etapas de processamento, pois uma base de dados definida de forma errônea pode gerar resultados sem muita qualidade de informações.

3.3.2 Definição de abordagem dos dados

De acordo com a literatura, a definição de abordagem dos dados pode ser feita por dois tipos de análises: a Semântica e a Estatística. A abordagem adotada neste trabalho é a Estatística, e para a representação de documentos é usado a abordagem *Bag of Words* (saco de palavras). A seguir é apresentado uma breve descrição sobre a análise escolhida.

A Análise Estatística, ao contrário da Análise Semântica (que analisa a sequência das palavras), é definida através da quantidade de vezes que a palavra se repete no texto. O processo de Análise Estatística é composto pelas seguintes etapas: Codificação dos dados, Estimativa dos dados e Modelos de representação de documentos. A abordagem utilizada para representação de documentos é chamada de *Bag of Words*, por que ela não leva em consideração a ordem que as palavras aparecem no texto e nem a pontuação, considerando unicamente a quantidade de vezes que a palavra se repete.

3.3.3 Indexação e normalização

Nessa fase, a base de dados passa por algumas modificações na tentativa de otimizar o futuro processo de identificação dos termos. Para tal, são normalizadas as variações de morfologia e problemas com palavras sinônimas. O processo de indexação e normalização é composto pelas seguintes fases:

1. **Identificação de Termos**: identifica os termos (simples e compostos) dentro dos textos. Para uma identificação eficaz algumas alterações podem ser feitas no texto, como: converter maiúsculo ou minúsculo, caracteres especiais podem ser removidos, dentre outros.

2. **Remoção de *Stopwords***: processo que remove do texto palavras irrelevantes como preposições, pronomes, artigos e etc. Essas palavras são classificadas como *Stopwords*, e normalmente não são usadas como termos para consultas.
3. **Normalização Morfológica**: variações morfológicas são eliminadas com a retirada de prefixos e sufixos das palavras, considerando somente os radicais. Outras características como gênero, número e grau das palavras podem ser removidas nessa etapa.

3.3.4 Cálculo da relevância e seleção dos termos

Nessa fase, é calculado a importância dos termos em cada texto. O cálculo é baseado na frequência, análise estrutural ou posição sintática da palavra. Já o processo de selecionar palavras importantes do texto, pode ser baseado no peso dos termos ou posição sintática. A técnica usada neste trabalho é a "filtragem baseada no peso de termo". Essa técnica elimina termos que aparecem poucas vezes nos textos.

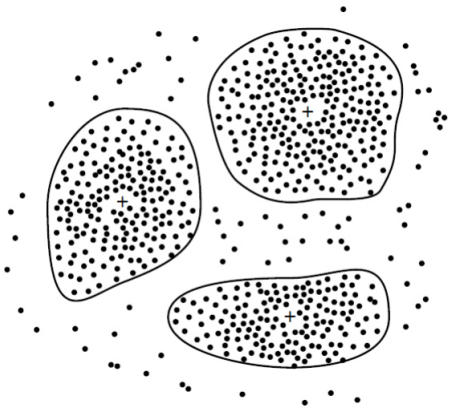
3.3.5 Pós-processamento

Execução do algoritmo que irá consumir os termos e resultados obtidos na Mineração de Texto. Dentre as possibilidades de aplicação para mineração textual encontram-se as clusterizações, ou seja, algoritmos que buscam agrupar dados semelhantes. Tem o intuito de detectar padrões e descobrir *insights* que possam ser utilizados para tomar decisões e responder a questionamentos referentes a base de dados minerada.

3.4 Algoritmos de clusterização

Algoritmos de clusterização tem como tarefa identificar e agrupar automaticamente dados pelo seu Grau de Similaridade. Um *cluster* (ou agrupamento) é um conjunto de dados semelhantes. Um exemplo abstrato de um *cluster* é mostrado na Figura 1, que ilustra uma base de dados classificada com 3 *clusters*.

Figura 1 – 3 *Clusters*



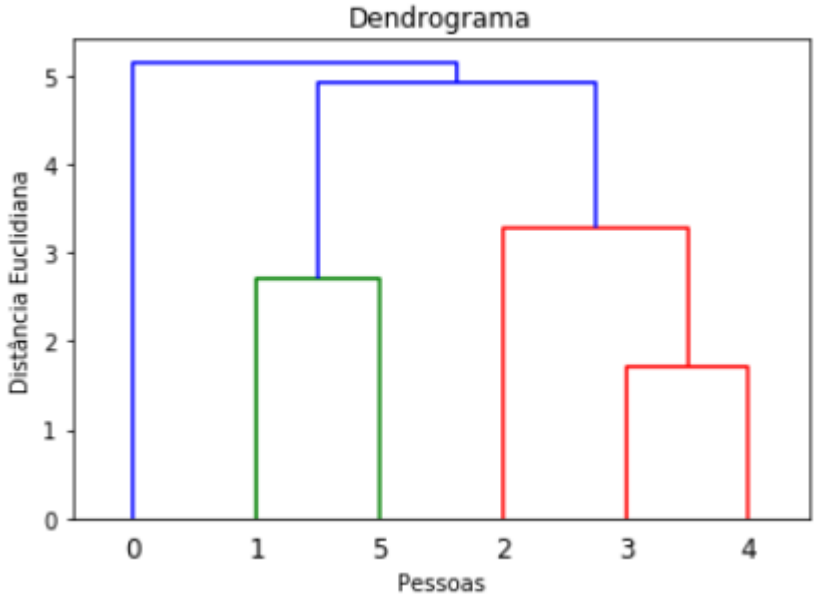
Fonte: (HAN *et al.*, 2011)

3.4.1 Hierárquicos

Os *clusters* Hierárquicos fundamentam-se nos agrupamentos e desagrupamentos dos elementos. A principal característica de modelo hierárquico é o dendrograma (diagrama bidimensional) formado pelo modelo. De acordo com Metz e Monard (2005) "Essa técnica permite analisar os *clusters* em diferentes níveis de granularidade, pois cada nível do dendrograma descreve um conjunto diferente de agrupamentos." A partir do dendrograma formado, basta escolher a quantidade de *clusters* de interesse.

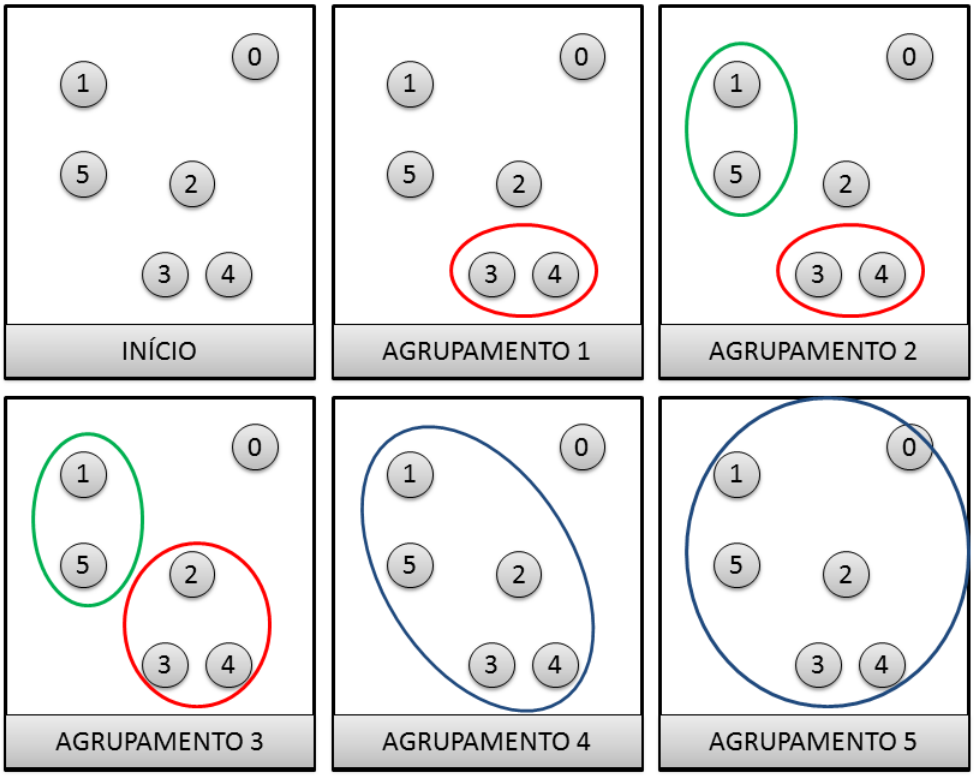
O algoritmo segue os seguintes passos: inicialmente cada elemento é um grupo, e a cada iteração, um grupo é ligado a outro grupo baseado em sua similaridade, até que todos os elementos estejam no mesmo grupo. A Figura 2 apresenta um exemplo de um dendrograma gerado pelo vizinho mais próximo. A seguir, a Figura 3 apresenta a sequência de agrupamentos realizada pelo algoritmo.

Figura 2 – Dendrograma



Fonte: Elaborado pelo Autor (2020).

Figura 3 – Sequência de agrupamentos



Fonte: Elaborado pelo Autor (2020).

4 TRABALHOS RELACIONADOS

Este capítulo descreve trabalhos da literatura mais relevantes para a contextualização do problema proposto nesta monografia.

4.1 *Redes sociais e a força de suas conexões*

Com o crescente uso das redes sociais, a possibilidade de identificar a força de uma conexão entre pares de pessoas se intensificou no últimos anos, sendo uma medida muito utilizada para: formação de grupos de trabalho, aplicação em negócio e no *marketing digital*. O trabalho de Gilbert e Karahalios (2009) determina a força de uma conexão através da combinação de sete dimensões, que indicam o nível de relacionamento entre pares de pessoas. Pessoas que apresentam determinados graus de confiança e que fazem parte do mesmo ciclo de amizade, são consideradas pessoas com uma forte conexão.

No trabalho em questão os autores usaram dados extraídos de 35 contas da rede social *online Facebook*. No modelo, a Força de Conexão é uma combinação entre sete dimensões definidas na literatura (GILBERT; KARAHALIOS, 2009): Intensidade, Intimidade, Duração, Serviços recíprocos, Estrutural, Apoio emocional e Distância social. Através da análise das interações coletadas e um questionário para medir a Força de Conexão, foram encontradas uma correlação linear entre 74 variáveis. Essas variáveis são as ações encontradas no *Facebook como*: palavras trocadas no mural; número de amigos em comum; *posts* no mural de amigos entre outros.

O trabalho apresenta 15 das 74 variáveis preditas encontradas: dias desde a última comunicação; dias desde a primeira comunicação; intimidade x estrutural; palavras trocadas; amigos em comum; diferença educacional entre outros. O modelo foi capaz de prever a força de conexão com uma precisão de 90% a partir dos relacionamentos estabelecidos em uma rede social. Com isso o estudo mostra que a força de conexão também se manifesta nos meio de comunicação social.

Esta pesquisa difere do trabalho de Gilbert e Karahalios (2009) no seguinte ponto: Gilbert e Karahalios (2009) usou um questionário como apoio para determinar a força de conexão. A proposta dessa pesquisa é usar somente as contas da Rede Social *Online* de cada pessoa. A semelhança entre os trabalhos, estar na proposta de encontrar as variáveis importantes para o modelos.

