Pontifícia Universidade Católica de Goiás

Programa de Mestrado em Engenharia de Produção e Sistemas

ANÁLISE DA REDE SOCIAL TOCANTINS DIGITAL, UTILIZANDO O ALGORITMO *k*-MÉDIAS E CENTRALIDADE DE INTERMEDIAÇÃO

Carolina Palma Pimenta Furlan

2014

ANÁLISE DA REDE SOCIAL TOCANTINS DIGITAL, UTILIZANDO O ALGORITMO *k*-MÉDIAS E CENTRALIDADE DE INTERMEDIAÇÃO.

Carolina Palma Pimenta Furlan

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Engenharia de Produção e Sistemas da Pontifícia Universidade Católica de Goiás, como parte dos requisitos para obtenção do título de Mestre em Engenharia de Produção e Sistemas.

Orientador: Clarimar José Coelho, Dr.

Goiânia

Setembro de 2014

ANÁLISE DA REDE SOCIAL TOCANTINS DIGITAL, UTILIZANDO O ALGORITMO *k*-MÉDIAS E CENTRALIDADE DE INTERMEDIAÇÃO.

CAROLINA PALMA PIMENTA FURLAN

Esta Dissertação julgada adequada para obtenção do título de Mestre em Engenharia de Produção e Sistemas, e aprovada em sua forma final pelo Programa de Pós-Graduação em Engenharia de Produção e Sistemas da Pontifícia Universidade Católica de Goiás em Setembro de 2014.

_____

Prof. Ricardo Luiz Machado, Dr.
Coordenador do Programa de Pós-Graduação em Engenharia de Produção e Sistemas

Banca Examinadora:

_____

Prof. Clarimar José Coelho, Dr.
Orientador

_____

Prof. Sibélius Lellis Vieira, Dr.

_____

Prof. Marcelo Lisboa Rocha, Dr.

Goiânia-Goiás
Setembro de 2014

Dados Internacionais de Catalogação da Publicação (CIP)
(Sistema de Bibliotecas PUC Goiás)

Furlan, Carolina Palma Pimenta.
F985a Análise da Rede Social Tocantins Digital, utilizando o Algoritmo *k*-médias e centralidade de intermediação [manuscrito] / Carolina Palma Pimenta Furlan. – Goiânia, 2014.
90 f. ; il. ; 30 cm.

Dissertação (mestrado) – Pontifícia Universidade Católica de Goiás, Programa de Mestrado em Engenharia de Produção e Sistemas, 2014.
"Orientador: Prof. Dr. Clarimar José Coelho".
Referências Bibliográficas: p. 21-42.

1. Redes sociais on-line. I. Coelho, Clarimar José. II. Pontifícia Universidade Católica de Goiás, MEPROS, Programa de Mestrado em Engenharia de Produção e Sistemas. III. Título.

CDU 316.77:004.738.5(043)

AGRADECIMENTOS

Primeiramente agradeço à Deus por avisar que me daria a força no início deste trabalho e por me acompanhar em todos os momentos.

Aos meus pais por me darem a formação que recebi. À minha mãe por cuidar tão bem da minha filha Beatriz quando eu precisei viajar para Goiânia para assistir às aulas do mestrado.

Ao meu esposo pela paciência, pelo companheirismo, por nunca me deixar desistir e cujo amor foi fundamental para este desafio!

Aos meus sogros por me ajudarem quando eu precisava estudar!

Ao meu orientador Prof. Dr. Clarimar por ter me aceitado como sua aluna de mestrado, pela orientação, disponibilidade e paciência.

À minha amiga Micéia Garrido Lopes sempre disposta a me ajudar nos momentos de apuro.

À Fundação de Amparo à Pesquisa do Estado do Tocantins por conceder um ano de bolsa de estudos deste mestrado.

A todos os professores do programa de Mestrado em Engenharia de Produção e Sistemas da PUC de Goiás pelos ensinamentos e aos membros das bancas de qualificação e defesa.

Deus obrigada por colocar essas pessoas tão especiais na minha vida!

Resumo da Dissertação apresentada ao MEPROS/ PUC Goiás como parte dos requisitos necessários para a obtenção do grau de Mestre em Engenharia de Produção e Sistemas (M.Sc.).

ANÁLISE DA REDE SOCIAL TOCANTINS DIGITAL, UTILIZANDO O ALGORITMO k-MÉDIAS E CENTRALIDADE DE INTERMEDIAÇÃO

Carolina Palma Pimenta Furlan

Setembro de 2014

Orientador: Prof. Clarimar José Coelho, Dr.

O advento da Internet possibilitou vários meios de comunicação entre as pessoas e dentre elas destacam-se as plataformas de redes sociais, tornando-se uma nova forma de relacionamento. Explorar esse novo ambiente tornou-se cada vez mais apreciado para os pesquisadores, tanto quanto para os gestores de um modo em geral. Esse ambiente proporciona a formação de comunidades e através delas é possível a identificação de formação de grupos através dos seus interesses. A maioria dos algoritmos de visualização de redes sociais são representados em grafos. O ambiente dessa pesquisa consiste no subgrupo Tocantins digital da rede social Facebook, o qual possui mais de 10.000 membros. Neste ambiente utilizou-se a ferramenta API, na qual foi possível desenvolver aplicações para se coletar informações das postagens dos membros do grupo pesquisado. Neste trabalho foi aplicado o algoritmo *k*-médias para o agrupamento de dados representados pelos cinco grupos como a melhor solução encontrada, e também a medida centralidade de intermediação onde revelou a existência de três membros com maior influência em suas postagens, e quatro produtos ou serviços mais visualizados dentro do subgrupo.

PALAVRAS-CHAVE: centralidade de intermediação, *k*-médias, Facebook, redes sociais.

Summary of Dissertation submitted to MEPROS/PUC Goiás as part of the requirements for the degree of Master of Engineering in Production Systems (M. Sc.)

SOCIAL NETWORK ANALYSIS TOCANTINS DIGITAL, USING THE K-MEANS ALGORITHM AND BETWEENNESS CENTRALITY.

Carolina Palma Pimenta Furlan

September 2014

Advisor: Prof. Clarimar José Coelho, Doctor.

The advent of Internet enabled various means of communication among people and among them there are the social networking platforms, becoming a new form of relationship. Explore this new environment has become increasingly appreciated for researchers as well as for managers in a way in general. This environment provided the formation of communities and through them it is possible to identify the formation of groups through their interests. Most visualization algorithms for social networks are represented in graphs. The environment of this research consists in the subgroup Tocantins digital social of network Facebook, which has more than 10.000 members. In this paper the k-means algorithm for clustering of data represented by the five groups as the best solution found, and also the centrality measure of intermediation which revealed the existence of three most influential members in your posts, and four products or services was applied most Viewed within the subgroup.

KEYWORDS: betweenness centrality, k-means, Facebook, social network.

SUMÁRIO

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

API: *Application Programming Interface* (Interface de Programação de Aplicativos)

CSV: *Comma-separated Values* (Valores separados por vírgula)

DM: *Data Mining* (Mineração de Dados)

FQL: *Facebook Query Language* (Linguagem de Consulta do Facebook)

FSSP: *Flow Shop Scheduling Problem* (Problema da Programação do Fluxo de Loja)

HTML: *Hypertext Markup Language* (Linguagem de Marcação de Hipertexto)

HTTP: *Hypertext Transfer Protocol* (Protocolo de Transferência de Hipertexto)

JSON: *Java Script Object Notation* (Notação de Objeto de Java Script)

JUNG: *Java Universal Network/Graph Framework* (Rede Universal Java)

KDD: *Knowledge Discovery in Databases* (Descoberta de Conhecimento em Bases de Dados)

PHP: *Personal Home Page* (Página Pessoal)

SDK: *Software Development Kit* (Kit de Desenvolvimento de Software)

SNA: *Social Network Analysis* (Análise de Redes Sociais)

SQL: *Structured Query Language* (Linguagem de Consulta Estruturada)

XML: *eXtensible Markup Language* (Linguagem de Marcação)

WEKA: *Waikato Environment for Knowledge Analysis* (Ambiente para a análise do conhecimento da Universidade de Waikato)

1.INTRODUÇÃO

Este trabalho apresenta um software desenvolvido para a identificar agrupamentos de produtos e serviços similares que são mais postados, destacar quais membros possuem maior influência nas postagens de oferta de prestação de serviços ou produtos e apresentar quais produtos e serviços são mais visualizados do subgrupo do Facebook denominado Tocantins digital utilizando para isto, o algoritmo *k*-médias (MACQUEEN, 1967; BISHOP, 1995; JAIN *et al*, 2000; TAN; STEINBACH; KUMAR, 2009) e a medida de centralidade de intermediação (EVERIT, 1974; FREEMAN, 1977; NEWMAN, 2004; MARTELETO; SILVA, 2004; MIKA, 2011).

O subgrupo Tocantins digital é um espaço privado formado por membros com interesses comuns na rede social *online* Facebook. É um ambiente de postagens de classificados onde seus membros podem publicar ofertas, por meio dos *posts*, e também recebê-las de outros membros. Neste contexto, o subgrupo Tocantins digital possui todas as características do ambiente Facebook. O Facebook disponibiliza para programadores e desenvolvedores uma Interface de Programação de Aplicativos (*Application Programming Interface,* API), que possibilita o desenvolvimento de aplicações, que podem ser integradas ao ambiente Facebook sem o envolvimento com os detalhes dos programas originais do ambiente Facebook (KIRKPATRICK, 2007, PATRÍCIO, 2009). No caso desse trabalho, as novas aplicações são integradas ao Tocantins digital. Essa API oferece recursos para a coleta de informações provenientes da plataforma Facebook que são publicadas pelos próprios membros do Tocantins digital (FACEBOOK, 2013).

Em ambientes da rede mundial (*World Wide Web, Web*) como o Facebook os dados são desestruturados (LOPES; HIRATANI, 2008). As atividades de criação de

novas aplicações envolvem a criação de um novo ambiente organizado em uma nova estrutura para o armazenamento dos dados extraídos do Facebook. Após a extração dos dados, inicia-se a tarefa de agrupamento que pode ser realizado por diversos algoritmos de agrupamento aplicados na área de mineração de dados (MITCHELL, 1997).

Os dados usados pelo software desenvolvido nesse trabalho são provenientes das tabelas com informações sobre as postagens e perfis dos usuários que fazem parte do Tocantins Digital, atualmente com mais de 10000 membros. Para a extração dos dados basta que o usuário tenha perfil de administrador

A estrutura de dados do Facebook é organizada em um grafo que representa os relacionamentos entre um conjunto de entidades (BARABASI, 2003; BROOKSHEAR, 2005; TAN; STEINBACH; KUMAR, 2009). Um grafo é uma estrutura que contém vértices e arestas (FOROUZAN; MOSHARRAF, 2011; MOKARZEL; SOMA, 2008). Nesse trabalho, os vértices do grafo representam as classes de produtos e serviços e arestas representa o fluxo entre as classes.

O subgrupo Tocantins digital (https://facebook.com/groups/tocantinsdigital/) foi escolhido para uso neste trabalho por se caracterizar em um ambiente que proporciona um canal de comunicação entre seus membros por meio de interesses em comum. Atualmente, o Facebook não dispõe de métodos de análise para identificar agrupamentos e medir a influência de um vértice, isto é, a influência de uma postagem em um dado subgrupo como o Tocantins digital. A plataforma disponibiliza apenas ferramentas para manuseio e extração de grandes conjuntos de dados.

O algoritmo *k*-médias separa os objetos em grupos segundo uma medida de distância ou similaridade entre eles (MACQUEEN, 1967). A tarefa de agrupamento (*clustering*) de dados resume-se em utilizar a informação encontrada nos dados e seus

relacionamentos para dividir os dados em grupos, que consistem em elementos de um grupo relacionados entre si, mas não relacionados aos elementos de outros grupos (FERREIRA, 2012), bem como, possibilita a observação, registro e análise do comportamento do usuário e de suas reações (BARABÁSI, 2002).

A motivação para a escolha do algoritmo *k*-médias para o agrupamento de classes de produtos e serviços é devida a representação dos dados no formato de um grafo onde as classes mantêm interação e possibilitam a utilização de uma técnica de agrupamento que encontra grupos de tamanhos variados. O *k*-médias é um algoritmo muito conhecido que vem sendo aplicado com sucesso nos problemas práticos de agrupamento de dados (DUBES; JAIN, 1980; COELHO FILHO *et al*, 2013).

Já a centralidade de intermediação é uma medida da influência, em decorrência do fluxo de movimentação de dados, entre os vértices de um grafo, cujo cálculo envolve a quantidade de menores caminhos, dentre todos existentes, que passam em um determinado vértice (FREEMAN, 1977). O problema do caminho mínimo consiste na minimização do custo de travessia de um grafo entre dois vértices. O custo é dado pela soma dos pesos de cada aresta percorrida (FOROUZAN; MOSHARRAF, 2011; MOKARZEL; SOMA, 2008). Por exemplo, a interação entre dois vértices, A e B, podem depender de outro vértice C localizado no caminho da comunicação entre A e B. A informação, para ser compartilhada entre os vértices A e B precisa, na maioria das vezes, passar pelo vértice C. Portanto, o vértice C está presente na maioria dos menores caminhos existentes no grafo (VIEIRA, 2011) e possui uma alta centralidade de intermediação, por ter uma influência considerável, em virtude de seu controle sobre a informação que passa entre os outros vértices (NEWMAN, 2010).

Na análise do Tocantins digital, torna-se necessário observar que todos os pares de vértices trocam informações. Estes pares estão realmente conectados por um caminho onde as informações sempre tomam um caminho mais curto (distância geodésica). Para calcular a centralidade de intermediação, para cada par de vértices do grafo, calcula-se quantas vezes um determinado vértice aparece no caminho mais curto daquele par. O vértice que aparecer o maior número de vezes nos caminhos geodésicos de cada par do grafo possui o maior valor de centralidade de intermediação (NEWMAN, 2010). Os vértices com maior valor de centralidade podem ser cruciais por conectarem diferentes regiões da rede, e atuarem como pontes, fazendo parte de um maior número de menores caminhos no grafo (FREEMAN, 1977; NEWMAN 2001). As arestas são elementos importantes para diminuir a média de comprimento dos caminhos entre os vértices, para acelerar a difusão de informações e para aumentar o tamanho de parte do grafo em uma dada distância de um vértice. Contudo, em grafos com muitas pontes são mais frágeis e menos agrupadas (BARRAT *et al*, 2008).

Os processos dentro de uma rede social são considerados dinâmicos, devido a essa rede sofrer constantes alterações através das relações entre as pessoas, grupos ou organizações (STRÖELE; ZIMBRÃO; SOUZA, 2012). A relevância desse estudo se dá não apenas pelo fato que a rede social Facebook ser atualmente um ambiente que promove a interação entre as pessoas, mas também por ser considerada a rede social mais acessada no Brasil (EMARKETER, 2013).

A motivação para a escolha desse tema resulta dos interesses em se tratar com uma grande base de dados *online* e posteriormente realizar uma análise de dados, aplicando-se um método de agrupamento e um instrumento que mede a influência do fluxo de movimentação de informações. Devido ao grande volume de informações, existem problemas de como combinar e visualizar a relação entre a falta de estrutura

dessas informações. Com os algoritmos utilizados neste trabalho, é possível indexar as informações solicitadas com maior precisão e organizar esses dados onde estiverem escritos, de tal maneira que seja possível a manipulação mediada por computador. Trata-se de obter uma forma possível de extrair e integrar informações tanto dos usuários quanto das informações que fluem no subgrupo da rede social *online*.

Com o resultado deste trabalho, gestores da área de marketing podem verificar, através deste cenário, o comportamento dos membros da rede social. Portanto, estes gestores podem fazer uso de maneira desejada para se encontrar o cliente que se almeja, definindo uma publicidade adequada de determinados tipos de produtos e serviços que podem ser oferecidos dentro da rede social pesquisada e evitando, em certos momentos, a propaganda maçante àquelas pessoas que não têm interesse nesses mesmos produtos e serviços.