4.2 *Afinidade de grupos de trabalho*

Tarefas colaborativas são muito comuns na vida de qualquer ser humano, e a seleção de pessoas para compor grupos faz parte desse contexto. Tendo essa problemática como foco, o trabalho de Castilho *et al.* (2014) busca entender quais os fatores que influenciam na decisão de selecionar alguém para fazer parte de um grupo. A análise do comportamento das pessoas em ambientes virtuais são indícios de como se comportam na vida real, logo é possível determinar o seu comportamento em decisões de colaboração.

Os autores selecionaram uma sala de aula com 31 estudantes da graduação e os aplicaram um teste, perguntando como era o relacionamento profissional com todos os outros alunos da sala de aula. Para uma análise precisa, coletaram informações sobre o desempenho em sala de aula, e dados sobre a interação social com outros alunos. Cada aluno respondia um questionário com a pergunta: "Você gostaria de trabalhar com esta pessoa?" para cada um dos colegas de sala e as opções eram: "sim", "não" ou "indiferente".

A conclusão mostram que os resultado providos pelo *Facebook*, são mais informativos sobre com quem os alunos desejam trabalhar. Há uma relação entre a formação de equipe *offline* e as interações *online*, ou seja as relações na vida real são refletidas no comportamento das redes sociais.

A semelhança entre os trabalhos é buscar identificar se as conclusões de Castilho *et al.* (2014) se aplica a base de dados usada este trabalho, ou seja será que os interesses apresentados pelos usuários são refletidos nas amizades formadas nas redes sociais?

4.3 *Modelagem da força de relacionamento*

Assim como no trabalho anterior, a pesquisa desenvolvida por Xiang *et al.* (2010) propõe um modelo para inferir a Força de Conexão como objetivo de distinguir as relações fortes de fracas. O modelo é baseado nos padrões de interação e aproximação dos usuários. O mesmo é implementado com um esquema de otimização, para extrair os pontos fortes e os parâmetros dos relacionamentos, para um conjunto de pares de usuários.

Para construir o *dataset* usado no trabalho foram selecionados aleatoriamente contas da Rede Social *Online Facebook*, e resultou uma amostra no total de 4500 nós. Inicialmente, para uma amostragem de usuários foi observada as variáveis de interação como: se existe uma conexão entre os usuários; compartilhamento de fotos e mensagens; frequenta grupos em comum;

equivalência de profissão e por fim se tem amigos em comum. A força do relacionamento é uma variável dependente das variáveis de interação citadas inicialmente.

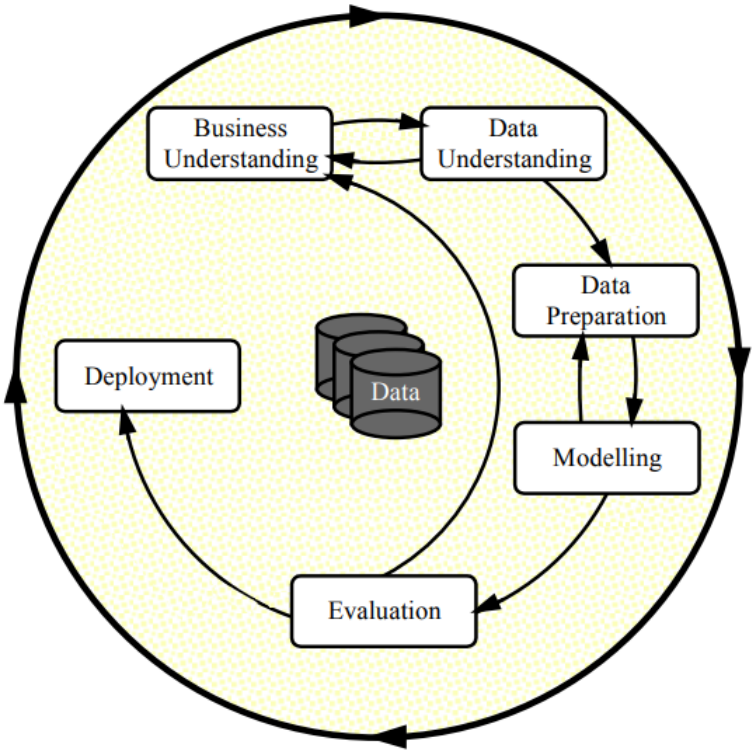
Os experimentos mostraram que o gráfico gerado pelos pontos fortes estimados do relacionamento, são melhores (melhor desempenho de classificação e maior autocorrelação) do que os gráficos gerados pelos dados brutos. O modelo pode ser usado para melhorar o desempenho das redes sociais de várias maneiras, pois resulta em um gráfico ponderado em que as ligações importantes são destacadas e as ligações fracas são reduzidas.

Diferente do trabalho de Xiang *et al.* (2010), a proposta desta pesquisa é criar uma métrica para medir a Força de Conexão (em número) e não classificar a relação, como uma força fraca ou forte. Assim como Xiang *et al.* (2010), o projeto irá selecionar aleatoriamente as contas para serem analisadas.

5 PROCEDIMENTOS METODOLÓGICOS

A metodologia deste trabalho foi baseada no modelo CRISP-DM (*Cross-Industry Standard Process of Data Mining*), por ser considerado o padrão de maior aceitação e devido a ampla literatura disponível (RIVO *et al.*, 2012). A Figura 4 ilustra as fases do modelo:

Figura 4 – Fases do CRISP-DM



Fonte: (WIRTH; HIPP, 2000)

5.1 Entendimento do negócio (*Business Understanding*)

A primeira fase consistiu em entender o problema, ou seja, quais objetivos era desejado atingir com a Mineração dos Dados. Neste trabalho, foi realizado um estudo sobre como as redes sociais têm sido usadas nos últimos anos e quais informações são possíveis inferir nesta gigantesca rede de interação.

Através de um estudo de viés científico/literário, foi realizado uma busca sobre os padrões de interação que influenciam a força de uma conexão, ou seja, quais ações executadas na Rede Social *Online Facebook* podem apresentar indícios de existência de relações sobre os indivíduos analisados. Foi definido a utilização de 4 dimensões para medir a Força de Conexão, como mencionado no Seção 3.2.

5.2 Entendimento dos dados (*Data Understanding*)

Nesta fase foi realizada uma análise exploratória sobre os dados disponíveis, organizando-os e documentando-os. Durante a fase de entendimento dos dados, as informações importantes foram identificadas e analisadas, buscando mensurar os padrões existentes na base de dados estudada. Os dados usados nesta pesquisa foram obtidos a partir de uma extração de posts públicos da Rede Social *Online Facebook*. Após a anonimização dos dados, estes foram organizados relacionando cada perfil a um usuário anônimo. Para um melhor entendimento das próximas etapas metodológicas, é apresentado a seguir a descrição dos termos comumente utilizados para definição de algumas ações na Rede Social *Online Facebook:*

1. **Perfil**: é uma conta no *Facebook* associada a uma pessoa física conhecido como usuário;
2. **Amigos**: cada perfil pode estar ligado a outros perfis por uma relação de "amizade";
3. **Publicações**: cada perfil existente no *Facebook* pode fazer publicações. As publicações podem contém textos, imagens, vídeos, links, dentre outras formas de compartilhamento de informações. Um *repost* da publicação de outra pessoa também é considerado uma publicação. Uma publicação tem alguns atributos como: título, mensagem, data, imagem, etc;
4. **Reações**: amigos do perfil podem reagir a uma publicação feita pelo mesmo. A reação é uma forma de mostrar a opinião sobre determinado tema presente na publicação.

5.2.1 Estrutura de dados utilizada para gerenciamento da base de dados

Os dados usados neste trabalho são representados no formato *JavaScript Object Notation* (JSON) - Notação de Objetos JavaScript. O formato JSON abaixo descreve a estrutura da base de dados usada no trabalho e posteriormente na Tabela 1 há uma descrição, e o tipo de cada variável.

```
{"identidadeFacebook": "perfil_1",
 "amigos": [
   "perfil_2",
   "perfil_3"
 ],
 "postagens": [
   {"titulo:"titulo_publicacao",
```

```
8       "data:"data_publicacao",
9       "mensagem:"mensagem_publicacao",
10      "linkDaPostagem:"link_publicacao",
11      "reacoes": [
12        {"usuario": "perfil_2", "tipo": "tipo_reacao"}
13      ]}
14   ]
15  }
```

Tabela 1 – Descrição das variáveis

Variável	Descrição	Tipo
identidadeFacebook	Armazena o identificador do perfil no *Facebook*. Esse identificador é único para cada perfil.	String
amigos	Armazena os amigos do perfil, ou seja, contém uma lista de identificadores.	Array
postagens	Armazena uma lista de publicações feitas pelo perfil.	Array
titulo	Cada publicação tem um título, normalmente esse título guarda o nome do perfil.	String
data	Armazena a data da publicação.	String
mensagem	Armazena o texto da publicação.	String
linkDaPostagem	Armazena um link que dá acesso a publicação.	String
reacoes	Armazena a lista de reações que a publicação recebeu.	Array
usuario	Armazena o identificador do perfil que reagiu.	String
tipo	Armazena o tipo da reação feita pelo usuário, que pode ser: curti, amei, força, haha, uau, triste, grr.	String

Fonte: Elaborado pelo Autor (2020).

5.2.2 Tamanho da base

- **Quantidade de perfis**: neste trabalho foi utilizado uma base com 1.011 documentos, ou seja, a descrição de 1.011 perfis distintos. Cada um desses perfis tem uma lista de amigos. No total foram descobertos 46.829 perfis distintos;
- **Quantidade de publicações**: A base contém no total 25.685 publicações. Algumas dessas publicações têm a variável "mensagem" igual a *null*, isso acontece quando a publicação possui somente uma imagem sem nenhuma descrição. Publicações assim não foram

analisadas, e portanto removidas da base de dados, pois não é possível tirar nenhuma informações das mesmas. Ao total a base contém 15.748 publicações com a variável "mensagem" diferente de *null*.

5.3 Preparação dos dados (*Data Preparation*)

Nesta fase os dados foram preparados para modelagem (próxima etapa) que consistirá em transformar os dados brutos no conjunto de dados final. A mineração dos dados foi realizada considerando diversos fatores como a relevância, qualidade, restrições e limites no tipo dos dados.

5.3.1 Mineração de texto

Como mencionado anteriormente, o processo de Mineração de Texto foi baseado na metodologia apresentada na Fundamentação Teórica (Seção 3.3), a seguir são apresentadas as tomadas de decisão de cada etapa metodológica.

- **Seleção de texto**: A construção da personalidade de cada perfil foi baseada nas publicações do mesmo. Desta forma, o foco foi a mineração de publicações textuais utilizando a variável "mensagem". Como não faz sentido analisar textos vazios, o processo de mineração foi realizado somente nas 15.748 mensagens existentes na base de dados que são diferente de *null*.
- **Definição de abordagem dos dados**: Usando Análise Estatística, foi calculada a frequência de cada palavra presente nas publicações. Ao total foram encontradas 21.408 palavras diferentes.
- **Indexação e normalização**: Todas as publicações passaram por alguns tratamentos para normalização dos dados. Na fase de identificação de termos, o texto foi convertido para letras minúsculas. Acentos e caracteres especiais foram removidos. A fase de remoção de *Stopwords* também foi realizada, porém para evitar uma diminuição na precisão de busca, a fase de normalização morfológica não foi executada.
- **Cálculo da relevância dos termos e Seleção dos termos**: Após os tratamentos realizados nas fases anteriores, o cálculo da relevância foi realizado considerando as palavras com frequência maior que quatro. Desta forma, base de dados passou de 21.408 palavras para 1.006 palavras (termos selecionados).

5.3.2 Construção de categorias

O processo de construção das categorias consistiu na seleção e divisão das palavras em grupos. De acordo com os termos selecionados na fase de Mineração de Texto, foram criadas pela autora 19 categorias. Essas categorias estão de acordo com as dimensões que mensuram a Força de Conexão, apresentadas por Burt (2009), conforme especificado na Fundamentação Teórica (Seção 3.2). Por tanto as seguintes categorias foram criadas:

- **Intimidade**: Trabalho, Doutrina e Relacionamento;
- **Serviços** : Tecnologia, Rede Social e Pandemia;
- **Estrutural**: Futebol, Esporte, Artes, Lúdico, Signo, Música, Comida, Animal, Maquiagem e Droga;
- **Distância Social**: Política, Religião e Escolaridade.

Cada termo selecionado foi analisado podendo ser enquadrado em pelo menos uma das categorias criadas ou então descartado, caso não se enquadrasse em nenhuma categoria. Desta forma, o conjunto de 1.006 termos foi categorizado, retornando um total de 345 termos que se enquadram em ao menos uma das categorias definidas. O Apêndice A, apresenta os termos relacionados com a sua categoria. Para abranger ainda mais a busca, outros termos semelhantes aos selecionados foram adicionados e algumas palavras foram colocados no plural. Após estas verificações de abrangência, o número de palavras pertencentes a pelo menos uma categoria passou de 345 para 433.

5.3.3 Seleção das publicações

A fase de seleção das publicações pode ser considerada uma das etapas mais importante para preparação dos dados, pois ela elimina as mensagens sem informações válidas. Como citado anteriormente, a base de dados contém 15.748 mensagens diferente de *null*, mas foi constatado que nem todas essas publicações poderiam ser utilizadas para determinar a personalidade dos perfis. Para realizar essa seleção, foi utilizado o grupo de termos definidos na fase anterior. A seguinte regra para seleção foi estabelecida: se a mensagem (da publicação) contiver pelo menos uma palavra dos termos selecionados, a publicação é classificada como uma publicação válida. Ao total foram selecionadas 3.618 publicações válidas.

5.3.4 Seleção de Perfis

Com objetivo de construir a personalidade de cada perfil, baseado nas reações em cada publicação, definiu-se que a seleção de perfis seria padronizada de acordo com as reações à publicações que se enquadrassem em ao menos uma das categoria definidas na Seção 5.3.2. Dessa forma, a seleção de perfis foi realizada através da contabilidade de categorias distintas que cada perfil reagiu, a seguir é apresentado a quantificação dos perfis selecionados.

- 35.570 perfis não reagiram a nenhuma publicação que contivesse termos de pelo menos uma categoria;
- 7.403 perfis reagiram a publicações relacionadas apenas uma categoria;
- 2.367 perfis reagiram a publicações relacionadas a duas categorias distintas;
- 719 perfis reagiram a publicações relacionadas a três categorias distintas;
- 316 perfis reagiram a publicações relacionadas a quatro categorias distintas;
- 160 perfis reagiram a publicações relacionadas a cinco categorias distintas;
- 149 perfis reagiram a publicações relacionadas a seis categorias distintas;
- 58 perfis reagiram a publicações relacionadas a sete categorias distintas;
- 42 perfis reagiram a publicações relacionadas a oito categorias distintas;
- 16 perfis reagiram a publicações relacionadas a nove categorias distintas;
- 29 perfis reagiram a publicações relacionadas a dez ou mais categorias distintas.

Logo, quanto mais publicações de categorias diferentes um perfil reagir, mais informações é possível obter sobre este perfil. Isso acontece porque cada categoria informa um interesse do perfil a respeito de algo, como por exemplo: se um perfil 'curtiu' publicações que contém nome de times é provável que esse perfil goste de futebol. Por tanto, para análise deste trabalho foram escolhidos os perfis que reagiram a publicações relacionadas à 3 ou mais categorias distintas, totalizando 1.489 perfis.

5.4 Modelagem (*Modelling*)

Nessa fase alguns algoritmos foram criados para obter determinados objetivos, sendo testado também alguns modelos inteligentes para realização de agrupamentos. Para alcançar estes objetivos, algumas matrizes contendo informações importantes para o funcionamento dos algoritmos e modelos estudados foram definidas.

5.4.1 Matriz Amizade

Antes de definir o Grau de Similaridade entre os perfis, foi mapeado o relacionamento de amizade entre eles. Para tal, foi criado uma matriz de adjacência, nomeada como Matriz Amizade A_{nxn}, onde n é o número de perfis estudado. Desta forma, caso haja uma amizade entre o perfil i e o perfil j, $a_{i,j} = 1$ e caso contrário $a_{i,j} = 0$. A Figura 5 demonstra um exemplo real da Matriz Amizade dos 10 primeiros perfis estudados. Neste trabalho, a Matriz Amizade A é simétrica, ou seja, $a_{i,j} = a_{j,i}$.

Figura 5 – Matriz Amizade

	0	1	2	3	4	5	6	7	8	9	...	1479	1480	1481	1482	1483	1484	1485	1486	1487	1488
0	0	1	1	1	1	1	1	1	1	1	...	0	0	0	0	0	0	0	0	0	0
1	1	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
2	1	0	0	0	0	1	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
3	1	0	0	1	0	1	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
4	1	0	0	0	0	1	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
5	1	0	1	1	1	1	0	0	0	1	...	0	0	0	0	0	0	0	0	0	0
6	1	0	0	0	0	0	1	0	0	0	...	0	0	0	0	0	0	0	0	0	0
7	1	0	0	0	0	0	0	1	0	1	...	0	0	0	0	0	0	0	0	0	0
8	1	0	0	0	0	0	0	0	0	1	...	0	0	0	0	0	0	0	0	0	0
9	1	0	0	0	0	1	0	1	1	1	...	0	0	0	0	0	0	0	0	0	0

Fonte: Elaborado pelo Autor (2020).

5.4.2 Matriz Perfil-Categoria

A Matriz Perfil-Categoria C_{nxm}, onde n é o número de perfis estudado e m o número de categorias, foi criada para representar quais perfis reagiram a quais categorias. Os valores contidos nessa matriz fazem referência a quantidade de vezes que cada perfil reagiu a uma categoria. A Matriz Perfil-Categoria C é definida como:

$$C_{ij} = \begin{cases} 10.0 + X * 0.1, & \text{tal que } X \text{ é o número de vezes que o perfil } i \text{ reagiu a categoria } j \\ 0.0, & \text{caso } i \text{ não tenha reagido a nenhuma publicação da categoria } j \end{cases} \quad (5.1)$$

A Figura 6 demonstra um exemplo real dos dados contidos na Matriz Perfil-Categoria. Na imagem tem a demonstração das reações dos 10 primeiros perfis.

Figura 6 – Matriz Perfil-Categoria

	Polít	Relig	Estud	Traba	Doutr	Relac	Futeb	Espor	Artes	Lúdic	Signo	Músic	Comid	Anima	Maqui	Droga	Pande	Tecno	Faceb
0	10.3	0.0	0.0	0.0	0.0	10.3	0.0	0.0	10.4	10.1	10.1	10.2	10.1	10.2	10.1	0.0	0.0	0.0	10.3
1	10.2	0.0	0.0	0.0	0.0	10.1	10.1	0.0	10.1	10.1	0.0	10.1	0.0	0.0	0.0	0.0	0.0	0.0	10.1
2	10.2	0.0	0.0	0.0	0.0	10.2	0.0	0.0	10.1	0.0	0.0	0.0	0.0	10.1	0.0	0.0	0.0	10.1	10.1
3	10.2	0.0	0.0	0.0	10.1	10.2	0.0	10.2	10.1	10.1	10.1	10.2	0.0	10.2	0.0	0.0	0.0	0.0	10.1
4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	10.1	0.0	0.0	0.0	0.0	10.1	0.0	0.0	0.0	0.0	10.3
5	10.2	0.0	0.0	10.1	0.0	0.0	0.0	0.0	10.2	10.5	10.1	10.1	10.1	0.0	10.1	0.0	0.0	0.0	10.1
6	10.2	10.2	0.0	0.0	10.2	10.7	0.0	10.1	10.2	10.4	0.0	10.2	10.1	10.1	0.0	10.1	10.4	10.1	10.7
7	10.1	0.0	10.1	0.0	0.0	10.1	0.0	0.0	10.1	10.1	0.0	0.0	0.0	10.2	0.0	0.0	10.1	0.0	10.1
8	10.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	10.1	10.2	0.0	0.0	0.0	10.2	0.0	0.0	0.0	10.1	10.1
9	10.2	10.3	0.0	0.0	0.0	10.2	0.0	0.0	10.1	10.1	0.0	0.0	0.0	0.0	0.0	0.0	10.1	10.1	10.3

Fonte: Elaborado pelo Autor (2020).

Como demonstrado anteriormente há um intervalo de pelo menos 10 pontos entre uma categoria não reagida e outra reagida. Essa decisão foi tomada para enfatizar que a reação de um perfil a uma categoria é extremamente relevante em consideração a um perfil que não tenha reagido a categoria mencionada. Na Tabela 2 há um exemplo com 3 perfis fictícios, considerando somente a categoria Política.

Tabela 2 – Exemplo com 3 Perfis

	Política
Perfil 1 (P1)	0
Perfil 2 (P2)	10.1
Perfil 3 (P3)	10.9

Fonte: Elaborado pelo Autor (2020).

- P1, não reagiu nada sobre política;
- P2, reagiu 1 publicação sobre política;
- P3, reagiu 9 publicações sobre política.

Note que desta forma, pode-se dizer que o perfil P2 é mais parecido com o perfil P3, sendo estes dois perfis são considerados muito distintos do perfil P1, pois este não reagiu a nada sobre política. Como o perfil P1 não reagiu a nada sobre política é provável que ele não tem interesse em publicações que dizem respeito a este tema. Diferente do perfil P3 que reagiu a nove publicações sobre política, com isso pode-se imaginar que o perfil P3 gosta muito de publicações a respeito de política. Note também que o perfil P1 'curtiu' pelo menos uma publicação sobre política e por mais que seja um número bem inferior ao perfil P3 ele demonstrou algum interesse e isso faz com ele se aproxime mais de P3.

5.4.3 Matriz Distância

Um dos objetivos do trabalho é definir um Grau de Similaridade entre dois perfis. Esse Grau de Similaridade (Força de Conexão) foi definido baseado nos interesse de cada perfil, representado pelas reações em cada categoria, como demonstrado na Matriz Perfil-Categoria. Para tal, foi calculado a distância entre todos os pares de perfis. Este cálculo foi realizado entre pares de perfis através da Distância Euclidiana. A distância euclidiana é a distância geométrica no espaço multidimensional. O cálculo da distância entre dois elementos $X = [X_1, X_2, ..., X_P]$ e $Y = [Y_1, Y_2, ..., Y_P]$ é:

$$d_{xy} = \sqrt{(X_1 - Y_1)^2 + (X_2 - Y_2)^2 + ... + (X_p - Y_p)^2} = \sqrt{\sum_{i=1}^{p} (X_i - Y_i)^2} \tag{5.2}$$

O cálculo foi realizado entre as linhas da Matriz Perfil-Categoria. Para representação das distâncias foi criada uma matriz, denominada Matriz Distância D_{nxn}, onde n é o número de perfis estudado, para representar a distância entre o perfil i e o perfil j. A Figura 7 demonstra um exemplo real da Matriz Distância dos 9 primeiros perfis estudados.