Empresas de marketing e anunciantes têm aumentado seus orçamentos para anúncios e campanhas em redes sociais, através de mecanismos que regem o funcionamento da difusão de informações nas redes sociais e da influência que ocorre entre os usuários das redes. Aplica-se a ferramentas de análise de dados nos fóruns de mensagens utilizando um modelo de difusão de influência. Dado um grafo representando uma rede social, cada modelo simula como uma ideia ou informação se propaga na rede (CAVALCANTI *et al,* 2012).

Os estudos sobre agrupamento em redes sociais *online,* de acordo com o modelo de Watts e Strogatz (WATTS; STROGATZ, 1998), identificam grupos na rede social Orkut, cujos membros participam de outros grupos (RECUERO, 2004). No trabalho de Cheng *et al* (2008), foram apresentadas características da rede social YouTube

quanto ao compartilhamento de vídeos, aplicando-se a técnica de agrupamento (*clustering*).

Numa abordagem feita no provedor Universo *OnLine*, Benevenuto *et al* (2009) utilizaram o algoritmo X-means que é um algoritmo eficiente que estende o popular *k*-médias para estudar perfis de usuários que acessam servidores de compartilhamento de vídeos. Como resultado, encontraram quinze grupos distintos como melhor escolha no estudo de caso, onde foi possível destacar o ranking de requisições de usuários e a distribuição de tempo entre chegadas das requisições.

Em uma rede social científica, ambiente onde pesquisadores possuem o mesmo objetivo de pesquisa, a abordagem do *k*-médias resultou na identificação de comunidades de pesquisas, pesquisadores que detêm maior influência e compreender a evolução social dos pesquisadores ao longo do tempo (STRÖELE; ZIMBRÃO; SOUZA, 2012). Através de uma comparação entre métodos de segmentação, Quinteiro (2011), utilizou o método do *k*-médias para identificar grupos de pessoas que gostam de música, aplicando-se um questionário aos membros de um grupo da rede social Facebook. Nesta mesma rede, foi possível validar o indicador de vulnerabilidade de usuários, alcançando melhor resultado de um número de *clusters* com o uso do *k*-médias no grupo pesquisado (HANNE *et al*, 2012). No trabalho de Santos (2012), este algoritmo foi aplicado na identificação de agrupamentos de acordo com o perfil das pessoas que possuem preferência por um determinado tipo de assunto de uma variedade de notícias.

No trabalho de Botelho e Sousa (2011), os autores apresentaram uma revisão bibliográfica sobre o algoritmo *k*-médias e a centralidade de intermediação na detecção de comunidades em grafos. Em uma comunidade de professores, o *k*-médias agrupou equipes semelhantes e a centralidade de intermediação mostrou que os valores de

centralidade são baixos entre os professores, pois os mesmos possuem conexão direta uns com os outros não dependendo de terceiros. Através dos resultados foi possível obter informações sobre o desempenho das equipes e dos gestores (LIN; CHEN; TSAI, 2003).

No programa da Rede Europeia de Observação sobre a Coesão e Desenvolvimento Territorial (European Spatial Planning Observation Network, ESPON), rede científica composta por universidades, centros de pesquisas, consultorias e órgãos públicos, considerados beneficiários desta rede científica, Madeira (2010) teve como resultado seis agrupamentos de beneficiários semelhantes, utilizando o *k*-médias. Em seguida, a centralidade de intermediação mediu o posicionamento de cada beneficiário para determinar padrões de colaboração.

O trabalho de Rodrigues (2009) no ambiente de um sistema de correio eletrônico, o algoritmo *k*-médias foi aplicado para identificar grupos de professores que utilizam regularmente este sistema. Já a centralidade de intermediação assumiu um papel importante para determinar em que ponto se obteve a divisão ótima da rede deste grupo de professores.

A centralidade de intermediação é uma dentre as métricas de Análise de Redes Sociais (*Social Network Analysis*, SNA), onde no trabalho de Albuquerque *et al* (2013), foi aplicada para se mapear a relação entre os usuários da rede social *Twitter*. Em 2011, Barbosa *et al* (2011), apresentaram um estudo de medidas de centralidade e detecção de comunidades a partir de dados da Plataforma Lattes do CNPq, de forma a determinar os vértices centrais da rede. Neste trabalho, a medida de centralidade é aplicada para quantificar a importância de cada vértice dentro do grafo do Tocantins digital.

No trabalho de Gulini e Misaghi (2012), a centralidade de intermediação, teve como propósito avaliar o comportamento dos participantes de uma rede social online, onde pode-se visualizar qual membro do grupo possui maior prestígio tanto no sentido de buscá-lo como referência de suas postagens, quanto no aspecto de suas respostas aos demais. Para Smolka (2006), uma análise por meio da centralidade serviu para medir a relação de cooperação entre empresas de base tecnológica, ligações com universidades e centros de pesquisa para o desenvolvimento de produto, baseada nos resultados de entrevistas realizadas em redes sociais inter-empresariais.

O estudo de caso proposto neste trabalho tem como base metodológica o processo Descoberta de Conhecimento em Banco de Dados (*Knowledge Discovery in Databases*, KDD), que é o processo de conversão de dados brutos em informações úteis, combinando métodos tradicionais de análise de dados com algoritmos sofisticados para processar grandes volumes de dados (BRAGA, 2005; TAN; STEINBACH; KUMAR, 2009). Portanto, neste trabalho, envolveram as etapas descritas a seguir: na primeira etapa foram realizadas pesquisas bibliográficas acerca dos algoritmos a serem aplicados na mineração dos dados e posteriormente das ferramentas oferecidas pelo subgrupo Tocantins Digital para a extração dos dados. Na segunda etapa, foram extraídos os dados experimentais, por meio da ferramenta API do Facebook, que disponibiliza as tabelas de dados do subgrupo Tocantins digital para consulta aos dados por meio da Linguagem de Consulta do Facebook (*Facebook Query Language*, FQL). Dessas tabelas, foram extraídas as informações relacionadas às postagens deste subgrupo e posteriormente armazenados em um banco de dados da Linguagem de Consulta Estruturada (*Structured Query Language*, SQL). Na sequência, os valores utilizados como parâmetros foram analisados para se encontrar os grupos, através da Interface de Programação de Aplicativos (*Application Programming Interface*, API), *Simple k-means,* que consiste

num algoritmo do Ambiente para a Análise do Conhecimento da Universidade de Waikato (*Waikato Environment for Knowledge Analysis*, WEKA), no qual faz o uso do *k*-médias para separar os grupos (*clusters*) de classes de produtos e serviços, de acordo com postagens realizadas pelos membros do subgrupo pesquisado. O WEKA, que está na versão 3.6, é uma ferramenta de domínio público que pode ser utilizada no KDD, pois apresenta uma coleção de ferramentas de mineração de dados e disponibiliza algoritmos de classificação não-supervisionada (técnicas de *clustering*) e de regras de associação (HAUS, 2012).

Após o processo de identificação dos grupos, outra pesquisa foi apresentada utilizando-se a medida de centralidade de intermediação, a qual utiliza o cálculo do caminho mínimo entre os vértices como parâmetro para se identificar qual ou quais membros do subgrupo possuem maior influência nas postagens e quais produtos/serviços são mais postados. Com um número crescente de pessoas trocando informações e postando algum tipo de oferta de produto ou serviço dentro dos grupos formados nas próprias redes sociais, foi possível perceber um novo ambiente de negócio, que despertou o interesse por uma análise. Contudo, tanto é possível, ao final dessa análise, obter padrões de preferências de produtos e serviços mais postados, como também detectar quais vértices possuem uma maior influência de suas postagens.

O Capítulo 2 abrange uma fundamentação teórica sobre Análise de Redes Sociais, Facebook, Mineração de Dados, Etapas do Processo de Descoberta de Conhecimento (*Knowledge Discovery in Database*, KDD) e Análise de Grupos. O Capítulo 3 apresenta a metodologia, as ferramentas usadas para obter os resultados e discute os resultados da identificação dos agrupamentos no encontro das classes de serviços, produtos e da influência dos membros. Por fim, no Capítulo 4 são apresentadas as considerações finais e trabalhos futuros.

2. REVISÃO BIBLIOGRÁFICA

Neste capítulo são descritos conceitos relativos à Análise de Redes Sociais, Facebook, Processo de Descoberta de Conhecimento (*Knowledge Discovery in Databases*, KDD), algumas das principais tarefas da mineração de dados (*Data Mining*, DM) e análise de grupos.

2.1 ANÁLISE DE REDES SOCIAIS

Uma rede social é constituída por um grupo de pessoas, organizações ou outras entidades sociais, e suas relações socialmente significativas, como amigos, colegas de trabalho ou troca de informações, que podem estar conectadas por redes de computadores (MARTELETO; SILVA, 2004).

A análise em uma rede social (*Social Network Analysis*, SNA) (WASSERMAN; FAUST, 1994) foi abordada por Freeman em 1996, dentro de um contexto sociológico, descrevendo os meios pelos quais a informação pode ser codificada. A SNA pode ser aplicada em várias áreas do conhecimento como gerenciamento de conhecimento, desenvolvimento organizacional, pesquisas em economia e ciências sociais, epidemiologia e segurança contra ataques terroristas, dentre outros (PIMENTEL; FUKS, 2011). A maioria das redes sociais são para fins pessoais e de socialização. No entanto, a análise de rede social leva ao mapeamento, medição e modelagem dos relacionamentos e fluxo entre as pessoas e grupos (AL-FAYOUMI; BANERJEE JR; MAHANTI, 2009).

Conforme Wattenberg (2006), a análise de redes sociais envolve três tarefas fundamentais:

- Identificar comunidades: os atores devem ser agrupados em comunidades, de acordo com os seus atributos. É importante a avaliação da densidade de uma comunidade em termos de conexão e identificar cliques e relacionamentos abertos;
- Identificar atores centrais: é necessária a identificação dos atores que possuem o maior número de conexões, assim como pontos de articulação - atores que formam pontes entre comunidades. Esta tarefa requer a visualização e compreensão global da rede;
- Analisar papel e posição de relacionamentos e indivíduos: essa tarefa requer interpretação da estrutura da rede e depende dos atributos de atores e relacionamentos.

A questão do relacionamento pessoal passou além daquele contato físico para a interação *on-line*. Algumas pessoas sentem mais liberdade em se expressar ou formar um relacionamento por meio do ambiente da Internet. De acordo com Velloso (2011), um site de relacionamento ou rede social é uma estrutura social composta por pessoas ou organizações, conectadas por um ou vários tipos de relações, que partilham valores e objetivos comuns.

Redes sociais na *Web* são ambientes virtuais onde os participantes interagem com outras pessoas e criam redes baseadas em algum tipo de relacionamento. Em um sistema de redes sociais na *Web*, cada membro possui sua própria rede social, o que forma uma teia de relacionamentos (PIMENTEL; FUKS, 2011). Dessa forma, podem operar em diferentes níveis, tais como, redes de relacionamentos (Facebook, Orkut, Myspace,

Twitter), redes profissionais (Linkedin), redes comunitárias (redes sociais em bairros ou cidades), redes políticas, dentre outras (VELLOSO, 2012). Estas redes sociais disponibilizam recursos como troca e visualização de mensagens e arquivos, proporcionando aos seus usuários novas formas de se expressar e de se relacionar.

Os usuários e os programadores das redes sociais passaram a ser codesenvolvedores na criação de novos aplicativos, tornado um ambiente de ideias, e até mesmo de reuniões, onde envolvem uma inteligência coletiva (PISANI; PIOTET, 2010).

As conexões entre participantes crescem em um ritmo acelerado, se tornando um fenômeno o número de pessoas que acessam as informações ou estabelecem algum tipo de relação. Dessa forma, há uma tendência ainda maior no crescimento das redes sociais, pois as pessoas gastam um tempo maior nos sites de relacionamentos.

Para explorar suas aplicações, as redes sociais são representadas por grafos. Sendo assim, uma característica fundamental nas redes sociais é a possibilidade de lidar com dados relacionais, ou seja, dados que expressam relações (conexões ou laços) entre objetos (vértices, indivíduos, grupos) diversos (HANNEMAN, 2000). Segundo Fortunato (2010), uma comunidade pode ser denominada *cluster* ou módulo, na qual é utilizada em diversas aplicações no contexto de redes reais e permitem fazer uma análise estrutural da mesma. Identificar módulos e seus limites permite classificar os vértices que o compõem, desde o ponto de vista da posição estrutural dentro do módulo.

Grande parte dos algoritmos de visualização de redes sociais são baseados em grafos, destacando relacionamentos entre indivíduos e grupos de indivíduos. Como exemplo, o trabalho de Andery (2010) propõe soluções para representar e explorar visualmente redes sociais com o uso de grafos que podem ser implementadas através de projeções multidimensionais.

Marteleto (2001) propôs o uso da Análise de Redes Sociais com o objetivo de perceber os fluxos de informação e as construções sociais e simbólicas dos grupos estudados, de forma a compreender a comunicação como instrumentos de mobilização nos movimentos sociais. Devido a SRA utilizar dados relacionais, como por exemplo, informações como tipos de contatos, vínculos, ligações de sujeitos e grupos, no trabalho de Aguiar (2007), a SRA foi aplicada para procurar padrões estruturados de interação entre indivíduos nos sites de redes sociais, fomentado por motivações comerciais, das articulações e agenciamentos das redes sociais de ONGs e dos movimentos sociais.

No trabalho de Recuero (2008), a SRA em uma análise realizada em fotologs brasileiros, apresentou agrupamentos que podem apresentar interações frequentes no tempo, gerando laços sociais. Mondini *et al* (2012) destacaram em seu trabalho, que a rede social *online* Twiter obteve a maior representatividade devido ser a ferramenta mais utilizada entre os membros de uma Instituição de Ensino Superior de Santa Catarina.

A análise de Redes Sociais tornou-se um recurso que respalda a gestão organizacional, identificando os atores mais influentes na rede, bem como se tornando, cada vez mais, um recurso estratégico na estruturação e criação de ligações importantes (CROSS; PARKER; BORGATTI, 2000).

2.2 O FACEBOOK

Lançada em 4 de fevereiro de 2004, a rede social Facebook foi criada por Mark Zuckerberg, sendo restrita apenas aos estudantes da Universidade de Harvard. Em

27 de fevereiro de 2006, passou a aceitar estudantes secundaristas e algumas empresas, sendo restrito aos usuários com idade abaixo de 13 anos (VELLOSO, 2011).

O Facebook, que no início era uma simples galeria eletrônica de fotos para estudantes universitários, expandiu-se enormemente por volta de 2006 e conheceu um grande sucesso (PISANI; PIOTET, 2010). De acordo com Russel (2011), em 2010 o Facebook foi o site mais visitado, considerando que esta rede social possui mais de 500 milhões de usuários que trocam mensagens, de modo semelhante a uma comunicação por e-mail, participam de *chats* em tempo real, compartilham fotos e demais arquivos, chegando a mobilizar grupos de pessoas em defesa ou crítica de uma causa. Em outubro 2013 (EMARKETER, 2013), de acordo com dados da *Experian Marketing Services*, 73% de todas as visitas realizadas em ambiente de redes sociais no Brasil foram para o Facebook.

Para participar da rede social Facebook, o usuário precisa cadastrar um perfil social onde dados pessoais são informados. Esse perfil é considerado uma conta individual, onde o usuário pode procurar amigos, participar de um ou mais grupos, de forma a ter à sua disposição um comunicador instantâneo, permissão para compartilhar arquivos, fotos, vídeos, entre outros. O Facebook também permite que se crie uma página, semelhante a um site de uma empresa ou organização, onde seu o conteúdo pode ser visto por qualquer pessoa, sem a necessidade que o administrador da página dê permissão para acessá-la.

Outro ambiente dentro do Facebook que pode ser criado é o subgrupo. A formação de um subgrupo se dá pelos usuários com os mesmos interesses, compartilhando opiniões e informações de forma semelhante a um fórum. Possui uma particularidade onde as pessoas podem participar mediante um convite enviado pelo

administrador, de acordo com a configuração do subgrupo. O administrador ainda pode remover publicações abusivas e até mesmo os membros (FACEBOOK, 2013).

Além desses recursos, o Facebook disponibiliza para o público desenvolvedor a sua Interface de Programação de Aplicativos (*Application Programming Interface*, API), então chamada de Plataforma Facebook, o que possibilita, a partir de então, o desenvolvimento de aplicações integradas ao Facebook que poderiam fazer uso das informações publicadas por seus usuários em suas interações na rede (KIRKPATRICK, 2007). Portanto, os usuários e os programadores das redes sociais passaram a ser co-desenvolvedores na criação de novos aplicativos, tornado um ambiente de ideias, e até mesmo de reuniões, onde envolvem uma inteligência coletiva (PISANI; PIOTET, 2010).