Figura 7 – Matriz Distância

	0	1	2	3	4	5	6	7	8
0	0.000000	22.633382	24.784067	20.253889	26.913937	17.674841	28.753435	24.785076	24.866644
1	22.633382	0.000000	22.584508	22.674435	24.781646	24.743282	30.481634	22.629406	22.629627
2	24.784067	22.584508	0.000000	24.821966	17.610508	30.469165	28.897924	20.200743	14.425672
3	20.253889	22.674435	24.821966	0.000000	26.874523	26.839151	24.919470	24.822168	24.863025
4	26.913937	24.781646	17.610508	26.874523	0.000000	28.747174	33.987939	22.585394	17.553062
5	17.674841	24.743282	30.469165	26.839151	28.747174	0.000000	33.840508	30.336447	26.762100
6	28.753435	30.481634	28.897924	24.919470	33.987939	33.840508	0.000000	28.689719	29.003965
7	24.785076	22.629406	20.200743	24.822168	22.585394	30.336447	28.689719	0.000000	20.200248
8	24.866644	22.629627	14.425672	24.863025	17.553062	26.762100	29.003965	20.200248	0.000000

Fonte: Elaborado pelo Autor (2020).

5.4.4 Análise da similaridade

Utilizando as matrizes apresentadas nas seções anteriores, foi criado um algoritmo para geração de uma nova métrica para a classificação do Grau de Similaridade (Força de Conexão) entre pares de perfis da Rede Social Online *Facebook*. Os padrões gerados pelo algoritmo são apresentados na Tabela 3, sendo os valores definidos através do cálculo da

Distância Euclidiana entre um perfil que não reagiu a nenhuma categoria (vetor-linha da Matriz Perfil-Categoria completamente zerado) com perfis que reagiram a: 1 categoria; 2 categorias ... até as 19 categorias.

Tabela 3 – Métrica de Similaridade

Distância Euclidiana	Grau de Similaridade	Quantidade de Categorias diferentes
Abaixo de 20.2	(5) Muito Similar	Até 3 categorias diferentes
Entre 20.2 e 28.55	(4) Similar	Entre 4 e 7 categorias diferentes
Entre 28.56 e 36.41	(3) Irrelevante	Entre 8 e 12 categorias diferentes
Entre 36.41 e 41.63	(2) Dissimilar	Entre 13 e 16 categorias diferentes
Acima de 41.64	(1) Muito Dissimilar	Acima de 17 categorias diferentes

Fonte: Elaborado pelo Autor (2020).

Desta forma, através do algoritmo que retorna as classificações de perfis da Rede Social Online *Facebook* conforme a Tabela 3, é possível definir o Grau de Similaridade entre pares de perfis, utilizando para tal a nova classificação para Força de Conexão definida neste trabalho. Esta classificação pode ser utilizada para análise de perfis em qualquer aplicação que tenha como principal interesse, obter informações sobre relacionamentos entre usuários da Rede Social Online *Facebook*.

Objetivando apresentar uma aplicação para o algoritmo proposto, o presente trabalho apresenta o seguinte questionamento, sobre relacionamentos entre usuários da Rede Social Online *Facebook*: “Interesses semelhantes realmente influenciam a amizade entre um grupo de usuários da rede social *Online Facebook*?”

Para essa análise, foi realizado o cruzamento entre a Matriz Amizade e a Matriz Distância. Como cada $a_{i,j} = 1$, significa que existe uma amizade entre o perfil i e o perfil j, foi utilizado sua combinação com a Matriz de Distância para definir o Grau de Similaridade entre o perfil i e o perfil j, como apresentado na Figura 8. Para definir o Grau de Similaridade geral de cada perfil, foi calculado a Moda do Grau de Similaridade dos amigos de cada perfil, como apresentado a Figura 9.

Figura 8 – Cruzamento das matrizes

Matriz Amizade

Aij	1	2	3	4	5	6
1	0	0	1	0	1	0
2	0	0	1	1	1	1
3	1	1	0	1	1	1
4	0	1	1	0	1	0
5	1	1	1	1	0	1
6	0	1	1	0	1	0

×

Matriz Distância

Dij	1	2	3	4	5	6
1	0	45.63	37.35	34.43	40.77	12.34
2	45.63	0	29.41	23.01	25.33	36.16
3	37.35	29.41	0	42.25	19.23	29.64
4	34.43	23.01	42.25	0	44.99	12.40
5	40.77	25.33	19.23	44.99	0	18.70
6	12.34	36.16	29.64	12.40	18.70	0

Fonte: Elaborado pelo Autor (2020).

Figura 9 – Cálculo da Moda

Moda

	1	2	3	4	5	6	MODA
1	-	-	2	-	2	-	2
2	-	-	3	4	4	3	4
3	2	3	-	1	5	3	3
4	-	4	1	-	1	-	1
5	2	4	5	1	-	5	5
6	-	3	3	-	5	-	3

LEGENDA:
5 - Muito Similar
4 - Similar
3 - Irrelevante
2 - Dissimilar
1 - Muito Dissimilar

Fonte: Elaborado pelo Autor (2020).

Através da análise da Moda é possível verificar qual o Grau de Similaridade do perfil considerando todos os seus amigos. Desta forma, no caso de um perfil que tenha vários amigos com Grau de Similaridade diferentes, será considerado o grau que mais aparece entre os amigos. Em caso de empate (como no perfil 2 da Figura 9), será considerado o valor mais alto. Esse resultado mostra que os amigos do perfil 5 (Figura 9), são muito similares a ele. Já os amigos do perfil 4 (Figura 9), são muito dissimilares a ele. Da mesma forma, as amizades dos perfis 3 e 6 são irrelevantes, ou seja, não são similares e nem dissimilares.

5.4.5 Formação de grupo

Outra importante aplicação do algoritmo para classificação de Força de Conexão apresentado neste trabalho é o suporte para criação de grupos similares com base nos dados da Rede Social Online *Facebook*. Com o algoritmo criado é possível medir a Força de Conexão

(similaridade) entre pares de usuários da rede. Desta forma, esta informação pode ser utilizada para geração de grupos (*clusters*) com perfis similares.

Dentre os algoritmos não supervisionados apresentados na literatura para construção de *clusters*, destaca-se o algoritmo Hierárquico, descrito no Capítulo 3, que clusteriza dados buscando a maior similaridade (menor Distância Euclidiana) possível entre seus elementos. Para exemplificar sua aplicação, foi escolhido 10 perfis aleatórios na base de dados para serem divididos em 2 grupos. Como resultado, o algoritmo retornou um grupo com 7 perfis (Figura 10) e outro com 3 perfis (Figura 11).

Figura 10 – Algoritmo Hierárquico - Grupo 1

	Polit	Relig	Estud	Traba	Doutr	Relac	Futeb	Espor	Artes	Lúdic	Signo	Músic	Comid	Anima	Maqui	Droga	Pande	Tecno	Faceb
0	10.2	0.0	0.0	0.0	10.1	10.3	10.1	0.0	10.2	10.4	10.1	10.1	10.1	0.0	0.0	0.0	0.0	0.0	10.1
1	10.1	0.0	0.0	0.0	0.0	10.3	0.0	0.0	0.0	0.0	10.1	10.1	0.0	10.2	0.0	0.0	0.0	0.0	10.2
2	0.0	10.1	0.0	0.0	0.0	10.1	0.0	0.0	10.3	10.5	0.0	0.0	0.0	10.3	0.0	0.0	0.0	0.0	10.5
3	10.2	0.0	0.0	0.0	0.0	10.2	0.0	10.1	10.1	10.3	0.0	10.1	10.1	0.0	0.0	0.0	0.0	10.1	10.3
4	0.0	0.0	0.0	0.0	0.0	10.1	0.0	10.1	0.0	10.1	0.0	0.0	10.1	10.1	0.0	0.0	0.0	0.0	10.4
5	10.2	10.1	0.0	10.1	0.0	10.3	0.0	0.0	10.1	10.2	10.1	10.1	10.2	10.1	10.1	0.0	10.4	0.0	10.3
6	10.1	0.0	0.0	0.0	0.0	10.2	10.1	0.0	0.0	10.1	0.0	10.1	0.0	0.0	0.0	0.0	10.2	0.0	0.0

Fonte: Elaborado pelo Autor (2020).

Figura 11 – Algoritmo Hierárquico - Grupo 2

	Polit	Relig	Estud	Traba	Doutr	Relac	Futeb	Espor	Artes	Lúdic	Signo	Músic	Comid	Anima	Maqui	Droga	Pande	Tecno	Faceb
0	0.0	0.0	10.1	0.0	0.0	0.0	0.0	10.1	0.0	0.0	0.0	0.0	10.2	0.0	0.0	0.0	0.0	0.0	0.0
1	0.0	10.1	10.1	0.0	0.0	10.1	0.0	0.0	10.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	0.0	0.0	10.1	0.0	0.0	0.0	0.0	0.0	10.1	10.3	0.0	10.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Fonte: Elaborado pelo Autor (2020).

Principais características dos grupos formados:

- A maioria dos perfis que reagiram a publicações sobre os temas Política, Relacionamento, Lúdico, Animais e *Facebook* encontram-se no Grupo 1;
- A maioria dos perfis que reagiram a publicações sobre o tema Estudante encontram-se no Grupo 2.

Uma característica do modelo de clusterização gerado, através do algoritmo Hierárquico, é a formação de grupos de tamanhos indefinidos. Esta questão pode se tornar um problema, quando se deseja formar grupos com a mesma quantidade de perfis. Como exemplo, pode-se citar a aplicabilidade empresarial, onde busca-se por grupos com tamanhos pré-definidos, comumente utilizados para geração de equipes de trabalho.

Tendo em vista esta problemática, surgiu o seguinte questionamento: "seria possível

formar grupos similares com a mesma quantidade de perfis?". Buscando responder este tipo de questionamento, foi realizada uma modificação no algoritmo Hierárquico estudado, objetivando gerar um modelo que pudesse garantir que todos os grupos contivessem a mesma quantidade de elementos, ao mesmo tempo que também fosse maximizado as similares entre os perfis. O algoritmo 1, a seguir apresenta as modificações realizadas.

Algoritmo 1: Algoritmo Hieráquico Modificado

Entrada: A Matriz Distância, contendo a distância entre todos elementos; $l =$ Uma lista com todos os elementos; $k =$ Quantidade de grupos; $tam =$ Tamanho dos grupos.
Saída: Os grupos formados.
$disponivel \leftarrow l$
$g \leftarrow 0$
while $g < k$ **do**
  Seleciona o par (x, y) mais próximos em $disponivel$
  Cria um novo grupo com o par (x, y) selecionado
  Remova os elementos x e y de $disponivel$
  $g \leftarrow g + 1$
end while
while $disponivel \neq vazio$ **do**
  Calcula a distância euclidiana de todos os grupos menores que tam para todos elementos em $disponivel$
  Adiciona o elemento de menor distância ao grupo
  Remove o elemento de $disponivel$
end while
return grupos

Utilizando o modelo gerado através do algoritmo Hierárquico modificado e a mesma base de dados com 10 perfis, do exemplo anterior, foi possível formar dois grupos como apresentados nas Figuras 12 e 13. Pode-se notar que as características dos grupos formados são semelhantes as definidas pelo modelo Hierárquico tradicional, ou seja:

- A maioria dos perfis que reagiram a publicações sobre o tema Política, Relacionamento, Lúdico, Animais e *Facebook* encontra-se no Grupo 1.
- A maioria dos perfis que reagiram a publicações sobre o tema Estudante encontra-se no Grupo 2;

Figura 12 – Algoritmo Hierárquico Modificado - Grupo 1

	Polit	Relig	Estud	Traba	Doutr	Relac	Futeb	Espor	Artes	Lúdic	Signo	Músic	Comid	Anima	Maqui	Droga	Pande	Tecno	Faceb
0	10.2	0.0	0.0	0.0	10.1	10.3	10.1	0.0	10.2	10.4	10.1	10.1	10.1	0.0	0.0	0.0	0.0	0.0	10.1
1	10.1	0.0	0.0	0.0	0.0	10.3	0.0	0.0	0.0	0.0	10.1	10.1	0.0	10.2	0.0	0.0	0.0	0.0	10.2
2	0.0	10.1	0.0	0.0	0.0	10.1	0.0	0.0	10.3	10.5	0.0	0.0	0.0	10.3	0.0	0.0	0.0	0.0	10.5
3	10.2	0.0	0.0	0.0	0.0	10.2	0.0	10.1	10.1	10.3	0.0	10.1	10.1	0.0	0.0	0.0	0.0	10.1	10.3
4	0.0	0.0	0.0	0.0	0.0	10.1	0.0	10.1	0.0	10.1	0.0	0.0	10.1	10.1	0.0	0.0	0.0	0.0	10.4

Fonte: Elaborado pelo Autor (2020).

Figura 13 – Algoritmo Hierárquico Modificado - Grupo 2

	Polit	Relig	Estud	Traba	Doutr	Relac	Futeb	Espor	Artes	Lúdic	Signo	Músic	Comid	Anima	Maqui	Droga	Pande	Tecno	Faceb
0	0.0	0.0	10.1	0.0	0.0	0.0	0.0	10.1	0.0	0.0	0.0	0.0	10.2	0.0	0.0	0.0	0.0	0.0	0.0
1	0.0	10.1	10.1	0.0	0.0	10.1	0.0	0.0	10.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	0.0	0.0	10.1	0.0	0.0	0.0	0.0	0.0	10.1	10.3	0.0	10.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0
3	10.2	10.1	0.0	10.1	0.0	10.3	0.0	0.0	10.1	10.2	10.1	10.1	10.2	10.1	10.1	0.0	10.4	0.0	10.3
4	10.1	0.0	0.0	0.0	0.0	10.2	10.1	0.0	0.0	10.1	0.0	10.1	0.0	0.0	0.0	0.0	10.2	0.0	0.0

Fonte: Elaborado pelo Autor (2020).

Por fim, com o objetivo de mensurar a qualidade dos grupos gerados pelos modelos propostos, utilizou-se o índice *Error Sum of Squares* (SSE) (Soma dos Erros Quadrados) que mede a soma do erro (ao quadrado) da distância de todos os elementos de um grupo para o centróide do mesmo, o SSE é definido como:

$$SSE = \sum_{i=1}^{k} \sum_{x_j \in C_i} d(x_j, \bar{x}_{C_i})^2,$$

onde k representa a quantidade de grupos e x_j um elemento contido no grupo, sendo $\bar{x}_{C_i}$ o centróide do grupo C_i.

A Tabela 4 compara o SSE dos *clusters* formados pelos algoritmo Hierárquico tradicional e o algoritmo Hierárquico modificado. Como o objetivo do Hierárquico modificado é formar grupos fortemente similares com a mesma quantidade de elementos, pode-se notar que embora o mesmo perca em similaridade em relação ao modelo gerado através do algoritmo Hierárquico tradicional, esta perda é consideravelmente pequena, podendo-se afirmar que os grupos encontrando ainda se mantém similares.

Tabela 4 – SSE

	Grupo 1	Grupo 2	SSE
Hieráquico tradicional	2118.92	480.11	2599.03
Hieráquico modificado	1689.71	1275.36	2965.07

Fonte: Elaborado pelo Autor (2020).

6 RESULTADOS

Este capítulo expõe os resultados dos algoritmos e modelos criados nesta monografia. Na Seção 6.1 mostra o resultado obtido pela análise das relações de amizade entre os perfis da base de dados. Enquanto a Seção 6.2 apresenta os resultados das execuções dos algoritmos de clusterização: Hierárquico e Hierárquico Modificado.

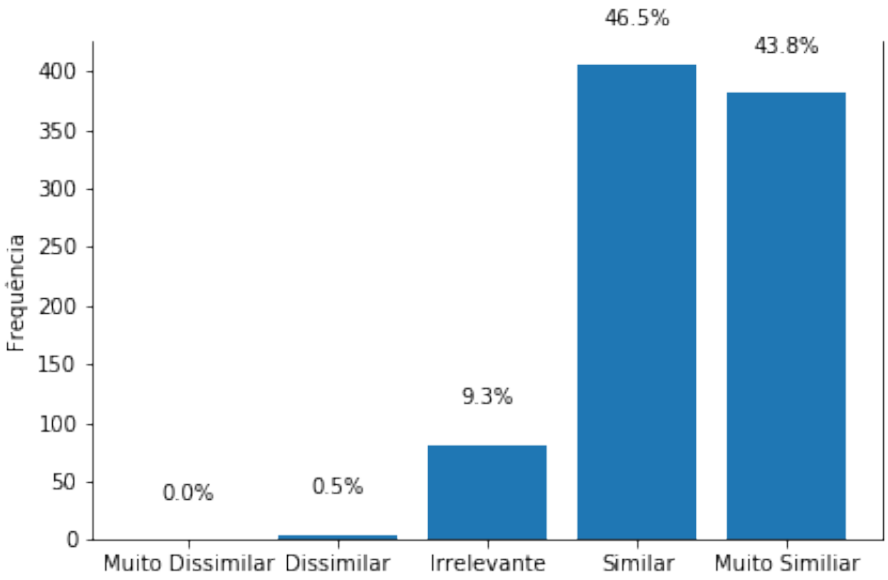
6.1 Análise da similaridade de todos os perfis

Castilho *et al.* (2014) concluiu no seu trabalho que os relacionamentos pessoais reais são refletidos nas interações em redes sociais. Logo, a similaridade entre duas pessoas na "vida real" é identificada também "vida virtual". Com base nessa conclusão, este trabalho apresentou um algoritmo para classificação da Força de Conexão entre usuários da Rede Social Online *Facebook*. Como exemplo de aplicação, a classificação obtida pelo algoritmo pode ser utilizada para responder se: "Interesses semelhantes realmente influenciam a amizade entre um grupo de usuários da rede social *Online Facebook*?". Como visto na Seção 5.3.4, a base de dados contém informações de 1.489 perfis, sendo a grande maioria destes perfis conectados pela relação de amizade do *Facebook*. Nesta rede apenas três perfis não estão conectados a nenhum outro perfil, ou seja, não possuem amigos. Os demais 1486 perfis estão divididos da seguinte forma:

- 873 perfis possuem somente um amigo;
- 530 perfis possuem de 2 a 10 amigos;
- 83 perfis possuem mais de 10 amigos.

Com o objetivo de responder o questionamento foram gerados alguns gráficos baseados na divisão apresentada acima. A Figura 14 contém o gráfico formado pelos perfis que possuem somente um amigo. Dos 873 perfis analisados, apenas 4 perfis (0.5%) possuem amizades Dissimilares, enquanto 81 perfis (9.3%) possuem amizades Irrelevantes. Em relação a similaridade, 406 perfis (46.5%) possuem amizades Similares e 382 perfis (43.8%) possuem amizades Muito Similares.

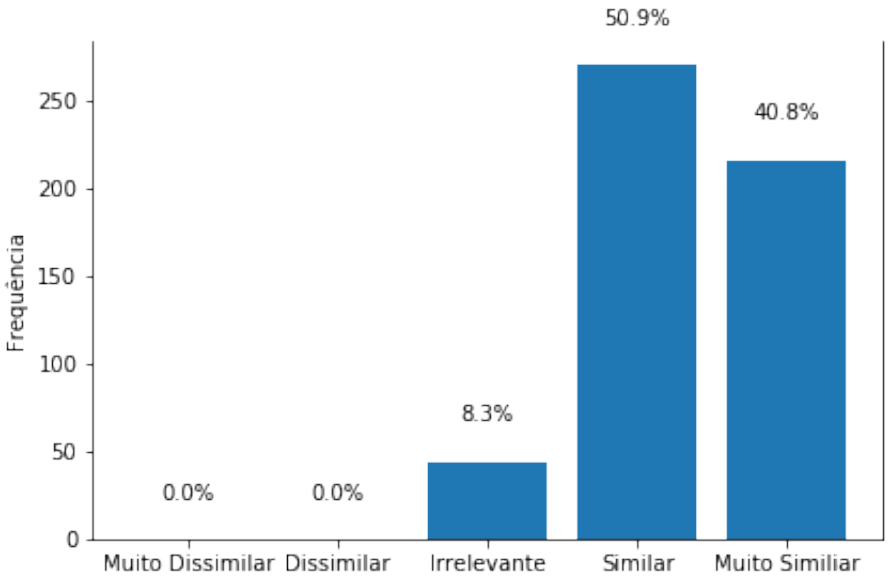
Figura 14 – Perfis com 1 amigo



Fonte: Elaborado pelo Autor (2020).

A Figura 15 apresenta o gráfico formado pelos perfis que possuem de 2 a 10 amigos. Dos 530 perfis analisados, 44 perfis (8.3%) possuem amizades Irrelevantes. Não havendo neste caso perfis que possuam amizades Dissimilares. Em relação a similaridade, 270 perfis (50.9%) possuem amizades Similares e 216 perfis (40.8%) possuem amizades Muito Similares.

Figura 15 – Perfis com 2 a 10 amigos

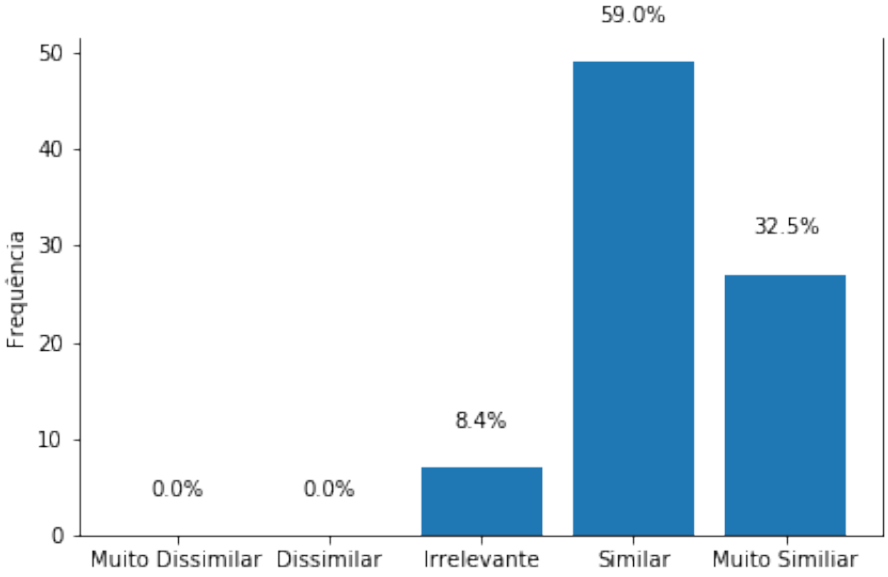


Fonte: Elaborado pelo Autor (2020).