2.3 MINERAÇÃO E DESCOBERTA DE CONHECIMENTO DOS DADOS

Antes de dar início ao assunto de extração de dados é importante esclarecer alguns dos principais termos utilizados, que são comumente encontrados no meio *Web*, os quais são: dado, informação, conhecimento e atributo. Davenport e Prusak (1999) conceituam de forma bem clara que dados são um conjunto de fatos distintos, relativos a eventos. Num contexto organizacional, dados são utilitariamente descritos como registros estruturados de transações. De acordo com Velloso (2011), dado é definido ainda como um elemento puro e quantificável que pode ser utilizado em um ambiente operacional.

Os dados podem ser transformados em informação quando é possível adicionar valor aos mesmos (DAVENPORT; PRUSAK, 1998). De acordo com Velloso (2011), a informação é um conjunto estruturado de dados, transmitindo conhecimento. A

informação pode ser definida como dados organizados, em tabelas consultadas, por uma ferramenta específica, diretamente no banco de dados.

A combinação de vários fatores como contexto, interpretação, experiência pessoal, aplicabilidade, e processo cognitivo incrementam a informação, transformando-a em conhecimento (SIQUEIRA, 2005). O conhecimento implica uma compreensão e experiência implícitas que podem significar a diferença entre seu bom ou mau uso. Com o tempo, a informação se acumula e se deteriora, mas o conhecimento evolui com experiência, estabelecendo conexões com novas situações e eventos em contexto (TURBAN; MCLEAN; WETHERBE, 2004).

Um conjunto de dados pode ser visto como registros, ponteiros, vetores, padrões, eventos, casos, exemplos, observações ou entidades. Muitas vezes, um conjunto de dados é um arquivo, onde os objetos são registros ou linhas. Do mesmo modo, o campo ou coluna corresponde a um atributo. Sendo assim, um atributo é uma propriedade ou característica de um objeto que pode variar, seja de um objeto para outro ou de tempo para outro (TAN; STEINBACH; KUMAR, 2009).

2.3.1 Mineração de Dados

Atualmente, é crescente tanto as conexões de acesso à Internet, quanto o armazenamento de dados, que são informações importantes que podem ficar em um repositório central de dados ou estarem em locais distribuídos. Esses dados podem estar em vários formatos, sejam eles em uma planilha, ou em arquivos os quais podem ser utilizados nas organizações para tomadas de decisões. Estas organizações precisam obter esse conhecimento que é valioso, mas em certos casos, estes dados não estão disponíveis

devido a falta de ferramentas adequadas para se realizar uma extração, desperdiçando estratégias das organizações.

Nos anos de 1980 o grande desafio era migrar dados para informações, por meio de Sistemas de Informações, que pudessem analisar dados e organizar informações para melhorar o processo decisório nas empresas. Já em 1990, criar sistemas capazes de representar e processar conhecimento para diferentes necessidades dos indivíduos ou grupos, passou a ser um novo desafio. A partir daí, os analistas de negócio passaram a usar ferramentas no processo de obtenção e análise das informações para extração de conhecimento (REZENDE, 2005). Dessa forma, a técnica de mineração de dados ingressou como uma ferramenta aliada a essas necessidades.

A mineração de dados é uma parte integral da Descoberta de Conhecimento em Banco de Dados (*Knowledge Discovery in Databases*, KDD), sendo uma ferramenta que começou a ser utilizada na década de 1980 para extração de dados a partir de grandes depósitos de dados, sendo atualmente um desafio nas pesquisas de Engenharia, Medicina, Ciência e Estatística.

A área de mineração de dados atrai ideias tais como: amostragem, estimativa e teste de hipóteses a partir de estatísticas e algoritmos de busca, técnicas de modelagem e teorias de aprendizagem da inteligência artificial, reconhecimento de padrões e aprendizagem de máquina (TAN; STEINBACH; KUMAR, 2009).

Em se tratando do ambiente para a realização da mineração deste trabalho, o os servidores da plataforma Facebook disponibilizam seu conteúdo em um repositório na *Web*. Dessa forma, esse tipo de mineração é considerada *Web Mining*. O termo *Web Mining*, introduzido por Pinheiro (2003), é um processo de descoberta e análise de informações úteis a partir de dados oriundos da Internet. O crescimento da

disponibilização de informações *online*, combinado com a falta de uma estrutura mínima para a maioria dos dados web, proporcionou o desenvolvimento de ferramentas poderosas, e computacionalmente eficientes, para a mineração de informações na Internet.

A descoberta de conhecimento em redes sociais, utilizando as técnicas de mineração de dados, inteligência artificial e aprendizado de máquina, vem sendo utilizadas num amplo cenário da Internet (FREITAS *et al*, 2008) desde a coleta até a análise de grande base d e dados (BENEVENUTO, 2011).

2.3.2 Processo de Descoberta de Conhecimento

O campo do KDD está associado com o desenvolvimento de métodos e técnicas, fazendo com que os dados tenham algum sentido. As empresas passaram a utilizar esses dados para ganhar vantagem competitiva, aumentar a eficiência e fornecer serviços mais importantes para os clientes. Dessa forma, é necessário recorrer às técnicas de computacionais, pois os seres humanos não conseguem reunir os dados para descobrir padrões significativos nas estruturas de uma grande base de dados (FAYYAD, 1996).

No mundo dos negócios, as principais áreas de aplicação de KDD incluem marketing, finanças (especialmente na área de investimento), detecção de fraude, manufatura, telecomunicações e agentes Internet. Estes são apenas alguns dos inúmeros sistemas que utilizam técnicas de KDD para produzir automaticamente informações úteis a partir de grandes massas de dados brutos (FAYYAD, 1996).

2.3.3 Etapas do KDD

A Figura 1 ilustra os passos adicionais no processo de KDD, tais como a preparação de dados, a seleção de dados, limpeza de dados, a incorporação do conhecimento prévio adequado, e uma interpretação correta dos resultados da extração é essencial para assegurar que o conhecimento útil é obtido a partir dos dados. Uma aplicação incorreta de métodos de mineração de dados pode ser uma atividade perigosa, que pode levar à descoberta de padrões sem sentido e inválidos (FAYYAD, 1996).

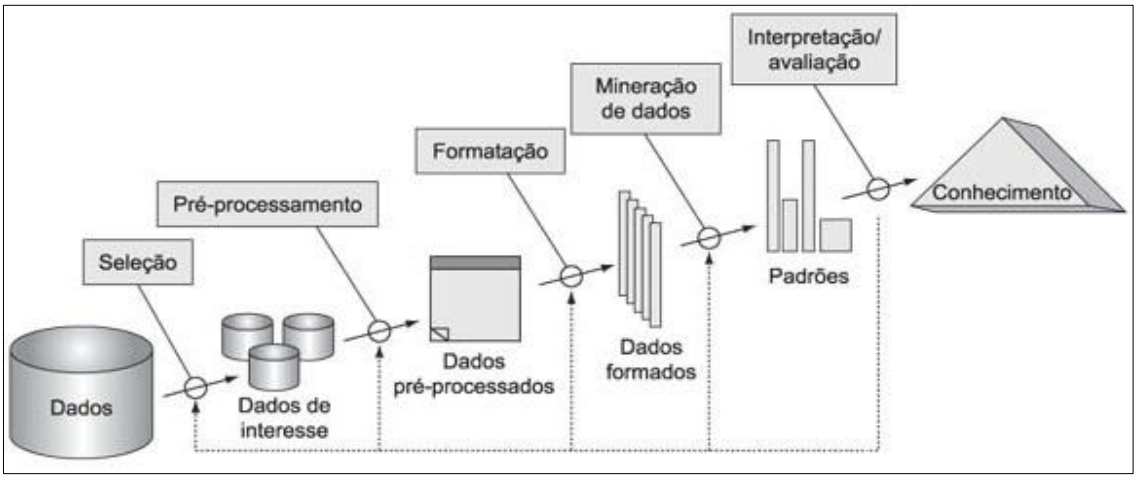

Figura 1: Etapas da descoberta de conhecimento em banco de dados.
Fonte: Fayyad (1996).

Ghosh *et al* (2008) descrevem também de forma sucinta as etapas do KDD:

1- Desenvolver uma compreensão do domínio da aplicação: o conhecimento prévio relevante, e os objetivos do usuário final.
2- Criar um conjunto de dados de destino: selecionar um conjunto de dados, ou simplesmente um subconjunto de variáveis ou de amostras de dados. Os conjuntos de dados resultantes dessa seleção são, então, pré-processados, ou

seja, recebem um tratamento para poderem ser submetidos aos métodos de extração de padrões.

3- Realizar limpeza de dados e pré-processamento: operações básicas, tais como a remoção de ruído ou de *outliers*.
4- Reduzir os dados e projetar: encontrar funcionalidades úteis para representar os dados dependendo do objetivo da tarefa. Usando redução de dimensionamento ou métodos de transformação para redução de um número efetivo de variáveis em questão ou para encontrar constantes de representação para os dados.
5- Escolher a tarefa de mineração de dados: decidir se o objetivo do processo de KDD é a sumarização, classificação, regressão, *clustering*, etc.
6- Escolher o algoritmo de extração de dados: método de seleção a ser utilizada para a busca de padrões nos dados. Isto inclui decidir quais modelos e parâmetros podem ser apropriados e combinando um método de mineração de dados em particular com os critérios gerais do processo de KDD.
7- Realizar a mineração de dados: em busca de padrões de interesse em uma forma de particular de representação ou um conjunto de tais representações: regras de classificação, regressão, *clustering*, e assim por diante. O usuário pode ajudar significativamente o método de mineração de dados, realizando corretamente os passos anteriores.
8- Interpretar os padrões minerados, possibilitando o retorno para todas as etapas 1-7 para posterior iteração. Este passo também pode envolver a visualização de padrões e modelos extraídos ou a visualização de dados extraídos dos modelos.

9- Consolidar a descoberta de conhecimento: incorporando esse conhecimento para o desempenho do sistema, ou simplesmente documentando e divulgando para as partes interessadas. Isso também inclui a verificação para resolver possíveis conflitos.

Uma organização precisa ter definidos, de forma bem clara, todos os passos do KDD, para posteriormente dar uma atenção especial na aplicação da mineração. A maior dificuldade, está em adquirir conhecimento nas extrações de informações válidas para se chegar ao processo de tomada de decisão.

2.3.4 Tarefas do KDD

As tarefas de mineração de dados, conforme Tan, Steinbach e Kumar (2009), são divididas em duas categorias que, são as tarefas de previsão e as tarefas descritivas. Nas tarefas de previsão o objetivo é prever o valor de um determinado atributo baseado nos valores de outros atributos. Este atributo a ser previsto é conhecido como a variável dependente ou alvo. Os atributos usados para fazer a previsão são conhecidos como as variáveis independentes ou explicativas. Já nas tarefas descritivas o objetivo é derivar padrões, que no caso são correlações, tendências, grupos, trajetórias e anomalias, que resumam relacionamentos subjacentes nos dados. Essas tarefas são, em muitos casos, de natureza exploratória, pois requerem técnicas de pós-processamento para validar os resultados.

As principais tarefas do KDD, de acordo com Fayyad (1996), são:

1- Classificação: aprender uma função que mapeia (classifica) um item de dados em uma das várias classes predefinidas.
2- Regressão: a aprendizagem de uma função que mapeia um item de dados para uma variável de previsão de valor real e a descoberta de relações funcionais entre as variáveis.
3- Agrupamento (*Clustering*): particionar os registros de um determinado banco de dados em vários subconjuntos de dados, produzindo um segmento de conjunto de registros de acordo com um critério pré definido. Tem o objetivo de agrupar dados semelhantes ou conseguir identificar algum tipo de exceção inerentes nos dados.
4- Sumarização: encontrar uma descrição compacta para um subconjunto de dados, por exemplo, a derivação de resumo ou de regras de associação e do uso de técnicas de visualização multivariados.
5- Modelagem de dependência: encontrar um modelo que descreva dependências significativas entre variáveis.
6- Mudança e Detecção de Desvio: descobrir as mudanças mais significativas nos dados medidos previamente ou valores normativos.

2.4 ANÁLISE DE GRUPOS

A análise de grupos divide os dados de uma amostra em grupos, conhecido como *clusters* ou módulos (FORTUNATO, 2010), que tenham significados, similaridades entre si e sejam úteis. Se esses grupos com significados forem o objetivo,

então os *clusters* devem capturar a estrutura natural dos dados. A análise de grupos tem desempenhado um papel importante nos campos das ciências sociais, biologia, estatística, reconhecimento de padrões, recuperação de informação, aprendizagem de máquina e mineração de dados. De acordo com a utilidade, essa técnica divide os dados em agrupamentos para compreensão, onde agrupamento (classe ou grupo) de objetos compartilham características comuns, sendo analisadas e descritas pelas pessoas, e agrupamento por utilidade, na qual fornece uma abstração de objetos individuais de dados para grupos nos quais esses objetos de dados residem. Estes podem ser caracterizados em termos de um protótipo de grupo que podem ser usados como base para análise ou processamento de dados (TAN; STEINBACH; KUMAR, 2009).

A técnica de análise de grupos agrupa objetos baseada em informações encontradas nos dados que descrevem os objetos e seus relacionamentos. O objetivo é que os objetos dentro de um grupo sejam semelhantes (ou relacionados) entre si e diferentes de (ou não relacionados aos) outros objetos de outros grupos. Quanto maior semelhança (ou homogeneidade) dentro de um grupo e maior a diferença entre grupos, melhor ou mais distinto será o agrupamento (TAN; STEINBACH; KUMAR, 2009).

2.4.1 O Algoritmo *k*-médias

Uma técnica simples que pode ser utilizada para análise de grupos é o *k*-médias (*k-means*), que foi proposto inicialmente por Macqueen em 1967, para várias aplicações, entre elas a de separar os objetos em grupos (*clusters*) segundo uma medida de distância ou similaridade entre eles (MACQUEEN, 1967). O *k*-médias é uma técnica particional, que realiza agrupamento por meio de otimização de uma função objetivo,

baseada em protótipos que tenta encontrar *n* número especificado pelo usuário de grupos *k*, que são representados pelos seus centróides (centros de gravidade) (TAN; STEINBACH; KUMAR, 2009).

O algoritmo *k*-médias consiste numa técnica não-hierárquica de agrupamento baseada em protótipos que criam um particionamento de um nível dos objetos de dados (TAN; STEINBACH; KUMAR, 2009). A técnica não hierárquica define o número *k* de classes que se pretende constituir e faz-se uma classificação inicial dos *n* indivíduos em *k* classes (JOHNSON; WICHERN 1998). O número de grupos *k* pode ser especificado antecipadamente ou durante o processo de agrupamento.

O *k*-médias possui fácil programação e é computacionalmente econômico, capaz de processar grandes volumes de dados, sendo que o armazenamento requerido é *O((m+K)n),* onde *m* é o número de pontos e *n* é o número de atributos (MACQUEEN, 1967). O *k*-médias é um algoritmo particional que se baseia no cálculo do erro quadrático (*Sum of Squared Erros*, SSE) como sua função objetivo. A soma do erro quadrático mede a qualidade de um agrupamento, onde é possível calcular o erro de cada ponto de dados, a distância Euclidiana, até aqueles centroides mais próximos e depois calcula-se a soma total dos erros quadrados. O erro quadrático é definido conforme equação SSE= $\sum_{i=1}^{K} \sum_{x \in Ci} dist\ (c_i, x)^2$, onde *dist* é a distância Euclidiana padrão entre dois objetos no espaço Euclidiano. O centróide que minimiza o SSE do grupo é a média. Este centróide do grupo de índice *i* é definido pela equação $c_i = \frac{1}{m_i} \sum_{x \in C_i} x$ (TAN; STEINBACH; KUMAR, 2009). O objetivo do algoritmo é, portanto, obter uma partição que minimiza o erro quadrático para um número *k* fixo de agrupamentos. Portanto o valor de *k* deve ser fornecido pelo usuário (BOTELHO; SOUSA, 2011). Dados os conjuntos diferentes de grupos que são produzidos pelas execuções do algoritmo *k*-médias, é preferível a com

menor erro quadrado, já que isso significa que os centroides deste agrupamento são uma representação melhor dos pontos do seu grupo (TAN; STEINBACH; KUMAR, 2009).