Na Figura 16 contém o gráfico formado pelos perfis que possuem mais de 10 amigos.

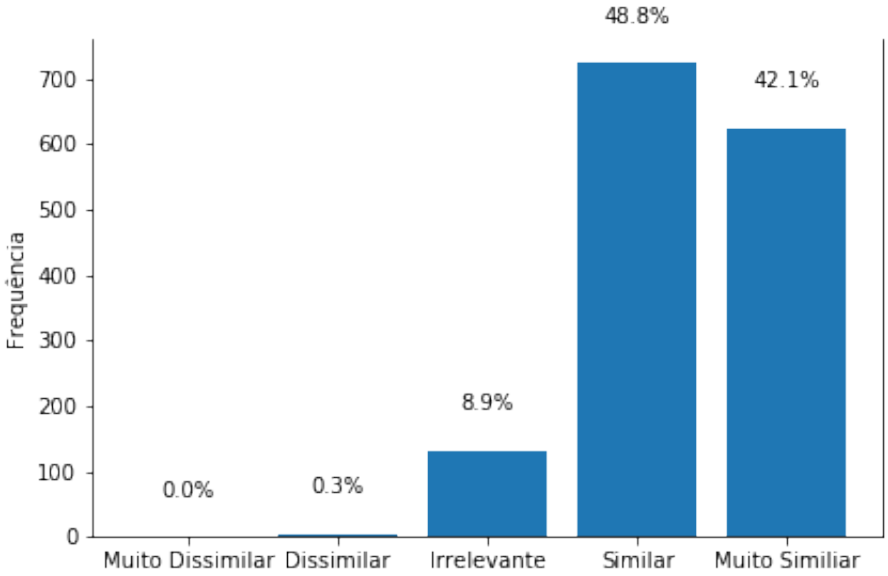
Dos 83 perfis analisados 7 perfis (8.4%) possuem amizades Irrelevantes. Não havendo neste caso perfis que possuam amizades Dissimilares. Em relação a similaridade, 49 perfis (59%) possuem amizades Similares e 27 perfis (32.5%) possuem amizades Muito Similares. Por fim, na Figura 17 contém um gráfico com todos perfis em uma análise geral.

Figura 16 – Perfis com mais de 10 amigos



Fonte: Elaborado pelo Autor (2020).

Figura 17 – Todos os perfis



Fonte: Elaborado pelo Autor (2020).

Os gráficos apresentam um padrão independente da quantidade de amigos que cada

perfil tenha, nenhum deles possui amigos “Muito Dissimilares”, pouquíssimos possui amigos “Dissimilares” ou “Irrelevantes”. A frequência de amigos “Similares” é sempre maior em todos casos analisados, seguido da classificação “Muito Similar”. Com isso é possível afirmar que, para a base de dados estudada, há uma relação entre os interesses dos usuários e as amizades por estes formadas na rede social, pois mais de 90% dos perfis tiveram suas amizades classificadas com “Similar” ou “Muito Similar”.

6.2 Análise dos algoritmos de clusterização

Para analisar a qualidade dos modelos inteligentes gerados a partir do algoritmo Hierárquico tradicional e modificado foram testados alguns conjuntos de perfis aleatórios da base de dados. Cada combinação (Quantidade de Perfis X Quantidade de Grupos) foi executada 5 vezes para ambos os algoritmos. As Tabelas 5, 6, 7, 8 mostram as execuções, para valores de 2, 5, 10 e 25 *clusters*, respectivamente. Nas tabelas apresentadas, cada coluna apresenta a quantidade de perfis em cada execução, enquanto cada linha apresenta as execuções de cada conjunto de perfis diferentes, apresentando o SSE do modelo gerado a partir do algoritmo Hierárquico tradicional (HT) e do algoritmo Hierárquico modificado (HM), assim como a diferença (em porcentagem) dos modelos ($HM - HT$), sendo o $HM - HT$ definido da seguinte forma:

$$HM - HT = (100 * SSE(HM))/SSE(MT) - 100$$

De posse do valor $HM - HT$, pode-se gerar as seguintes interpretações:

- Quanto mais próximo de 0% o valor de $HM - HT$ for, menor é diferença entre o SSE dos algoritmos;
- Se o valor de $HM - HT$ for negativo, significa que o modelo gerado pelo algoritmo Hierárquico modificado gerou grupos com o SSE menor do que algoritmo Hierárquico Tradicional. Logo, o agrupamento gerado pelo HM é mais similar do que o agrupamento gerado por HT;
- Analogamente, se o valor de $HM - HT$ for positivo, significa que o agrupamento de HT é mais similar que o HM;
- Se o valor de $HM - HT$ for igual a 0%, os valores de SSE dos grupos gerados pelos dois modelos são iguais.

Tabela 5 – 2 *Clusters*

		10 PERFIS	50 PERFIS	100 PERFIS	500 PERFIS
1	HT	1200.97	12518.01	25509.40	136014.05
	HM	1677.10	13163.44	25971.67	133879.73
	HM-HT	39.65%	5.16%	1.81%	-1.57%
2	HT	1926.96	12274.72	28094.48	138465.41
	HM	2009.09	12852.11	28086.66	137082.92
	HM-HT	4.26%	4.70%	-0.03%	-1.00%
3	HT	2057.59	12308.41	27240.42	140935.10
	HM	2125.13	13289.25	27468.56	139129.57
	HM-HT	3.28%	7.97%	0.84%	-1.28%
4	HT	2515.44	13321.90	26533.99	137057.45
	HM	2749.19	13750.26	26882.93	137417.75
	HM-HT	9.29%	3.22%	1.32%	0.26%
5	HT	1677.17	13014.21	26632.01	135483.62
	HM	1724.50	13233.16	26066.53	135052.58
	HM-HT	2.82%	1.68%	-2.12%	-0.32%

Fonte: Elaborado pelo Autor (2020).

Tabela 6 – 5 *Clusters*

		10 PERFIS	50 PERFIS	100 PERFIS	500 PERFIS
1	HT	448.89	10076.22	22881.02	124746.81
	HM	512.13	11165.43	22252.99	122431.50
	HM-HT	14.09%	10.81%	-2.74%	-1.86%
2	HT	957.47	9108.82	24663.41	123105.65
	HM	971.33	10191.16	23596.69	127045.28
	HM-HT	1.45%	11.88%	-4.33%	3.20%
3	HT	943.60	9821.30	23937.84	125480.30
	HM	1176.20	11382.41	22539.44	126075.47
	HM-HT	24.65%	15.90%	-5.84%	0.47%
4	HT	1101.54	11408.74	23463.71	128474.61
	HM	1227.36	11223.49	23404.32	123109.43
	HM-HT	11.42%	-1.62%	-0.25%	-4.18%
5	HT	666.12	10781.22	22922.10	126027.32
	HM	973.27	10495.14	22527.60	124109.76
	HM-HT	46.11%	-2.65%	-1.72%	-1.52%

Fonte: Elaborado pelo Autor (2020).

Tabela 7 – 10 *Clusters*

		50 PERFIS	100 PERFIS	500 PERFIS
1	HT	7526.24	17991.01	104126.27
	HM	8952.78	18417.84	108428.88
	HM-HT	18.95%	2.37%	4.13%
2	HT	6812.43	19274.81	104260.52
	HM	7783.06	19719.60	114636.69
	HM-HT	14.25%	2.31%	9.95%
3	HT	6453.09	17931.90	106357.77
	HM	8244.72	19131.15	113160.40
	HM-HT	27.76%	6.69%	6.40%
4	HT	7696.21	18712.34	104499.28
	HM	8614.63	18730.15	111235.49
	HM-HT	11.93%	0.10%	6.45%
5	HT	7743.09	18426.28	101719.64
	HM	8608.65	17720.40	108612.23
	HM-HT	11.18%	-3.83%	6.78%

Fonte: Elaborado pelo Autor (2020).

Tabela 8 – 25 *Clusters*

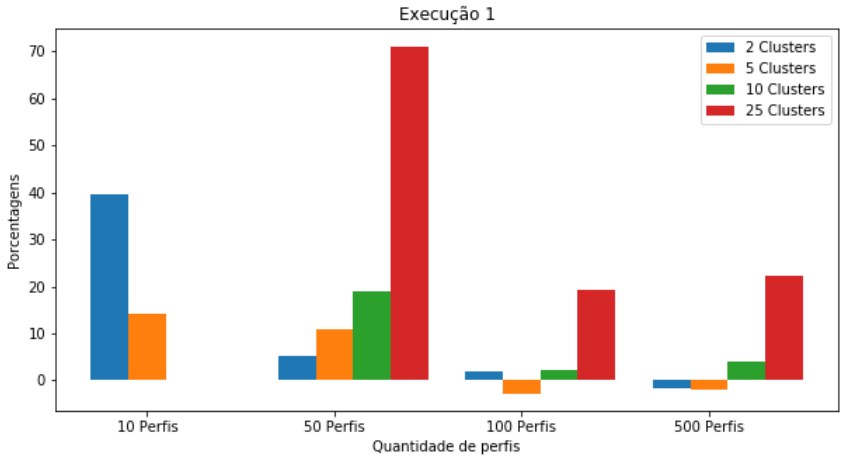
		50 PERFIS	100 PERFIS	500 PERFIS
1	HT	2252.67	9845.94	74155.58
	HM	3853.74	11743.89	90588.08
	HM-HT	71.07%	19.28%	22.16%
2	HT	1924.16	11325.62	77633.74
	HM	3348.96	12774.75	94386.35
	HM-HT	74.05%	12.80%	21.58%
3	HT	2457.44	9731.13	76059.31
	HM	3842.09	13564.37	94336.88
	HM-HT	56.35%	39.39%	24.03%
4	HT	2620.06	9916.25	75613.27
	HM	3962.26	13313.08	91810.04
	HM-HT	51.23%	34.26%	21.42%
5	HT	2710.15	9866.71	70773.14
	HM	3563.13	12096.45	89121.59
	HM-HT	31.47%	22.60%	25.93%

Fonte: Elaborado pelo Autor (2020).

Por fim, para analisar como os modelos apresentados se comportam com diferentes requisições de *clusters* e perfis, para cada uma das 5 execuções foram gerados gráficos de barra com objetivo de obter uma melhor visualização dos dados. Em cada gráfico, as barras se dividem em 4 categorias: 2, 5, 10, 25 *clusters*. O eixo *x* do gráfico representa quantidade

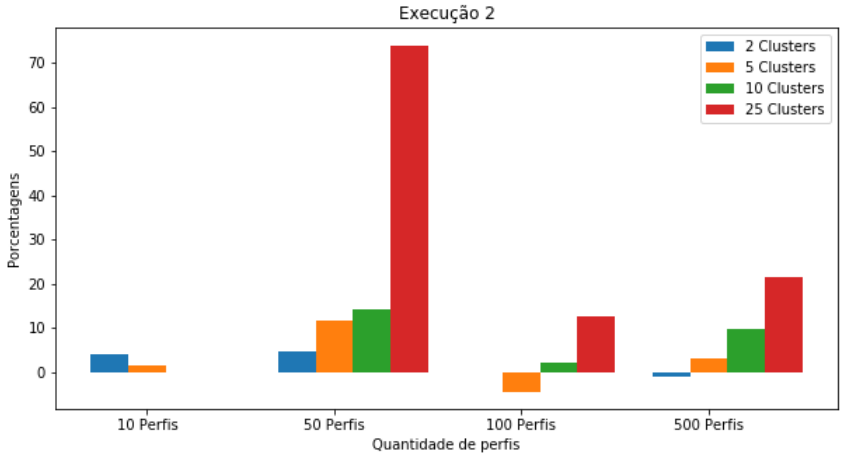
de perfis utilizados na clusterização: 10, 50, 100 e 500 perfis, enquanto o eixo *y*, os valores de $HM - MT$ obtidos pela clusterização. Para cada conjunto de perfis é possível verificar a quantidade de *clusters* que gera melhor resultado, bastando verificar qual barra possui maior valor de $HM - MT$. As Figuras 18, 19, 20, 21 e 22 contém, respectivamente, os gráficos da 1º, 2º, 3º, 4º e 5º execução.