Numa base de dados ampla, o *k*-médias não pode ser eficiente na determinação de soluções viáveis de boa qualidade quando na sua inicialização não for bem sucedida e os centroides iniciais representantes dos grupos ficarem mal posicionados no espaço de busca (COELHO FILHO *et al*, 2013). Outra desvantagem do *k*-médias é que, em termos de desempenho o algoritmo não garante o resultado global ótimo, pois a qualidade da solução final depende muito dos conjuntos inicias de clusters, podendo, na prática, vir a se afastar muito do ótimo global (SALDANHA; FREITAS, 2009). Outro inconveniente deste algoritmo é que como o número de agrupamentos é um parâmetro de entrada uma escolha inapropriada de *k* pode retornar em resultados pobres (SALDANHA; FREITAS, 2009).

Conforme ilustrado na Figura 3, o algoritmo do *k*-médias converge para uma solução para algumas combinações de funções de proximidade e tipos de centróides, atingindo um estado no qual nenhum ponto muda de um grupo para outro e, assim, os centróides não mudam. Caso isso não ocorra, a condição da linha 5 do algoritmo é muitas vezes substituída por uma condição mais fraca, por exemplo, repetir até que apenas 1% dos pontos mudem de grupo (TAN; STEINBACH; KUMAR, 2009).

```
1 Selecione K pontos como centroides iniciais
2 repita
3     Forme K grupos atribuindo cada ponto ao seu centroide mais próximo
4     Recalcule o centroide de cada grupo/
5 até que os centroides não mudem
```

Figura 2: Algoritmo *k*-médias básico
Fonte: Tan; Steinbach; Kumar, (2009).

Para identificar grupos de usuários que compartilham padrões comportamentais é utilizado o algoritmo de agrupamento *k*-médias (MACQUEEN, 1967). No que se refere a detectar grupos com formas ou tamanhos diferentes, o *k*-médias apresenta algumas dificuldades tais como, não lida com grupos não globulares, de tamanhos e densidades diferentes, mas, por outro lado, é bastante eficiente em múltiplas execuções com frequência. Sendo assim, o processo de mover os centroides para quando todos os pontos não mudam suas associações ou apenas uma pequena parcela deles muda (FERREIRA, 2012).

O *k*-médias pode ser utilizado como técnica para agrupar textos (GOMES; PARDO, 2009), registrar dados que possuem características semelhantes em um repositório de dados (*data mart*) (SARTORI, 2012), bem como para classificação não-supervisionada, que por sua vez são especificadas em quantas classes (*clusters*) as instâncias de um arquivo devem ser agrupadas por critérios induzidos do próprio algoritmo (SANTOS, 2005). Assim, este algoritmo visa particionar um conjunto de elementos numa coleção de *k* agrupamentos, onde cada elemento é alocado no agrupamento de cujo centróide se encontra mais próximo (JOHNSON; WICHERN 1998).

2.4.2 Medidas de Centralidade

A centralidade é um dos conceitos mais básicos no estudo de redes. Um índice de centralidade busca descobrir quais vértices são mais importantes dentro da rede (WASSERMAN; FAUST, 1994).

As medidas de centralidade permitem determinar a posição de um determinado vértice na estrutura de um grafo. Assim, é possível descobrir quais os vértices mais centrais, ou seja, com mais ligações. Estes vértices são importantes pois é através deles que a informação consegue fluir com mais rapidez. As medidas baseadas no conceito de centralidade são: grau de conectividade (*degree*), proximidade (*closeness*), e grau de intermediação (*betweenness*) (FERREIRA, 2013).

O grau de conectividade (*degree*) de um vértice é simplesmente o número de vértices que estão conectados a ele. Em um grado direcionado, o grau de conectividade é a soma do número de arestas que entram mais o número de arestas que saem do vértice (VIEIRA, 2011). O grau de proximidade (*closeness*) é a medida média de proximidade de um nó em relação a todos os atores de uma rede. Calcula-se contando todas as distâncias de um ator para se ligar aos restantes (FREEMAN, 1977).

A centralidade de intermediação (*Betweenness Centrality*, BC) foi proposta inicialmente por Freeman no ano de 1977. Esta centralidade é obtida com a contagem de arestas realizadas por cada vértice (FREEMAN, 1977). Na centralidade de intermediação para cada vértice da rede são verificados quantos relacionamentos são intermediados (QUEIROZ, 2012). Contudo, é possível mostrar o quão importante o vértice é para o grafo no sentido de que se for retirado do grafo quebrará muitas relações dificultando o fluxo do grafo (NASCIMENTO, 2013).

O vértice que se encontra localizado estrategicamente num caminho mais curto de comunicação, entre pares de indivíduos, está numa posição mais central da *Web*. Portanto, esse vértice é mais influente e de maior visibilidade, considerado grande disseminador de informações e possivelmente o de maior aceitação (MIKA, 2007).

Em uma rede social, um membro pode realizar postagens de mensagens e notícias que são repassadas de uma pessoa para outra. Considera-se então, a maior quantidade de vezes que esse membro transmitiu uma postagem. Esse membro é um vértice que pode ter uma elevada centralidade, tendo assim, uma influência considerável dentro de uma rede, em virtude de seu controle sobre a informação que passa entre outros membros (NEWMAN, 2010). Conforme é ilustrado na Figura 3, os vértices A e B estão conectados por dois caminhos geodésicos. Já o vértice C encontra-se em ambos os caminhos.

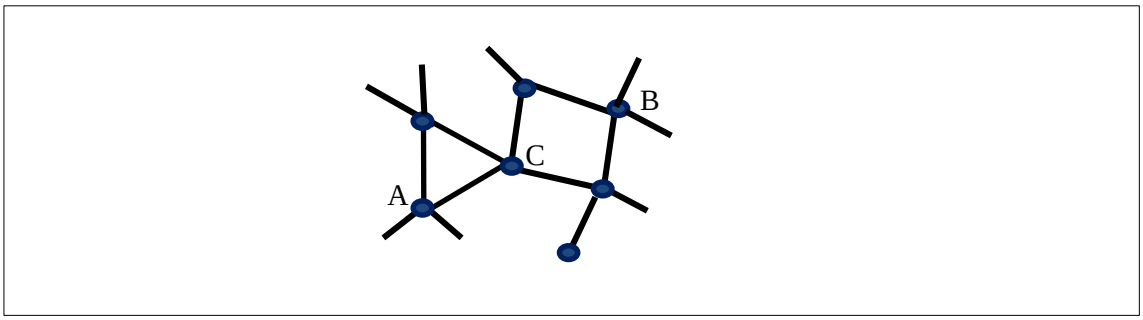

Figura 3: Vértices conectados por dois caminhos geodésicos.
Fonte: Newman (2010)

A centralidade de intermediação produz bons resultados e é recomendado em redes menores (NEWMAN, 2004) e quanto mais centrais, mais importantes são determinados atores nesta rede (ROSSONI; GUARIDO FILHO, 2007). Dessa forma, é avaliada a dependência de classes de serviços e produtos não adjacentes de outras que atuam como uma espécie de ponte para a efetivação da interação entre eles (FREEMAN, 1979). Esta centralidade é uma técnica hierárquica divisiva, que calcula o menor caminho entre todos os pares de vértices, pela contagem do número que cada aresta é percorrida e posteriormente remove as mais utilizadas até que cada vértice forme uma comunidade (NEWMAN; GIRVAN 2003). A centralidade de intermediação de um vértice pode ser

entendida como quantos caminhos mínimos, onde o caminho mais curto de um vértice para outro, que passam por um vértice quando se conta esses caminhos entre todos os pares possíveis de vértices da rede. Se algo flui em uma rede, seja informação, dinheiro, ou mesmo transporte, os vértices ou ligações com alta intermediação são aqueles que detém a maior taxa desse fluxo (FLECHA *et al*, 2011).

Nos grupos de vértices, também chamados de comunidades, são aplicados algoritmos com a medida de centralidade de intermediação, na tentativa de se extrair informações na *Web*, nos quais se utiliza o cálculo do caminho mínimo entre os vértices para fundamentar a detecção de comunidades (GIRVAN; NEWMAN, 2002).

A comunicação entre dois vértices não adjacentes (FOROUZAN; MOSHARRAF, 2011; MOKARZEL; SOMA, 2008), *a* e *b*, depende dos vértices pertencentes ao caminho que conecta *a* a *b*. A medida de centralidade de intermediação é representada pela equação: $c_B(v) = \sum_{a,b \neq v} \frac{\sigma_{avb}}{\sigma_{ab}}$, onde σ_{ab} é o número de geodésicas (caminhos de tamanhos mínimo) entre *a* e *b*, e σ_{avb}é o número dessas que passam por *v*. (FREEMAN, 1977).

De acordo com o exemplo de Queiroz (2012), na Figura 5, aplica-se o conceito para o vértice 1 da rede, onde σ_{23}=1, pois não há mais de um menor caminho com custos iguais entre os vértices 2 e 3. Da mesma maneira, σ_{24}= 1 e σ_{34}=1 e assim, $\sigma_{23}(1)$=1, $\sigma_{24}(1)$=1 e $\sigma_{34}(1)$=0, pois o menor caminho entre os vértices 3 e 4 não inclui o vértice 1. Pela equação da centralidade de intermediação, que segue: $c_B(1)$ = $\sigma_{23}(1)/\ \sigma_{23}$+$\sigma_{24}(1)/\ \sigma_{24}$+$\sigma_{34}(1)/\ \sigma_{34}$ = 1/1+1/1+0/1 = 2 Desta maneira, conclui-se que $c_B(1)$=2, ou seja, do total de 3 menores caminhos entre os nós 2, 3 e 4 da rede da Figura 4, 2 passam pelo vértice 1.

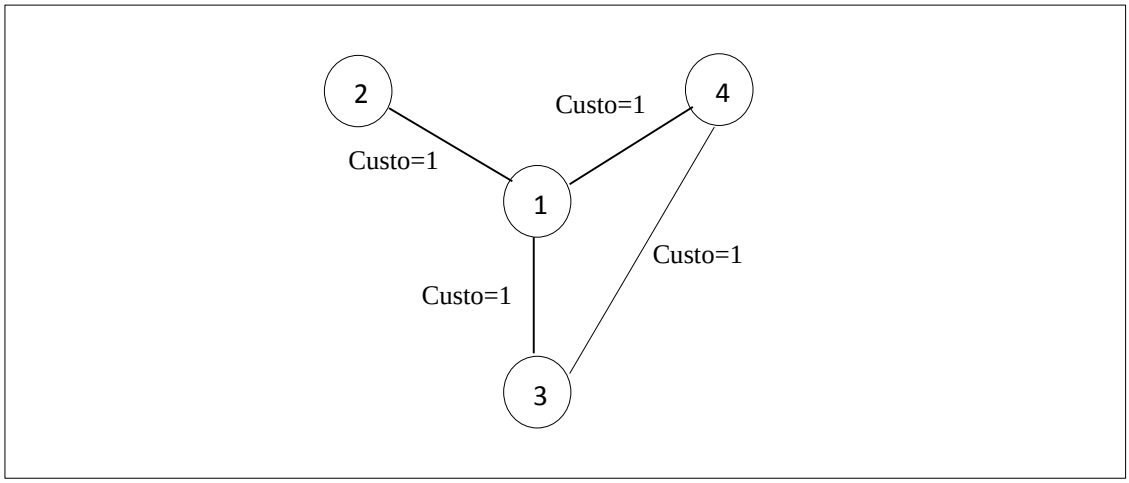

Figura 4: Rede de exemplo para cálculo de centralidade.
Fonte:Queiroz (2013)

O algoritmo proposto por Girvan e Newman (2002) para a medida da centralidade de intermediação, utilizado neste trabalho, segue na Figura 6. Os autores destacam que nas redes sociais é possível observar a existência de comunidades sem que tal apareça claramente definida nos métodos estatísticos tradicionais (GIRVAN; NEWMAN, 2002).

```
1 Início
2       Calcula centralidade de todas as arestas do grafo;
3       repita.
4               Remova aresta de maior centralidade;
5               Recalcula os valores das centralidades;
6       até encontrar K comunidades;
7 fim
```

Figura 5: Algoritmo de centralidade de intermediação
Fonte: Girvan e Newman (2002)

Este algoritmo é capaz de detectar comunidades a partir das ligações que sejam menos centrais às comunidades, dividindo o grafo original e o particionamento através da remoção das arestas menos importantes (GIRVAN; NEWMAN, 2002). Sendo assim, a medida de centralidade é adotada para quantificar a importância de cada uma das ligações do grafo.

O método de Girvan e Newman (2002) é uma boa maneira conceitual para pensar sobre particionamento de um grafo, e ele funciona bem para redes de tamanho moderado de até alguns milhares de vértices. No entanto, para as redes maiores, a necessidade para recalcular valores de centralidade de intermediação a cada passo torna-se computacionalmente muito caro, uma vez que realiza cálculos recursivos para remoção iterativa de arestas que possuem alto grau da medida, apresentando, no pior caso, complexidade de tempo de $O(M^2N)$, para uma rede de M arestas e N vértices (EASLEY; KLEINBERK, 2010).

3. ESTUDO DE CASO

Este Capítulo apresenta um estudo de caso da aplicabilidade do processo de KDD nos dados experimentais coletados na página do subgrupo Tocantins digital, no ambiente da plataforma da rede social Facebook, do qual fazem parte do universo da pesquisa mais de dez mil membros seguidores. O estudo de caso é um trabalho de caráter empírico que investiga um dado fenômeno dentro de um contexto real contemporâneo por meio de análise aprofundada de um ou mais objetos de análise (MIGUEL *et al*, 2012). Dessa forma, são descritos a seguir, de forma detalhada, os experimentos utilizados na mineração e análise dos dados.

O subgrupo Tocantins digital permite, através de suas próprias ferramentas API, realizar a extração de informações e permitir que outros programas utilizem das suas funções. A API do Tocantins digital possui funções que são utilizadas para complementar as aplicações, por meio dos dados extraídos dos relacionamentos dos seus usuários e informações contidas nas postagens, das quais são armazenados em formato de tabelas.

O fato dos usuários compartilharem conteúdo, como mensagens de texto, imagens, vídeos, etc, atraíram uma quantidade surpreendente de pessoas e tornou-se um ambiente ideal para estudos aplicando a extração de informações. Dessa forma, se existem informações fluindo nesse grupo, é possível identificar agrupamentos que são comunidades formadas pelos interesses em postagens de produtos e serviços e destacar os vértices com um alto número de intervenções dos quais detém a maior taxa desse fluxo.

3.1 FERRAMENTAS DA REDE SOCIAL FACEBOOK

Inicialmente, os dados só podem ser extraídos mediante um cadastro de membro no subgrupo Tocantins digital. Em seguida, é possível desenvolver aplicações fornecidas por esta plataforma, como a ferramenta Grafo API (*Graph API*), que permite a consulta aos elementos do grafo e as conexões do subgrupo pesquisado. As ferramentas APIs são portas voluntariamente abertas pelos elaboradores dos programas para permitir a outros programadores apropriar-se de funções interessantes e acrescentá-las a seus próprios aplicativos (PISANI; PIOTET, 2010). O Grafo API é baseado em Protocolo de Transferência de Hipertexto (*Hypertext Transfer Protocol*, HTTP), para consulta de dados e permite o acesso ao grafo social do Tocantins digital, o que representa de maneira uniforme os objetos no grafo e as conexões entre eles (FACEBOOK, 2013). Sendo assim, das postagens dos membros é construído um grafo *G* formado por um conjunto de vértices *V*, representados pelos membros, produtos e serviços. As conexões entre os vértices são representadas pelas arestas *E* (GOODRICH; TAMASSIA, 2007). Por meio desta API, é possível extrair dados das primeiras postagens até àquelas publicadas no mês de dezembro do ano de 2013.

O Grafo API é uma plataforma que fornece ferramentas para construção de aplicações por terceiros a serem oferecidas aos membros do subgrupo Tocantins Digital. Com o API é possível usar informações das conexões do usuário, assim como as informações do seu perfil, buscando tornar a aplicação mais envolvente, assim como possibilita a publicação de novas interações do usuário, tanto em seu *feed*[1] de notícias como nas páginas dos amigos do usuário (AQUINO; BRITO, 2012).

[1] Lista contínua de atualizações na página inicial que mostra as novidades de seus amigos e das páginas que você segue no Facebook.

O Facebook dispõe de outra API que permite o acesso ao grafo social, chamada de Linguagem de Consulta do Facebook (*Facebook Query Language*, FQL), que utiliza uma linguagem de consulta (*queries)* semelhante a Linguagem de Consulta Estruturada (*Structured Query Language*, SQL). A FQL permite a execução de múltiplas chamadas em uma única consulta e também a possibilidade de escolher o formato de retorno da resposta, sendo Notação de Objeto de Java Script (*Java Script Object Notation*, JSON) ou Linguagem de Marcação Extensível (*eXtensible Markup Language*, XML). Os termos utilizados no Grafo API são (FACEBOOK, 2013):

- Facebook *Login*: é uma ferramenta personalizada e segura para as pessoas acessarem seu aplicativo. O protocolo OAuth é utilizado para confirmar a identidade de uma pessoa, por meio de uma autenticação, e dando-lhe autorização do controle sobre direito de acesso às suas informações.

- Autorização: qualquer pessoa deverá autorizar o acesso a suas informações. A Autorização é um aplicativo que autoriza o acesso a essas informações.

- Permissão: através do Facebook *Login*, as pessoas concedem o acesso a dados básicos, lista de amigos e até mesmo das suas próprias ações.

- *Token* de acesso: Os *tokens* de acesso são sequências aleatórias que dão acesso seguro. É gerado no final do processo de autorização, mas temporário para as APIs. Representa um conjunto de permissões que foram concedidas e podem ser utilizados no contexto de uma aplicação específica. Os *tokens* são utilizados em diferentes casos como: *token* de usuário (para ler, modificar ou gravas dados de uma pessoa), *token* de aplicativo (para modificar e ler as configurações de aplicativos), *token* de página (fornecem permissão para APIs ler, escrever ou modificar os dados pertencentes a uma página).

Para se explorar o subgrupo Tocantins digital, a plataforma dispõe do Grafo API Explorer, que consiste em uma ferramenta gráfica utilizada para consultar, adicionar e remover dados das tabelas com informações com o perfil do membro do subgrupo e as informações contidas nas postagens. Com esta ferramenta é possível gerar *tokens* de autenticação e testar resultados de requisição HTTP, conforme ilustra a Figura 6 (FACEBOOK, 2013).

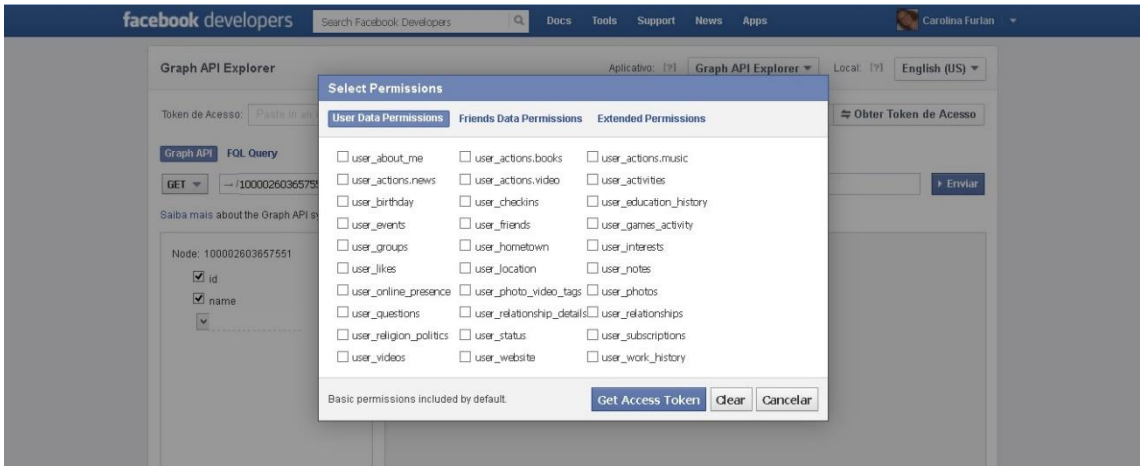

Figura 6: Grafo API Explorer
Fonte: Facebook (2013)

No que tange a privacidade, o Tocantins digital disponibiliza um controle sobre os dados, de forma que os usuários escolham o querem compartilhar com as aplicações (*app*). Dessa forma, mesmo que se tenha permissão de acesso aos dados, é possível que o desenvolvedor não consiga os dados desejados (LIANG *et al*, 2011).

Para permitir aos desenvolvedores um conjunto de funcionalidades para acessar a API do Tocantins digital, incluindo acesso à todos os recursos do Grafo API e FQL, o Facebook dispõe do aplicativo *opensource*[2] Kit de Desenvolvimento de Software (*Software Development Kit*, SDK). O SDK é constantemente utilizado para realizar operações como um administrador de aplicação, mas também pode ser utilizado para

[2] Programa fonte é aberto

realizar operações em nome do usuário atual da sessão. Ao eliminar a necessidade de gerenciar *tokens* de acesso manualmente, o SDK simplifica muito o processo de autenticação e autorização de usuários para seu aplicativo. Os SDK's oficiais para se integrar ao Facebook estão disponíveis para iOS, Android, JavaScript e PHP[3] (FACEBOOK, 2013).

As APIs não são somente um meio de abrir o código para enriquecer uma aplicação. Elas são também potencialmente uma fonte de receitas derivadas para o editor que autoriza a utilização de seu sistema (PISANI; PIOTET, 2010).

3.2 EXTRAÇÃO DOS DADOS DO TOCANTINS DIGITAL

Os dados foram obtidos através de um conjunto de dados extraídos por meio da ferramenta API do subgrupo Tocantins digital. Esta plataforma disponibiliza diversas tabelas para serem realizadas consultas aos dados postados pelos seus usuários, por meio da FQL. Na Figura 7 estão representados os procedimentos adotados na realização da extração dos dados.

[3] Linguagem de *script* livre de uso geral.

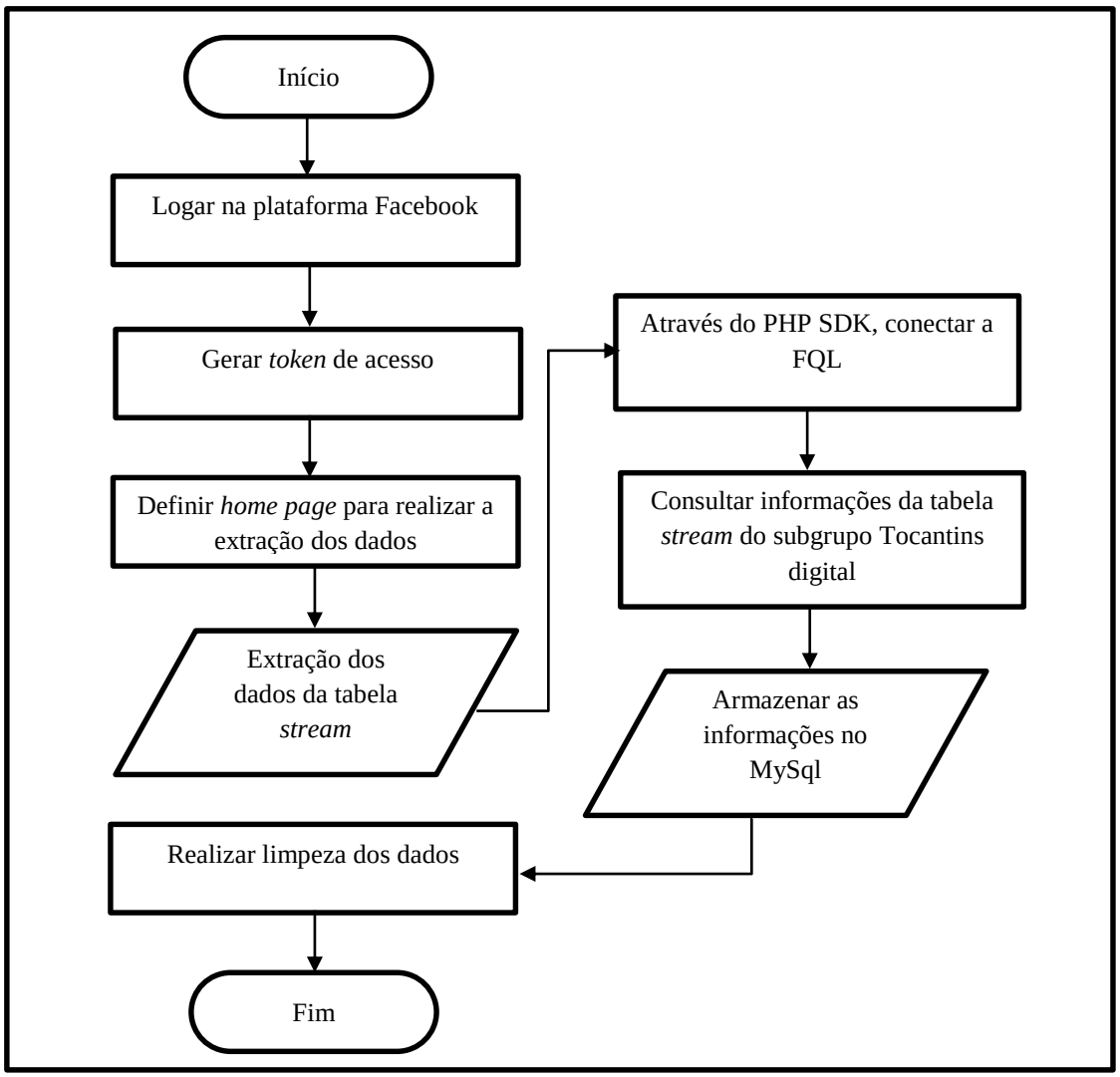

Figura 7: Fluxograma da execução da extração dos dados
Fonte: Autoria própria

Neste trabalho foi utilizada apenas a tabela *stream* que contém informações dos *posts*. A *stream* pode ser consultada através da FQL para retornar uma lista de informações dos posts de um usuário, de forma que o usuário deverá conceder permissão de acesso (*read_stream* e *read_insights*) à *app*. A Tabela 1apresenta as colunas da tabela *stream* com suas respectivas descrições.

Tabela 1: Colunas da tabela stream

Nome	**Tipo**	**Descrição**
action_links	*array*	Um *array* que contém o texto e a URL para cada link de ação.
actor_id	*string*	O ID do usuário, página, grupo ou evento que publicou o *post*.
app_data	*array*	Um *array* de informações do aplicativo específico, opcionalmente, fornecido para criar a ligação com o *post*.
app_id	*int*	Para os *post* publicadas por *apps*. Nesse caso é o ID desse *app*. Se o valor estiver vazio, isso indica que o *post* originou-se do Facebook.
attachment	*array*	Um *array* das informações anexadas sobre o *post*.
attribution	*string*	Para as *posts* publicadas por *apps*. Nesse caso, será o nome completo da *app*.
claim_count		Contagem para quantas pessoas receberam a oferta.
comment_info	*array*	Informações sobre os comentários deixados sobre este *post*.
created_time	*time*	Data e hora em formato UNIX *timestamp*.
description	*string*	Texto de histórias não intencionalmente gerados por usuários, como aqueles gerados quando dois usuários se tornam amigos. Você deve ter a migração "Incluir histórias atividades recentes" habilitado em seu aplicativo para obter nesse campo.
description_tags		A lista de *tags* na descrição do *post*.
expiration_timestamp	*string*	timestamp UNIX de quando a oferta expira
feed_targeting		Alimentação das informações
filter_key	*string*	Chave de filtro para buscar dados.
impressions	*int*	Número de impressões sobre o *post*. Esta informação só é visível se você tiver permissão *read_insights* de um proprietário da página.
is_exportable	*unsigned int32*	Se o *post* é exportável.
is_hidden	*bool*	Se um *post* foi definido como oculto.
is_published	*bool*	Se o *post* é publicado
like_info	*struct*	Informações sobre os gostos neste *post*.
message	*string*	A mensagem escrita no *post*.
message_tags	*object*	A lista de *tags* na mensagem de correio.
parent_post_id	*string*	ID da postagem de origem
permalink	*string*	A URL do *post*
place	*id*	ID do lugar associado ao *post*.
post_id	*string*	O ID do *post*
privacy	*struct*	As configurações de privacidade para um *post*
promotion_status	*string*	Status da promoção, se a mensagem foi promovido
scheduled_publish_time	*timestamp*	*Post* programado para ser publicado no formato timestamp Unix
share_count	*unsigned int32*	Número de vezes que o *post* foi compartilhado.
share_info	*struct*	Informações sobre as ações do *post*
source_id	*id*	O ID do usuário, página, grupo ou evento cujo o *post* é referenciado.
subscribed	*bool*	Se usuário está inscrito para o *post*.
tagged_ids	*array*	Um *array* de IDs marcados na mensagem do *post*
target_id	*id*	O usuário, página, grupo ou evento a quem a mensagem foi direcionada.
targeting	*struct*	Anúncios direcionados com informações do *post*
timeline_visibility	*string*	Informações da visibilidade do post na linha do tempo
type	*int32*	O tipo de história
updated_time	*timestamp*	O tempo que o post foi atualizado, o que ocorre quando um usuário comenta sobre o *post*. É expressado timestamp UNIX.

via_id	*numeric string*	ID do usuário ou página que compartilhou o *post*.
viewer_id	*id*	O ID do usuário da sessão atual.
with_location	*bool*	Se existe um local associado com o *post*
with_tags	*array*	Um *array* de IDs de entidades (por exemplo, usuários) marcados no *post*.
xid	*string*	Associado às notícias caixa de transmissão ao vivo

Fonte: Facebook (2013)

As colunas *post_id* (identificador do *post*), *source_id* (identificador do usuário, da página ou do grupo cujo *post* faz referência) e *filter_key* (identificador dos filtros do usuário) são campos indexáveis exigidos na parte *WHERE* da consulta. Para encontrar informações como conteúdo da mensagem e *actor_id* (número de identificação do usuário, página ou grupo que publicou o post) de um determinado *post*, deve-se substituir a variável "A" pelo ID do *post* ou do *source_id*, conforme ilustra o trecho de FQL na Figura 8:

```
SELECT message, actor_id FROM stream WHERE post_id = A

SELECT message, actor_id FROM stream WHERE source_id = A

SELECT message, actor_id FROM stream WHERE filter_key = A
```

Figura 8: Trecho do FQL para obter informações de um post.
Fonte: Facebook (2013)

Para o acesso à tabela *stream*, é necessário obter um token de acesso temporário. Este *token* pode ser obtido através de uma série de métodos, tais como: *User Access Token, Client Token, Page Access Token* e *App Access Token.* O *User Access Token* é o *token* de acesso de usuário que é necessário todas as vezes que o aplicativo chama uma API para ler, modificar ou gravar dados de uma pessoa específica do subgrupo

Tocantins Digital em seu nome. Nesse caso, ele requer de uma pessoa a permissão para sua aplicação. O *Client Token* é um identificador que pode ser instalado em celulares ou aplicativos de desktop que serve para acessar APIs de nível aplicativo, mas em um subconjunto limitado. O *Page Access Token* fornece permissão para APIs que lêem, escrevem ou modificam os dados pertencentes a uma página no Tocantins Digital.

Neste trabalho foi utilizado o método *App Access Token*, que é um tipo de *token* de acesso que modifica e lê as configurações de aplicativos. Na *home page* da Plataforma Tocantins digital, *developers.facebook.com*, foi criada a *App cfurlan* onde obtém o ID de identificação do usuário e a *app secret* que é a senha gerada pelo próprio subgrupo. Dessa forma, com este *token* de acesso criado, foi possível definir a home page *cfurlan.grupog3brasil.com.br*, que é um sub-domínio particular da pesquisadora, para realizar a extração dos dados. Entretanto, pode-se utilizar qualquer sub-domínio para realizar a extração.