Figura 18 – Valores de HM-HT para todos os elementos e *clusters* da execução 1



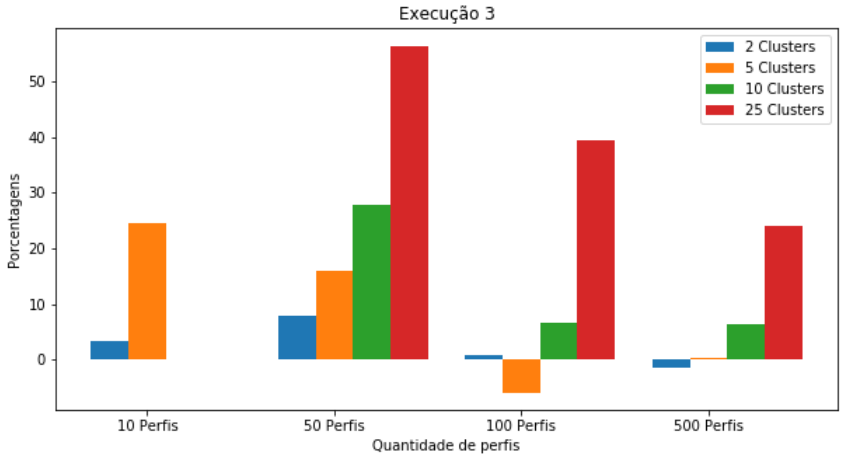
Fonte: Elaborado pelo Autor (2020).

Figura 19 – Valores de HM-HT para todos os elementos e *clusters* da execução 2



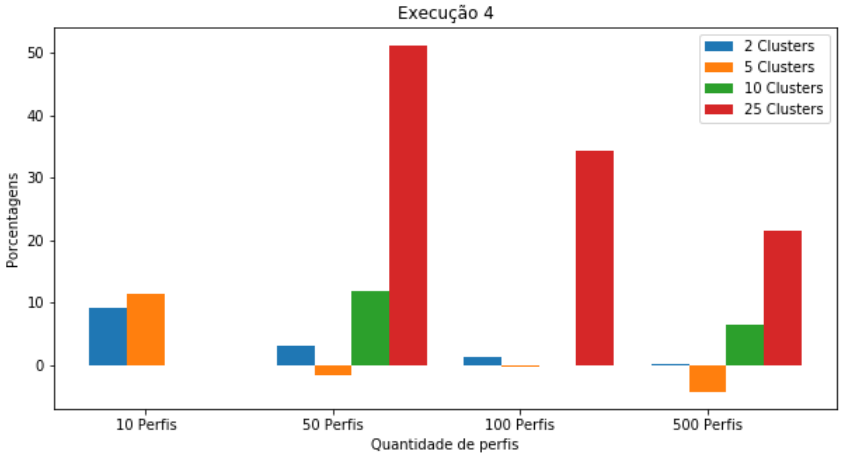
Fonte: Elaborado pelo Autor (2020).

Figura 20 – Valores de HM-HT para todos os elementos e *clusters* da execução 3



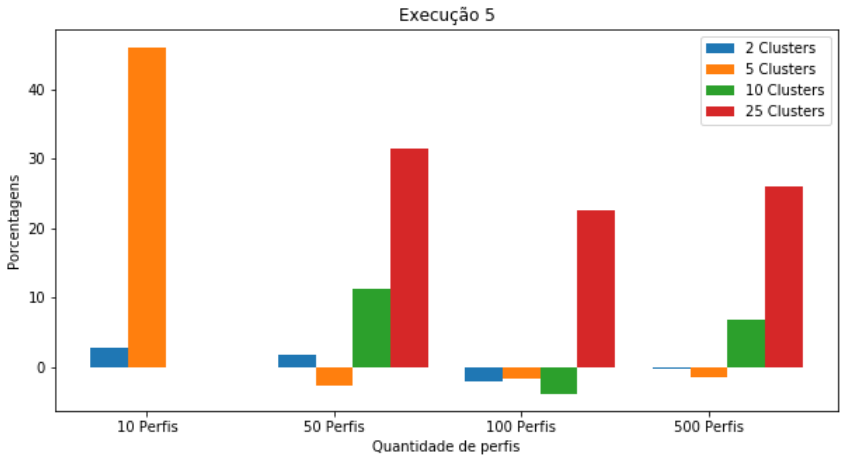
Fonte: Elaborado pelo Autor (2020).

Figura 21 – Valores de HM-HT para todos os elementos e *clusters* da execução 4



Fonte: Elaborado pelo Autor (2020).

Figura 22 – Valores de HM-HT para todos os elementos e *clusters* da execução 5



Fonte: Elaborado pelo Autor (2020).

Com análise dos gráficos é possível verificar que o algoritmo proposto tem bons resultados quando são divididos muitos perfis em poucos *clusters*. Como os casos apresentados para 100 e 500 perfis divididos em 2, 5 e 10 *clusters*. Porém o algoritmo apresentou resultados ruins quando divididos em muitos *clusters*.

Esse resultado ruim pode ser justificado pela própria característica dos algoritmos de Clusterização quando o número de *clusters* é grande. Considerando os dados do trabalho, é muito comum o algoritmo Hierárquico retornar poucos *clusters* com muitos elementos e muitos *clusters* com poucos elementos, muitas vezes até com um único elemento. Como demonstrado na Seção 5.4.5 o cálculo do SSE considera o erro do elemento com base no centróide do *clusters*. Logo, quando um *clusters* só tem um elemento, o valor do SSE será 0, por que ele representa o centróide do grupo. Nesse caso o valor SSE não irá influenciar o resultado final. Como o Hierárquico modificado sempre vai gerar *clusters* com a mesma quantidade de elementos, então ele nunca irá retornar *cluster* de tamanho 1, logo o valor SSE irá influenciar no resultado final.

Essa mesma característica justifica os bons resultados para muitos elementos em poucos *clusters*. Neste caso, dificilmente o algoritmo hierárquico irá retornar *clusters* com um elemento, por que serão classificados muitos elementos, então ele tende a dividir os elementos em poucos *clusters*. O modelo gerado pelo algoritmo modificado tende a ser próximo do modelo gerado pelo algoritmo tradicional.

7 CONCLUSÃO E TRABALHOS FUTUROS

7.1 Conclusões

Este trabalho apresenta uma metodologia para Mineração de Dados da Rede Social Online *Facebook*, assim como um algoritmo para classificação da Força de Conexão entre pares de usuários desta rede. Para tal, o foco deste trabalho foi a Mineração de Textos, utilizada para construir categorias com base em quatro dimensões altamente difundidas na literatura para definição de Forças de Conexão: Intimidade, Serviços, Estrutural e Distância Social. As categorias criadas foram utilizadas como entrada do algoritmo proposto neste trabalho, que através do cálculo da Distância Euclidiana entre os pares de perfis dos usuários, determina a Força de Conexão dos mesmos.

Buscando exemplificar possíveis aplicações para o algoritmo apresentado, o mesmo foi utilizado para determinar se a relação de “amizade” na Rede Social Online *Facebook* é fundamentada pelas categorias de interesses de seus usuários. Os resultados obtidos pelo algoritmo proposto demonstraram que a classificação “Similar” e "Muito Similar"aparece com mais frequência em todas as análises, da mesma forma que a classificação “Muito Dissimilar” não é contabilizada nenhuma vez e a classificação "Dissimilar"aparece apenas em pouquíssimos casos. Desta forma, pode-se afirmar que mais de 90% dos perfis estão conectados a outros perfis similares, ou seja, os interesses do grupo de usuários estudado nesta monografia realmente influenciam as amizade dos mesmos na Rede Social Online *Facebook*.

Outra aplicação exemplificada neste trabalho é a formação de grupos/equipes buscando maximizar a similaridade dos indivíduos envolvidos nestes agrupamentos. Com o objetivo de representar casos de requisição de grupos e equipes em situações empresariais, onde o número de integrantes dos grupos e equipes é pré-definido, foi desenvolvido uma versão modificada do algoritmo Hierárquico. Através da análise dos modelos de clusterização gerados a partir do algoritmo Hierárquico tradicional e do algoritmo Hierárquico modificado, foi possível afirmar que o modelo gerado pelo algoritmo Hierárquico modificado, proposto neste trabalho, apresenta bons resultados, quando se considera o requisito similaridade. Apenas para os casos considerando a formação de 25 *clusters*, ou seja, agrupar os perfis em muitos grupos, o modelo proposto não obteve resultados satisfatórios.

Por fim, é importante ressaltar que os resultados gerados neste trabalho podem ser utilizados de forma geral, ou seja, tanto a metodologia proposta para mineração de dados,

quanto os algoritmos criados podem ser utilizados em qualquer conjunto de dados de usuários da Rede Social Online *Facebook*. Desta forma, os resultados obtidos pelos algoritmos podem ser utilizados para resolução de diversos problemas/questionamentos empresariais, como exemplo, pode-se citar: agrupamentos de usuários que possuam interesse em determinados produtos. Este tipo de agrupamento é de suma importância para equipes de marketing, que podem a partir dos dados obtidos criar propagandas centralizadas. Outro exemplo de aplicação, seria a análise a similaridade entre usuários que ainda não se conhecem na Rede Social Online *Facebook*, para uso em aplicativos de serviços compartilhados.

7.2 Trabalhos Futuros

Em trabalhos futuros espera-se analisar outros algoritmos com o objetivo de determinar a quantidade ideal de *clusters* com base em um conjunto de perfis. Até o momento, os algoritmos apresentados neste trabalho exigem receber, como entrada, a quantidade *cluster* a serem gerados. Um outra possível contribuição, seria a inclusão de novas categorias na base de interesses, com intuito de contemplar uma maior gama de perfis. Além de novas categorias, outras informações como idade, sexo e ocupação podem ser analisadas para definir a Força de Conexão entre os perfis. Neste trabalho, a base de interesse foi construída considerando as reações em publicações, um outro objetivo futuro poderia ser a inclusão da classificação dos comentários existentes em cada publicação.

REFERÊNCIAS

BARBOSA, M. T. S.; BYINGTON, M. R. L.; STRUCHINER, C. J. Modelos dinâmicos e redes sociais: revisão e reflexões a respeito de sua contribuição para o entendimento da epidemia do hiv. **Cadernos de Saúde Pública**, SciELO Public Health, v. 16, p. S37–S51, 2000.

BURT, R. S. **Structural holes: The social structure of competition**. [S.l.]: Harvard university press, 2009.

CASTILHO, D.; MELO, P. O. V. de; QUERCIA, D.; BENEVENUTO, F. Working with friends: Unveiling working affinity features from facebook data. In: **Eighth International AAAI Conference on Weblogs and Social Media**. [S.l.: s.n.], 2014.