Por meio desse *token*, com as permissões concedidas *read_mailbox* (visualiza a caixa de entrada do usuário), *read_stream* (visualiza os posts) e *read_insights* (recupera métricas para todas as páginas e domínios pertencentes ao usuário), foi utilizado o PHP SDK, o qual dispõem de duas classes, sendo uma classe *cfurlan* que fornece uma implementação concreta que utiliza sessões PHP para armazenar IDs de usuários e *tokens* de acesso, e a classe *BaseFacebook*, que fornece acesso à plataforma. Essa classe fornece a maioria das funcionalidades necessárias, mas é uma classe abstrata, porque é projetada para ser uma subclasse de modo que você pode definir como os dados devem ser armazenados na aplicação. A classe Tocantins digital implementa esses métodos abstratos, utilizando sessões PHP (FACEBOOK, 2013). As descrições dos métodos da classe são descritas na Tabela 2. Após a extração dos dados, os mesmos foram

armazenados em tabelas no mesmo site da *App* em um sistema de gerenciamento de banco de dados MySQL.

Tabela 2: Métodos da classe Facebook SDK para PHP

Nome	Descrição
Api	Chamar o método Graph API ou uma consulta de FQL usando o SDK PHP.
destroySession	Quebrar a sessão atual.
getAccessToken	Obter o token de acesso atual que está sendo usado pelo SDK.
getAppId	Obter o App ID que o SDK está usando atualmente.
getApplicationAccessToken	Obter o token de acesso que deve ser usado para usuários registrados quando nenhum código de autorização está disponível.
getAppSecret	Obter o App secret que o SDK está usando atualmente.
getFileUploadSupport	Verificar se o suporte de upload de arquivo está habilitado no SDK.
getLoginUrl	Obter uma URL que o usuário pode clicar para iniciar sessão, autorizar o aplicativo, e ser redirecionado de volta para o aplicativo.
getLogoutUrl	Este método retorna uma URL que, quando clicado pelo usuário, irá registrá-los fora de sua sessão do Facebook e depois redirecioná-los de volta para a sua aplicação.
getSignedRequest	Obter o pedido assinado atual que está sendo usado pelo SDK.
getUser	Esse método retorna o ID de usuário atual do Facebook, ou 0 se não há usuário logado
setAccessToken	Definir o token de acesso atual que está sendo usado pelo SDK.
setAppSecret	Definir o App secret que o SDK está usando atualmente
setAppId	Definir o App ID que o SDK está usando atualmente.
setExtendedAccessToken	Estende o *token* de acesso atual que está sendo usado pelo SDK para ser um *token* long-lived. Isso requer a existência de um *token* de acesso válido.
setFileUploadSupport	Definir o suporte de upload de arquivos no SDK.
setPersistentData	Este é um método protegido abstrato, que devem ser implementados em classes que se estendem.
getPersistentData	Obter o valor armazenado de uma determinada chave, que foi criado com *BaseFacebook* :: *setPersistentData*
clearPersistentData	Apaga o valor armazenado de uma determinada chave.
clearAllPersistentData	Remover um aplicativo de um outra conta.
getResult	Obter o objeto que é o resultado do erro ou exceção retornada pelo servidor.
getType	Obter o tipo de erro ou exceção, por exemplo, OAuthException.

Fonte: FACEBOOK (2013)

3.3 LIMPEZA DE DADOS E PRÉ-PROCESSAMENTO

Após a extração, inicia-se uma limpeza de valores nulos através de uma consulta SQL. Essa consulta é destinada para operação de limpeza que tem em sua formulação a detecção e exclusão de valores nulos, que são postagens retiradas pelos

próprios membros. Posteriormente, foi possível realizar a tabulação de membros que realizaram algum tipo de postagem de prestação de serviço ou venda de produtos.

A informações brutas que foram coletadas, do subgrupo Tocantins Digital, da tabela *stream* da FQL, posteriormente foram armazenadas em uma nova tabela da *App cfurlan* no banco de dados MySQL. Posteriormente, foi realizada uma consulta SQL que buscou no campo *message* valores nulos e valores repetidos permitindo dessa forma, deletar os *posts* identificados nessa consulta da base da *App*. Os valores repetidos correspondem a inserção de uma mesma postagem realizada repetidas vezes pelo mesmo membro.

3.4 REDUÇÃO DOS DADOS E PROJEÇÃO

No banco de dados da *App cfurlan* foram criadas as seguintes tabelas SQL, conforme descritas na Tabela 1:

Tabela 3: Tabelas do App cfurlan em MySQL

Tabela	Função	Exemplo
class_produtos	Classificação dos produtos em categorias e a descrição de cada produto desta categoria.	Tipo de categoria: Informática Descrição do produto: notebook
class_servicos	Classificação dos serviços em categorias e a descrição de cada serviço desta categoria.	Tipo de categoria: serviço Descrição do serviço: eletricista
cliente_produto	Lista quantos *posts* um usuário publicou de um determinado tipo de serviço ou produto.	O ID 5334357XX postou 2 vezes um tipo de produto.
stream	Armazena algumas informações da *stream* da FQL.	Ator da publicação, data da postagem, mensagem, entre outras.

Fonte: Autoria própria

Para auxiliar na compreensão da tabela *class_produtos*, foram criadas as categorias informática, veículos, eletrodomésticos, eletroportáteis, móveis, celulares e

telefones para classificar os produtos ofertados nos classificados do subgrupo pesquisado. Como exemplo, tem-se um *post* que contém informações de venda de computador que é classificado no tipo de categoria Informática.

Na tabela *class_servicos* foi criada apenas a categoria serviço, que classifica prestadores de serviços como pintor, pedreiro, eletricista, encanador, doméstica, diarista, motorista, carpinteiro entre outros.

Na tabela *stream* foram listados 6.579 *posts* durante o período de coleta dos dados, sendo que após uma consulta SQL utilizando a cláusula *distintic*, foi possível identificar que apenas 72 membros realizavam algum tipo de postagem referente à oferta de serviços ou produtos. Em seguida, na tabela *cliente_produto* foram armazenadas informações de forma a classificar o quê e quantas vezes cada membro postou um determinado tipo de produto ou serviço. Conforme ilustrado na Tabela 2, a coluna *actor_id* são os números IDs dos membros do subrupo. As colunas identificadas com a inicial "p" representam algum tipo de serviço ou produto. Devido a grande quantidade de informações da tabela original, esta tabela possui uma representação parcial dos dados tabulados. Assim, o membro 5847576XX que postou 14 vezes ofertas do produto ou serviço p7 e o membro 6295014XX postou 10 ofertas do produto ou serviço p6, 12 do produto ou serviço p10 e 15 ofertas do produto ou serviço p15.

Tabela 4: Tabela cliente_produto

actor_id	p1	p2	p3	p4	p5	p6	p7	p8	p9	p10	p11	p12	p13	p14	p15
5251521XX	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0
5334357XX	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0
5847576XX	0	0	0	0	0	0	14	0	0	0	0	0	0	0	0
6011936XX	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6142021XX	0	1	0	0	0	0	1	0	0	0	0	0	0	0	1

6295014XX	0	0	0	0	0	10	0	0	0	12	0	0	0	0	15
6336199XX	0	1	0	0	0	0	0	0	0	7	0	0	0	0	0

Fonte: Autoria própria

Diante dos dados levantados, aliados ao fato que uma pessoa não consegue lidar com essa grande quantidade de informações das postagens do Tocantins digital, os dados da tabela *cliente_produto* foram exportados, em um arquivo no formato *csv*, para posteriormente serem utilizados neste trabalho para apresentar uma forma de explorar a técnica de agrupamento *k*-médias e a medida de centralidade de intermediação.

3.5 MINERAÇÃO DE DADOS COM *K*-MÉDIAS

Nesta etapa foi proposta a utilização do algoritmo *k*-médias, por ser um algoritmo de agrupamento (*clustering*) que será aplicado ao subgrupo Tocantins digital. Dessa forma, essa técnica pode ser aplicada ao problema de posicionamento dos vértices que correspondem aos nomes de produtos e de serviços postados no subgrupo.

Inicialmente, os dados da tabela *cliente_produto* foram exportados, em um arquivo no formato *csv*, para serem analisados em um programa escrito em Java, que utilizou as bibliotecas do WEKA, que possuem o algoritmo *Simple k-means*. O WEKA é uma ferramenta livre, que possui uma opção de análise de *clusters* e suas bibliotecas estão disponíveis no site do desenvolvedor. O código do programa em Java consta no Apêndice A.

Foi formulado o código na linguagem Java onde, basicamente, é feita a leitura do arquivo no tipo de dado *DataSource* que, posteriormente, foi utilizado como parâmetro

para o algoritmo em questão. Neste algoritmo foi definido o *seed*, que é um número aleatório usado para a escolha dos centróides iniciais. As impressões dos resultados foram apresentadas pela função *println*, que é um método que faz parte da classe *system* para imprimir uma informação em *string* na tela.

Para a realização de teste neste algoritmo, foi necessário informar o número *k* de grupos (*cluster*) que se deseja encontrar, mas que ainda não são conhecidos e também o valor de *seed* igual a 10 por ser um padrão da própria ferramenta. A taxa de erro quadrático de cada valor de *k* foi o critério de seleção de escolha do melhor número de agrupamento. O treinamento de agrupamento inicialmente foi realizado com o número mínimo de *clusters*, onde *k*=1 e no máximo 10. Conforme ilustra a Tabela 3, investigou-se a taxa de erro quadrático no resultado de cada *cluster* para ser o critério de seleção de melhor escolha do número de agrupamento.

Tabela 5: Comparativo de taxas de erro quadrático

Número de *cluster*	Taxa de erro quadrático
1	145.6
2	59.15
3	55.86
4	4.49
5	3.74
6	3.56
7	3.14
8	2.19
9	2.16
10	2.11

Fonte: Autoria própria

Após testar as diferentes possibilidades, percebeu-se que com agrupamento de 5 *clusters* resultou no melhor número de *clusters*, devido apresentar o erro quadrático no valor de 3.74. A partir dos resultados de erros quadráticos com agrupamentos de 6, 7, 8, 9 e 10, nota-se uma diferença mínima entre eles. É importante destacar que

agrupamentos com 10 *clusters*, com o menor valor de erro quadrático, não é ideal, pois aumenta a possibilidade de gerar vários clusters com apenas 1% dos membros. No Apêndice B estão detalhados os parâmetros utilizados no resultado deste algoritmo e o extrato dos dados coletados transformados para a mineração, que são os resultados obtidos no WEKA.

O algoritmo *Simple k-means* dividiu o grupo de postagens em 5 agrupamentos distintos, descritos como grupos de 0 a 4. Para a composição desses agrupamentos, o algoritmo definiu quatro conjuntos de dados formados por um padrão, que é baseado em conjuntos que se formaram de acordo com intervalo de produtos/serviços, sendo a faixa A com produtos ou serviços de P1 à P60, a faixa B de P61 à P120, a faixa C de P121 à P180 e a faixa D de P181 à P240, conforme as análises descritas abaixo e representadas na Figura 9:

- Agrupamento 0, com 1% dos membros que postaram produtos ou serviços que compõem as faixas A e B.
- Agrupamento 1, com 44% dos membros que postaram produtos ou serviços que compõem a faixa A.
- Agrupamento 2, com 36% dos membros que postaram produtos ou serviços que compõem as faixas B e D.
- Agrupamento 3, com 17% dos membros que postaram produtos ou serviços que compõem as faixas A e D.
- Agrupamento 4, com 1% dos membros que postaram produtos ou serviços que compõem as faixas A, B e D.

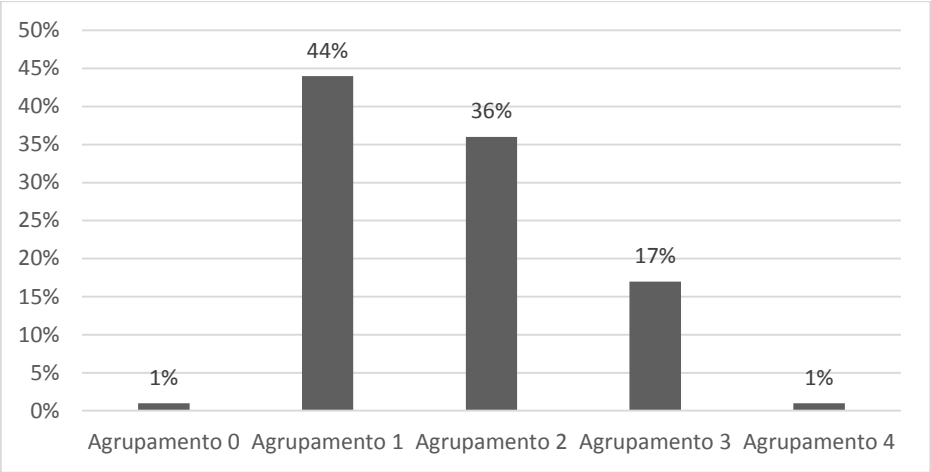

Figura 9: Resultado dos agrupamentos utilizando o algoritmo *k*-médias
Fonte: Autoria própria

Estes agrupamentos são considerados válidos pois são coesos. A coesão avalia o nível de similaridade entre os exemplos do grupo (PAIVA, 2014). Dessa forma, estes foram os principais produtos e serviços que se destacaram ao analisar cada grupo e ajudaram na definição do agrupamento.

A figura 10 visualiza as atribuições dos agrupamentos encontrados:

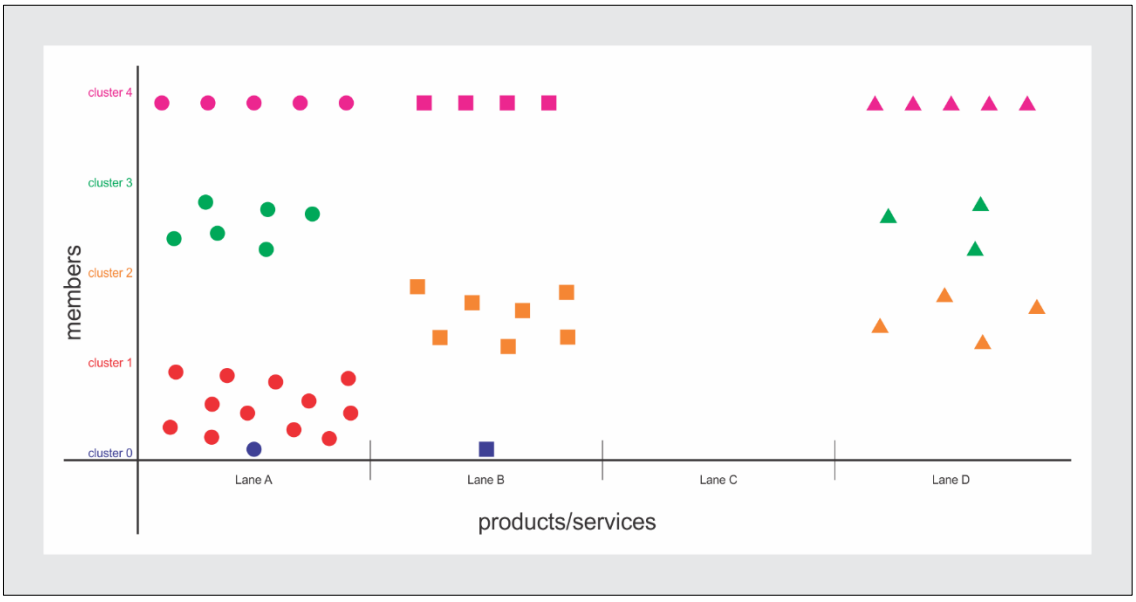

Figura 10: Representação dos agrupamentos.
Fonte: Autoria própria

Por esta análise, ilustrada na Figura 10, detectou-se uma tendência sob as postagens dos membros do Tocantins digital pelos produtos ou serviços dos agrupamentos 1, 2 e 3. Os resultados dos agrupamentos 0 e 4 apresentaram um índice baixíssimo de apenas 1% dos membros, sendo o agrupamento 4 com um membro que mais publicou produtos ou serviços nas faixas A, B e D. Na faixa C não houve publicações referentes aos produtos ou serviços deste subgrupo.