GILBERT, E.; KARAHALIOS, K. Predicting tie strength with social media. In: ACM. **Proceedings of the SIGCHI conference on human factors in computing systems**. [S.l.], 2009. p. 211–220.

GOMES, A. K. **Representação, extração e avaliação de interações entre usuários de redes sociais online**. Tese (Doutorado) — Universidade de São Paulo, 2013.

GRANOVETTER, M. S. The strength of weak ties. In: **Social networks**. [S.l.]: Elsevier, 1977. p. 347–367.

HAN, J.; PEI, J.; KAMBER, M. **Data mining: concepts and techniques**. [S.l.]: Elsevier, 2011.

LAZER, D.; PENTLAND, A.; ADAMIC, L.; ARAL, S.; BARABÁSI, A.-L.; BREWER, D.; CHRISTAKIS, N.; CONTRACTOR, N.; FOWLER, J.; GUTMANN, M. *et al.* Computational social science. **Science**, American Association for the Advancement of Science, v. 323, n. 5915, p. 721–723, 2009.

LIN, N.; ENSEL, W. M.; VAUGHN, J. C. Social resources and strength of ties: Structural factors in occupational status attainment. **American sociological review**, JSTOR, p. 393–405, 1981.

METZ, J.; MONARD, M. C. Clustering hierárquico: uma metodologia para auxiliar na interpretaç ao dos clusters. In: **XXIII Congresso da Sociedade Brasileira de Computação**. [S.l.: s.n.], 2005. v. 3, p. 347–395.

MISLOVE, A.; MARCON, M.; GUMMADI, K. P.; DRUSCHEL, P.; BHATTACHARJEE, B. Measurement and analysis of online social networks. In: ACM. **Proceedings of the 7th ACM SIGCOMM conference on Internet measurement**. [S.l.], 2007. p. 29–42.

MORENO, J. L. **Quem sobreviverá?: fundamentos da sociometria, psicoterapia de grupo e sociodrama....** [S.l.]: Dimensão Editora, 1992.

MOURA, M. Proposta de utilização de mineração de textos para seleção, classificação e qualificação de documentos. **Embrapa Informática Agropecuária-Documentos (INFOTECA-E)**, Campinas: Embrapa Informática Agropecuária, 2004., 2004.

RIVO, E.; FUENTE, J. de la; RIVO, Á.; GARCÍA-FONTÁN, E.; CAÑIZARES, M.-Á.; GIL, P. Cross-industry standard process for data mining is applicable to the lung cancer surgery domain, improving decision making as well as knowledge and quality management. **Clinical and Translational Oncology**, Springer, v. 14, n. 1, p. 73–79, 2012.

WELLMAN, B.; WORTLEY, S. Different strokes from different folks: Community ties and social support. **American journal of Sociology**, University of Chicago Press, v. 96, n. 3, p. 558–588, 1990.

WIRTH, R.; HIPP, J. Crisp-dm: Towards a standard process model for data mining. In: SPRINGER-VERLAG LONDON, UK. **Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining**. [S.l.], 2000. p. 29–39.

XAVIER, O. S. A sociometria na administração de recursos humanos. **Revista de Administração de Empresas**, SciELO Brasil, v. 30, n. 1, p. 45–54, 1990.

XIANG, R.; NEVILLE, J.; ROGATI, M. Modeling relationship strength in online social networks. In: ACM. **Proceedings of the 19th international conference on World wide web**. [S.l.], 2010. p. 981–990.

APÊNDICE A – TERMOS DAS CATEGORIAS

Política: ’apoia’, ’bolsominion’, ’bolsonaro’, ’brasil’, ’brasileira’, ’brasileiro’, ’brasileiros’, ’brasilia’, ’camara’, ’camilo’, ’comunista’, ’comunistas’, ’democracia’, ’denuncia’, ’denuncias’, ’direita’, ’ditadura’, ’educaao’, ’federal’, ’governador’, ’governadores’, ’governo’, ’governos’, ’lula’, ’ministerio’, ’ministro’, ’ministros’, ’municipais’, ’municipal’, ’municipio’, ’municipios’, ’nacionais’, ’nacional’, ’policia’, ’policiais’, ’politica’, ’politicos’, ’prefeito’, ’prefeitos’, ’prefeitura’, ’prefeituras’, ’presidente’, ’presidentes’, ’psol’, ’santana’,’utilidade’, ’utilidades’, ’voto’, ’votos’, ’votou’.

Religião: ’abenoado’, ’abenoando’, ’cristaos’, ’cristo’, ’espirito’, ’evangelico’, ’evangelicos’, ’fe’, ’gloria’, ’igreja’, ’igrejas’, ’jesus’, ’jesus’, ’louvor’, ’louvores’, ’misericordia’, ’missa’, ’ora’, ’oraao’, ’oraoes’, ’orar’, ’ore’, ’papa’, ’pascoa’, ’pastor’, ’pastores’, ’santo’, ’santos’, ’senhor’.

Estudo: ’6semestre’, ’aluno’, ’alunos’, ’aula’, ’aulas’, ’ciencia’, ’ciencias’, ’enem’, ’escola’, ’escolas’, ’estudante’, ’estudantes’, ’estudar’, ’faculdade’, ’faculdades’, ’fafidam’, ’ifce’, ’pesquisa’, ’pesquisas’, ’professor’, ’professores’, ’semestre’, ’semestre’, ’turma’, ’turma’, ’ufc’, ’ufc’, ’universidade’, ’universidade’.

Trabalho: ’emprego’, ’empregos’, ’enfermeira’, ’enfermeiras’, ’engenhariaeletrica’, ’professor’, ’professores’, ’trabalhador’, ’trabalhadores’, ’trabalhamos’, ’trabalhando’, ’trabalhar’, ’trabalho’.

Doutrina: ’aborto’, ’abortos’, ’feminista’, ’feministas’, ’machista’, ’machistas’, ’negro’, ’negros’, ’racismo’, ’racista’, ’racistas’.

Relacionamento: ’amiga’, ’amigas’, ’amigo’, ’amigos’, ’amizade’, ’amizades’, ’avo’,’avos’, ’bebe’, ’bebes’, ’casal’, ’casamento’, ’casamentos’,’casar’, ’criana’, ’crianas’, ’esposa’, ’esposas’, ’familia’,’familias’, ’filha’, ’filhas’, ’filho’, ’filhos’, ’gravida’, ’gravidas’, ’irma’, ’irmao’, ’irmaos’, ’irmas’, ’mamae’, ’mamaes’, ’marido’, ’maridos’, ’namora’, ’namorada’, ’namoradas’, ’namorado’, ’namorados’, ’namorar’, ’noiva’, ’noivas’, ’pai’, ’pais’, ’papai’, ’prima’, ’primas’, ’primo’, ’primos’, ’tia’, ’tias’, ’tio’, ’tios’, ’vovo’, ’vovos’.

Futebol: ’atleta’, ’atletas’, ’bola’, ’bolas’, ’corinthians’, ’flamengo’, ’futebol’, ’futebols’, ’gol’, ’gols’, ’hexa’, ’neymar’, ’ronaldinho’, ’ronaldo’, ’rumo’, ’seleao’, ’vozao’.

Esporte: ’campeonato’, ’campeonatos’, ’game’, ’games’, ’jiujitsu’, ’joga’, ’jogando’, ’jogar’, ’jogos’, ’treino’, ’treinos’.

Artes: ’arte’, ’artes’, ’artista’, ’artistas’, ’cultura’, ’culturas’, ’desenho’, ’desenhos’,

’imagem’, ’imagens’, ’livro’, ’livros’, ’tattoo’, ’tatuagem’, ’tatuagens’.

Lúdico: ’anime’, ’animes’, ’assistir’, ’batman’, ’bbb’, ’capitao’,’cinema’, ’cinemas’, ’disney’, ’episodio’, ’episodios’, ’filme’, ’filmes’, ’harry’, ’lol’, ’meme’, ’memes’, ’naruto’, ’netflix’, ’otaku’, ’personagem’, ’personagens’, ’potter’, ’serie’, ’series’, ’thrones’, ’trailer’.

Signo: ’aquario’, ’aries’, ’cancer’, ’capricornio’, ’escorpiao’, ’gemeos’, ’leao’, ’libra’, ’peixes’, ’sagitario’, ’signo’, ’touro’, ’virgem’.

Música: ’album’, ’albuns’, ’banda’, ’bandas’, ’baterista’, ’bateristas’, ’cantando’, ’carnaval’, ’festa’, ’festas’, ’forro’, ’funk’, ’hip’, ’hop’, ’jazz’, ’metal’, ’music’, ’musica’, ’musical’, ’musicas’, ’orquestra’, ’orquestras’, ’rock’, ’show’, ’tocando’, ’trilha’, ’trilhas’.

Comida: ’almoo’, ’arroz’, ’bacon’, ’batata’, ’batatas’, ’bolo’, ’bolos’, ’cafe’, ’chocolate’, ’chocolates’, ’churrasco’, ’churrascos’, ’coca’, ’comida’, ’comidas’, ’cuscuz’, ’delicia’, ’delicias’, ’feijao’, ’janta’, ’jantar’, ’ovo’, ’ovos’, ’paes’, ’pao’, ’pizza’, ’pizzas’, ’receita’, ’receitas’.

Animal: ’animais’, ’animal’, ’cachorro’, ’cachorros’, ’cadelinha’, ’cadelinhas’, ’cat’, ’cats’, ’dog’, ’femea’, ’femeas’, ’filhote’, ’filhotes’, ’gatinho’, ’gatinhos’, ’gato’, ’gatos’, ’pet’.

Maquiagem: ’adrielimakeup’, ’amomake’, ’amomaquiagem’, ’automaquiagem’, ’batom’, ’batons’, ’loucaspormaquiagem’, ’make’, ’makes’, ’makeup’, ’maquiador’, ’maquiadora’, ’maquiadores’, ’maquiagem’, ’maquiageminsta’, ’maquiagemlovers’, ’maquiagemoficial’, ’maquiagemprofissional’, ’maquiagemx’, ’maquiagens’, ’universomakeup’.

Droga: ’bebado’, ’bebados’, ’beber’, ’droga’, ’drogas’, ’maconha’, ’maconhas’, ’maconheiro’, ’maconheiros’.

Pandemia: ’auxilio’, ’corona’, ’coronavirus’, ’covid’, ’covid19’, ’hospitais’, ’isolamento’, ’lockdown’, ’mascara’, ’pandemia’, ’quarentena’, ’upa’, ’virus’.

Tecnologia: ’aplicativo’, ’aplicativos’, ’app’, ’apps’, ’celular’, ’celulares’, ’google’, ’internet’, ’link’, ’links’, ’spotify’, ’tecnologia’, ’tecnologias’.

Facebook: ’canais’, ’canal’, ’comentario’, ’comentarios’, ’compartilha’, ’compartilhar’, ’compartilhe’, ’curte’, ’curtir’, ’direct’, ’face’, ’facebook’, ’fake’, ’insta’, ’instagram’, ’online’, ’perfil’, ’perfis’, ’post’, ’posta’, ’postagem’, ’postagens’, ’postando’, ’postar’, ’postei’, ’posts’, ’publicaao’, ’publicaoes’, ’publicar’, ’sociais’, ’social’, ’status’, ’tbt’, ’tiktok’, ’tumblr’, ’twitter’, ’whatsapp’, ’wpp’, ’youtube’, ’zap’.