3.6 MEDIDA DE CENTRALIDADE DE INTERMEDIAÇÃO

Para quantificar as ligações do grafo que são consideradas mais importantes entre membros e entre membros e produtos, foi utilizada a medida centralidade de intermediação. Dessa forma, é possível destacar qual o vértice está mais conectado, sendo que este vértice pode ser um membro do subgrupo que realiza mais postagens, assim como um produto ou serviço que são mais visualizados.

Para o início dessa análise, também foram utilizados os dados da tabela *cliente_produto* para identificar os vértices e as distâncias (conexões) entre os vértices. A construção do grafo é realizado pela utilização das variáveis listadas na Tabela 4:

Tabela 6: Definições das variáveis para construção do grafo

Variável	Tipo	Função
Distinct_Vertex	LinkedList<String>	Todos os vértices do grafo
Source_Vertex	LinkedList<String>	Vértice de origem
Target_Vertex	LinkedList<String>	Vértice de destino
Edge_Weight	LinkedList<Double>	Distância da aresta entre um vértice e outro

Fonte: Autoria própria

Os vértices foram adicionados por meio da variável *Distinct_Vertex* de forma a identificar 72 vértices correspondentes aos atores (membros) e 240 vértices correspondentes às postagens sobre produtos ou serviços. Como resultado obteve-se 110 arestas. Os pontos nulos, que representam os atores que não realizaram algum tipo de postagem relacionada aos produtos ou serviços, foram removidos, bem como aqueles produtos ou serviços que em nenhum momento foi ofertado nas postagens desses membros.

Após a identificação dos vértices, foram adicionadas as ligações dos vértices de forma a identificar as variáveis: vértice de origem (*Source_Vertex*), vértice destino (*Target_Vertex*) e aresta que mede a distância entre os vértices (*Edge_Weight*). Estas variáveis foram passadas como parâmetros para a função *BetweenNess_Centrality_Score*, que é a responsável pela construção do grafo do tipo *UndirectedSparseGraph*. Em seguida, o cálculo da centralidade é realizado pela função *BetweennessCentrality*, que faz parte do pacote *edu.uci.ics.jung.algorithms.scoring* da biblioteca JUNG (*Java Universal Network/Graph Framework*), disponível na versão 2.0 no site do fabricante (http://jung.sourceforge.net/). O pacote *edu.uci.ics.jung.algorithms.scoring* possui mecanismos para a atribuição de valores, onde denotam importância, influência, centralidade, etc., para elementos com base em propriedades topológicas do grafo (JUNG, 2009). Por meio da visualização *VisualizationViewer*, disponível na biblioteca JUNG, juntamente com a API JFrame do Java foi possível desenhar o grafo construído, ilustrado na Figura 11. O código do programa utilizando a centralidade de intermediação consta no Apêndice C.

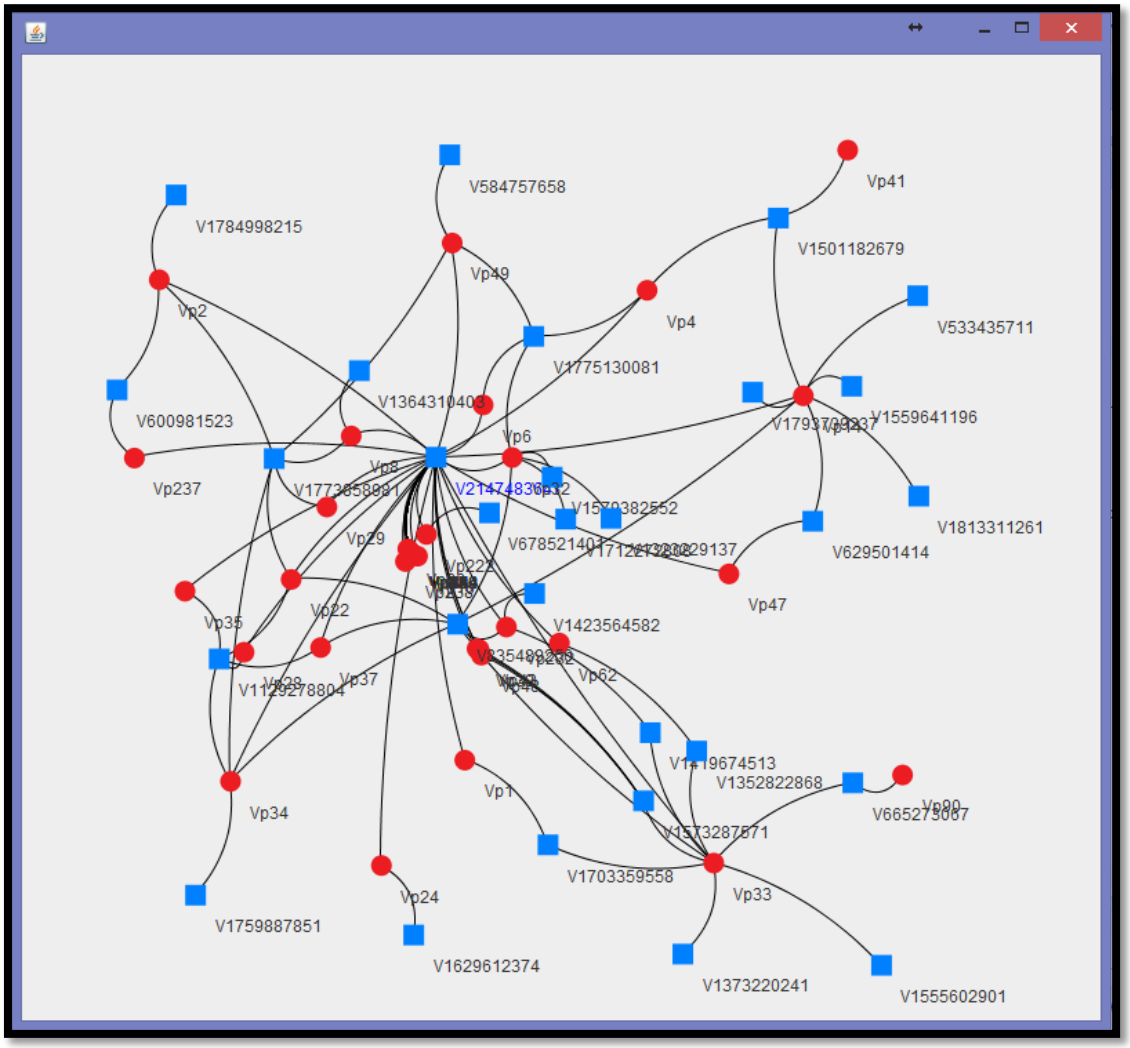

Figura 11: Representação do grafo utilizando a centralidade de intermediação.
Fonte: Autoria própria

No grafo que pode ser observado na Figura 11, os estilos dos vértices no formato quadrado são representados pelos atores (membros) e no formato de elipse, pelos produtos ou serviços. Quanto mais arestas saem de um vértice de origem no formato quadrado, mais publicações estão sendo postadas por este ator. Quanto mais arestas saem de um vértice de origem no formato elipse, mais vezes um tipo de produto ou serviço é postado por vários membros.

A Figura 12 ilustra uma ampliação de parte da Figura 11, onde é possível destacar o vértice V2147483647 com maior centralidade de intermediação.

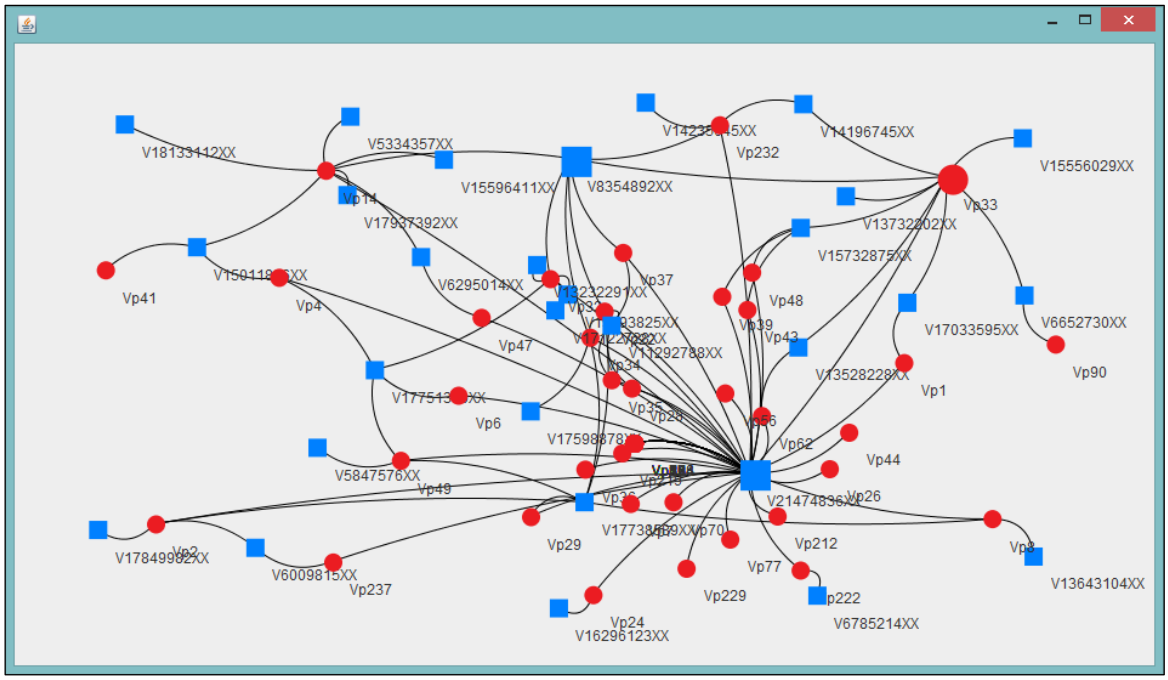

Figura 12: Imagem ampliada dos vértices mais influentes do grafo
Fonte: Autoria própria

O resultado do algoritmo que realiza o cálculo da medida da centralidade de intermediação é apresentado na Tabela 5, sendo que os vértices identificados com as iniciais “Vp” correspondem a algum produto ou serviço postado no classificado do subgrupo Tocantins digital. Os demais vértices são os membros identificados pelos números de seus IDs.

Tabela 7: Resultado da medida de centralidade de intermediação

Identificação de vértices	Vértice do Grafo	Centralidade de intermediação
1	V21474836XX	1785,84
2	V8354892XX	1110,65
3	V17738589XX	768,49
4	V15732875XX	188,70
5	V11292788XX	142,30
6	V17751300XX	103,32
7	V6652730XX	68,12
8	V17033595XX	60,73
9	V6009815XX	38,76
10	V15011826XX	34,30
11	V6295014XX	26,59
12	Vp22	789,80
13	Vp33	644,56
14	Vp29	513,04
15	Vp14	466,04
16	Vp32	198,95
17	Vp49	155,23
18	Vp232	122,25
19	Vp2	109,70
20	Vp4	75,13
21	Vp8	72,08
22	Vp24	68,86
23	Vp222	68,86
24	Vp34	67,05
25	Vp237	61,50
26	Vp37	45,81
27	Vp62	20,61

Fonte: Autoria própria

Avaliar as centralidades de produtos e serviços no subgrupo, oferece importantes informações nas tomadas de decisões para empresas desse setor. Os vértices representados por usuários apresentam maior capacidade sobre suas relações e tendem a estar melhores posicionados dentro de um grupo. Se destacam dentro desse subgrupo analisado os vértices V21474836XX, V8354892XX e V17738589XX que correspondem aos membros do subgrupo que possuem maior centralidade de intermediação. Os demais vértices que representam produtos e serviços que apresentam maior centralidade de intermediação são: Vp22, Vp33, Vp29 e Vp14. Os vértices que não foram elencados na Tabela 5 possuem valores de centralidade de intermediação igual a zero.

4. CONCLUSÃO

Os trabalhos realizados no ambiente de redes sociais possuem um grande volume de dados e, dessa forma, utilizam os processos da mineração de dados aplicando-se grafos, tendo em vista que são melhores representados pelos vértices, que correspondem aos objetos e suas arestas que formam os relacionamentos. A difusão de trabalhos de pesquisas na área de redes sociais demonstram o grande valor que as informações desse ambiente possui. É crescente também as formas de extração de dados que são proporcionadas pelo próprio ambiente da rede social.

Os objetivos desse trabalho foram alcançados com a identificação de um número de agrupamentos formados pela similaridade de produtos e serviços. A fim de otimizar o problema do agrupamento na busca do melhor conjunto de centros de grupos com o algoritmo *k*-médias, percebeu-se que quanto menor o número de agrupamentos, menor a semelhança entre os vértices. A partir disso, os cinco agrupamentos encontrados possibilitaram observar a tendência dos principais produtos e serviços ofertados das postagens realizadas pelos membros do Tocantins digital.

Em seguida, computados os parâmetros da centralidade de intermediação, foi possível construir uma topologia de rede com o objetivo de estudar o comportamento dos membros do subgrupo Tocantins digital baseado nos interesses que seus membros possuem ao visualizar os *posts* dos classificados. A visualização por meio do grafo permitiu destacar graficamente o comportamento dos vértices representados pelos membros do subgrupo Tocantins digital e os vértices representados pelos produtos e serviços mais visualizados.

As informações são advindas das postagens dos membros, que formam o conjunto de transações para estabelecer os relacionamentos entre os vértices. Dessa forma a intermediação medida sobre os vértices do subgrupo Tocantins Digital revelou que existem vértices em melhor posicionamento para transmitir e receber informações. Os vértices V21474836XX, V8354892XX e V17738589XX representam os membros mais influentes do subgrupo, que realizam mais postagens. Os vértices Vp22, Vp33, Vp29 e Vp14 correspondem aos produtos ou serviços mais visualizados pelos membros nos classificados de ofertas.

Com a organização dos dados, foi possível enfatizar a importância das pesquisas realizadas em um ambiente de rede social *online*, de forma a proporcionar de forma automatizada meios nos quais os gestores obtêm maiores informações para o planejamento de uma publicidade. Nota-se também que as interações entre os membros são geradas exclusivamente quando um produto ou serviço é anunciado no *post.* Através da identificação de quais produtos e serviços são mais visitados e das interações com maior fluxo entre os vértices, é possível otimizar um sistema de publicidade aos membros da rede social, que agora são conhecidos por meio dos seus interesses.

Os membros de uma rede social são potenciais futuros consumidores. Esses consumidores tendem a buscar produtos e serviços para atender às suas especificidades pessoais. Portanto, é interessante apontar que atualmente existe uma nova configuração na relação de consumo, mediado pela hipersegmentação, que resulta em consumidores que passam a ser também fornecedores de algum tipo de produto ou serviço, ou seja, podem ser fomentadas por indivíduos ou grupos com poder de liderança, que articulam pessoas em torno dos mesmos interesses.

Atualmente não existem mais grupos isolados de consumidores e sim uma rede de comunicação, onde as pessoas entram em contato com outras através das redes sociais, compartilhando opiniões por meio de uma comunicação rápida.

As publicidades frequentes e repetitivas sem critérios de escolha de clientes passarão por uma análise de acordo com o comportamento dos consumidores, de forma a atingir apenas aqueles interessados em produtos ou serviços de seu interesse. O excesso de propaganda é desnecessário, pois desperdiça o tempo de quem divulga aleatoriamente, e incomoda os consumidores que não desejam aquela oferta.

4.1 CONTRIBUIÇÕES

É importante ressaltar que o entendimento do comportamento do fluxo de uma rede constitui-se em um aporte relevante para gestores que buscam na publicidade, meios de atingir diretamente aos interesses dos consumidores e proporciona uma forma de otimizar alternativas de análise dos perfis entre pessoas que participam de grupos fechados ou abertos nas redes sociais.

Os resultados desta pesquisa proporcionam aos gestores um meio de estar sempre atualizado sobre as preferências que usuários das redes sociais possuem acerca dos produtos ofertados. Isso demonstra que os gestores não podem ficar limitados ao site da própria empresa para saber da opinião e até mesmo para divulgação de seus produtos e serviços. As redes sociais permitem uma abertura direta com o consumidor, pois as empresas podem criar sua própria página da rede social para ficarem mais próximas do

cliente, possibilitando a realização de mineração de dados em outros grupos já constituídos.

Os instrumentos de análise de redes sociais permitem revelar, por meio das ofertas de produtos e serviços, tanto no que se refere ao profissional informal, quanto às empresas, a possibilidade de troca de experiências entre outros grupos da mesma rede social pesquisada, assim como uma ampliação de divulgação.

No campo de análise de redes sociais, o particionamento de grafos serviu de ferramenta para análise, modelagem, predição e evolução dessas redes, sendo aplicado aos ramos de negócios, análise de mercado, marketing, rede de infra-estrutura, relacionamentos, comunicação, entre outras (EISEN et al, 1998).

4.2 LIMITAÇÕES

O grupo Tocantins Digital, até a data de conclusão deste trabalho, estava com cerca de 10000 membros cadastrados. Os dados, como localidade, autor, a serem obtidos por meio dos conteúdos dos *posts*, eram obtidos somente na condição de amigos da pesquisadora. Dessa forma, não foi possível identificar agrupamentos pela localidade dos membros do Tocantins digital, pois o Facebook limita a adição de amigos num curto prazo de tempo e, em certos casos, algumas pessoas não aceitaram o convite.

O Facebook, com a preocupação na segurança das informações, apresenta limitação de acesso aos dados de seus usuários por questões de privacidade. A extração se torna mais limitada para os usuários que não são administradores de um subgrupo.

Limita-se também a quantidade de consultas ao banco de dados para não sobrecarregar o servidor do Facebook.

4.3 TRABALHOS FUTUROS

Com base no referencial teórico e técnicas realizados, é possível dar continuidade no estudo de comunidades aplicando metaheurísticas e fazer uma análise comparativa com o resultado do *k*-médias para a solução de problemas de agrupamentos. Da mesma forma propõe-se também a utilização do processo de descoberta de conhecimento (KDD) utilizando os algoritmos *k*-médias e a centralidade de intermediação em outras redes sociais para detecção de suas comunidades.

As técnicas abordadas neste trabalho, podem servir em trabalhos de pesquisa que busquem padrões de preferência de produtos e serviços nos setores de comércio e varejo, mas também daqueles simples usuários que ofertam esporadicamente algum tipo de produto.

5. REFERÊNCIAS BIBLIOGRÁFICAS

AGUIAR, S. **Redes sociais na internet:** desafios à pesquisa. In: Congresso Brasileiro de Ciências da Comunicação, 30, Santos. Intercom – Sociedade Brasileira de Estudos Interdisciplinares da Comunicação, Santos: UFF, 2007.

AL-FAYOUMI, M.; BANERJEE JR, S.; MAHANTI, P. K. **Analysis of Social Network Using Clever Ant Colony Metaphor**. Proceedings of World Academy of Science, Engineering and Technology, v. 41, mai. 2009. Disponível em: <http://mmujallid.kau.edu.sa/Files/611/Researches/54202_24612.pdf> Acesso em: 27 dez. 2013,

ALBUQUERQUE, D. W. et al. **Estudo do uso do Twitter como Ferramenta de Análise de Opinião durante as Eleições Municipais de João Pessoa.** In: Brazilian Workshop on Social Network Analysis and Mining - Brasnam, 2, 2013. Maceió. Disponível em: < https://docs.google.com/a/unirg.edu.br/viewer?a=v&pid=sites&srcid= ZGVmYXVsdGRvbWFpbnxicmFzbmFtMjAxM3xneDo5OTM4MThmZDI4NjgwZTk >. Acesso em: 19 abr. 2014.

ALMEIDA, L.J. **Detecção de comunidades em redes complexas utilizando estratégia multinível**. 2009. Dissertação Mestrado em Ciências de Computação e Matemática Computacional. Instituto de Ciências e Matemáticas e de Computação, Universidade de São Paulo, São Carlos, 2009.

ANDERY, G. F. **Integrando projeções multidimensionais à análise visual de redes sociais**. 2010. Dissertação. Mestrado em Ciências de Computação e Matemática Computacional.- Instituto de Ciências Matemáticas e de Computação ICMC-USP, São Carlos, 2010.

AQUINO, A.; BRITO, A. **Estudo da Viabilidade do Uso do Facebook para Educação**. In: Congresso da Sociedade brasileira de Computação, 32, Curitiba, 2012,. Disponível em: < http://www.imago.ufpr.br/csbc2012/anais_csbc/eventos/wei/artigos/ Estudo%20da%20Viabilidade%20do%20Uso%20do%20Facebook%20para%20Educa %C3%A7%C3%A3o.pdf> Acesso em: 02 dez. 2013.

BARABÁSI, A. L. **Linked**: how everything is connected to everything else and what it means for business, science, and everyday life. Plume, 2003.

BARABÁSI, A. L. et al. Evolution of the social network of scientific collaborations. **Physica A: Statistical Mechanics and its Applications**, v. 311, p. 590-614, 2002. Disponível em: < http://arxiv.org/pdf/condmat/0104162.pdf>. Acesso em: 23 dez. 2013.

BARBOSA, D. A. B. L. et al. **Medidas de centralidade e detecção de comunidades em rede de co-autoria**. In: Simpósio Brasileiro de Pesquisa Operacional, 43, 2011, Ubatuba. Disponível em:< http://www.polinize.com.br/media/contents/Simp%C3%B3sio.pdf> Acesso em: 20 abr. 2014.

BARRAT, A.; BARTHÉLEMI, M.; VESPIGNANI, A. **Dynamical Processes on Complex Network.** New York: Cambridge University Press, 2008.

BENEVENUTO, F.; ALMEIDA, V.; PEREIRA, A.; ALMEIDA, J.; RODRIGUES, T.; GONÇALVES, M.. **Avaliação do perfil de acesso e navegação de usuários em**

ambientes web de compartilhamento de vídeos. In: Brazilian Symposium on Multimedia Systems and Web (WebMedia). 2009. p. 149-156.

BENEVENUTO, F.; ALMEIDA, J. M.; SILVA, A. S. **Explorando redes sociais online: Da coleta e análise de grandes bases de dados às aplicações**. Mini-cursos do Simpósio Brasileiro de Redes de Computadores (SBRC), 2011.

BISHOP, C. M. **Neural networks for pattern recognition**. New York: Oxford University Press, 482p, 1995.

BOTELHO, G. M.; SOUSA, F. B. **Detecção de comunidades em redes complexas**. 2011. Dissertação de Mestrado. Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, 2011.

BRAGA, L. P. V. **Introdução à Mineração de Dados**. 2. ed. Rio de Janeiro: E-papers, 2005.

BROOKSHEAR, J. G. **Ciência da Computação:** uma visão abrangente. 7. ed. São Paulo: Bookman, 2005.

CAVALCANTI, A. L. et al. **Um Modelo Híbrido LTM-ICM para Difusão de Influência em Redes Sociais**. In: Brazilian Conference on Intelligent System, BRACIS, 2012, Curitiba. Disponível em: <http://www.ppgia.pucpr.br/~enia /anais/wti/artigos/105280.pdf>. Acesso em: 27 dez. 2013.

CHENG, X; DALE, C; LIU, J. **Statistics and social network of YouTube videos**. In Int'l Workshop on Quality of Service (IWQoS), 2008.

COELHO FILHO, O. P.; MARTINHON, C. A.; CABRAL, L. A. F. **Uma abordagem melhorada do algoritmo de otimização por enxame de partículas para o problema de clusterização de dados**. In: Simpósio Brasileiro de Pesquisa Operacional, 45, 2013, Natal. 2013.

CORNACCHIONE JR., E. B. **Informática Aplicada às Áreas de Contabilidade, Administração e Economia.** 4 ed. São Paulo: Atlas, 2012.

CROSS, R.; PARKER, A.; BORGATTI, S. P. **A bird's-eye view: using social network analysis to improve knowledge creation and sharing**. Knowledge Directions, v.2, n.1, p.48-61, 2000. Disponível em: http://www.analytictech.com/borgatti/publications.htm. Acesso em 28 jul. 2004.

DATE, C. J. **Introdução a Sistemas de Banco de Dados**. 8 ed. Rio de Janeiro, Campus, 2004.

DAVENPORT, Thomas H.; PRUSAK, Laurence. **Conhecimento empresarial:** como as organizações gerenciam o seu capital intelectual. 6 ed. Rio de Janeiro: Campus, 1999.

DUBES, R.; JAIN, A. K. **Clustering and methodologies in exploratory data analysis**. In: Advances inComputers, v.19. NewYork: AcademicPress, 1980.

EASLEY, D.; KLEINBERK, J. **Networks crowds and markets**. Reasoning about a Highly Connected World. Cambridge: Cambridge University Press, 2010.

EMARKETER. **Facebook continues to lead social network usage in Brazil**. 6 dezembro 2013. Disponível em: < http://www.emarketer.com/Article/Badoo-Becomes-No-3-Social-Network-Brazil/1010436>. Acesso em: 23 dez. 2013.

EISEN, M. B.; SPELLMAN, P. T.; BROWN, P. O.; BOTSTEIN, D. **Cluster analysis and display of genome-wide expression patterns**. Proceeding of the National Academy of Sciences of the United States of America.(1998).

EVERITT, B. **Cluster Analysis**. Heinemann Educational Books, London, 1974.

FACEBOOK. **Developers**: The Graph API. Disponível em: http://www.facebook.com. Acesso em: 04 mar. 2013.

FAYYAD, U. M.; PIATETSKY-SHAPIRO, G.; SMYTH, P. **From data mining to knowledge discovery in databases**. California: AAAI/The MIT, 1996. Disponível em: <http://www.kdnuggets.com/gpspubs/aimag-kdd-overview-1996-Fayyad.pdf>. Acesso em: 24 out. 2013.

FERREIRA, L. N. **Técnica de agrupamento de dados baseada em redes complexas para o posicionamento de cluster heads em rede de sensores sem fio**.2012. Dissertação (Mestrado em Ciências de Computação e Matemática Computacional). Instituto de Ciências e Matemáticas e de Computação – ICMC-USP. 2012.

FERREIRA, T. F. P. Redes Sociais e Classificação Conceptual: Abordagem Complementar para um Sistema de Recomendação de Coautorias. Dissertação (Mestrado em Análise de Dados e Sistemas de Apoio à Decisão) – Universidade de Porto, 2012. Disponível em: http://repositorio-aberto.up.pt/bitstream/10216/70310/2/13987.pdf. Acesso em: 04 ago. 2014.

FLECHA, A. C.; BERNARDES, A. T.; SILVA, A. V. C. C. Medidas de centralidade como parâmetros para se avaliar os atores da rede de turismo: o caso da cidade de Ouro Preto. In: Simpósio de Administração da Produção, Logística e Operações Internacionais, 14, 2011, São Paulo. Anais. São Paulo: FGV, EAESP, 2011.

FREITAS C. M. D. S. *et al.* **Extração de Conhecimento e Análise Visual de Redes Sociais.** In: Simpósio Brasileiro de Computação, 28, Belém, Anais do XXVIII Congresso da SBC.. Belém do Pará: SEMISH, 2008. Disponível em: < http://www.lbd.dcc.ufmg.br:8080/colecoes/semish/2008/008.pdf >. Acesso em: 12 dez. 2013.

FOROUZAN, B.; MOSHARRAF, F. **Fundamentos da Ciência da Computação**. 2 ed. São Paulo: Cengage Learning, 2011.

FORTUNATO, S. **Community detection in graphs**. In: Complex Networks and Systems Lagrange Laboratory, ISI Foundation, 10133, Torino, 2010. Disponível em:< http://arxiv.org/pdf/0906.0612.pdf>. Acesso em: 24 dez. 2013.

FREEMAN, L. C. **A set of measures of centrality based on betweenness**. Sociometry, v. 40, n.1, 1977.

_____. **Centrality in social networks**: I. conceptual clarification. Social Networks, v. 1, p. 215-239, 1979.

_____. **Some antecedents of social network analysis**. Connections, v. 19, n. 1, p. 39-42, 1996.

GHOSH, A.; DEHURI, S.; GHOSH, S. **Multi-Objective Evolutionary Algorithms for Knowledge Discovery from Databases**. Springer, 2008.

GIRVAN, M.; NEWMAN, M. E. Community structure in social and biological networks. Proceedings of the National Academy of Sciences of the United States of America, v. 99, n. 12, 2002.

GOMES, F.T.; PARDO, T.A.S. **Classificação e agrupamento de textos para processamento multidocumento**. In: The Proceedings of the STIL Student Workshop on Information and Human Language Technology, 2009, São Carlos. Disponível em: <http://www.icmc.usp.br/pessoas/taspardo/TILic2009-GomesPardo.pdf>. Acesso em: 26 dez. 2013.

GOODRICH, M. T.; TAMASSIA, R. **Estruturas de dados e algoritmos em Java**. 4 ed. Porto Alegre: Bookman, 2007.

GULINI, A. S.; MISAGHI, M. **Redes Sociais Corporativas**: Avaliação da Centralidade de Grau Através de Índices Como Ferramenta de Gestão. Revista Eletrônica Produção em Foco, v. 2, n. 1, 2012.

HANNE, M. *et al*. **Analysis of vulnerability to facebook users**. In: Proceedings of the 18th Brazilian symposium on Multimedia and the web. ACM, 2012. p. 335-342.

HANNEMAN, R. **Introduction to Social Network Methods**. 2000. Disponível em: <http://faculty.ucr.edu/~hanneman/>. Acesso em: 12 dez. 2013.

HAUS, G. L. **Identificação de tráfego Bittorrent com fins periciais utilizando Weka**. 2012. 56f. Dissertação (Mestrado em Engenharia Elétrica) Departamento de Engenharia Elétrica da Faculdade de Tecnologia da Universidade de Brasília, Brasília, 2012.

JAIN, A. K.; DUIN, R. P. W.; MAO, J. **Statistical pattern recognition: a review.** IEEE Transactions on Pattern Analysis and Machine Intelligence, 2000.

JOHNSON, R. A.; WICHERN, D. W. **Applied Multivariate Statistical Analysis**. Prentice Hall, New Jersey, 1998.

JUNG Java Universal Network/Graph Framework. Libraries. 2009. Disponível em: < http://jung.sourceforge.net/download.html >. Acesso em: 04 mar. 2014.

KIRKPATRICK, D. **Facebook's plan to hook up the world**. CNN Money, 29 maio 2007. Disponível em: < http://money.cnn.com/2007/05/24/technology/facebook.fortune/ >. Acesso em: 24 dez. 2013.

LÉVY, Pierre. **A inteligência coletiva**: por uma antropologia do ciberespaço. São Paulo: Loyola, 1998.

LIN, Fu-ren; CHEN, Chun-Hung; TSAI, Kuo-Lung. **Discovering group interaction patterns in a teachers professional community**. In: System Sciences, 2003. Proceedings of the 36th Annual Hawaii International Conference on. IEEE, 2003. p. 8 pp.

LIANG, T. P.; HO, Y. T.; LI, Y-W.; TURBAN, E. **What drives social commerce**: the role of social support and relationship quality. International Journal of Electronic Commerce. Winter, 2011.

LOPES, T. J. P. L.; HIRATANI, G. K. L. **Mineração de opiniões e fatos aplicada à análise de investimentos.** São Paulo Senac, 2008.

MACQUEEN, J. **Some methods for classification and analysis of multivariate observations.** 1967**.** Disponível em: <www-m9.ma.tum.de/foswiki/pub/WS2010/ CombOptSem/kMeans.pdf>. Acesso em: 12 dez 2013.

MADEIRA, N. L. F. **A construção do conhecimento sobre o território europeu: análise da dinâmica de cooperação no programa ESPON**. Dissertação (Mestrado em Geografia) - Instituto de Geografia e Ordenamento do Território da Universidade de Lisboa, Lisboa, 2010.

MARTELETO, R. M. **Análise de redes sociais**. Aplicação nos estudos de transferência da informação. Ci. Inf., Brasília, v. 30, n. 1, p. 71-81, jan./abr. 2001.

MARTELETO, R. M.; SILVA, A. B. O. **Redes e capital social**: o enfoque da informação para o desenvolvimento local. Ciência da Informação, v.33, n.3, p.41-49, 2004.

MIGUEL, P. A. C, et al. **Metodologia de Pesquisa em Engenharia de Produção e Gestão de Operações**. 2 ed. Campus: São Paulo, 2012.

MIKA, P. **Social Networks and the Semantic Web**. Springer, 2007.

MITCHELL, T. M. **Machine learning**. McGraw-Hill Series in Computer Science, McGraw-Hill, 1997.

MONDINI, L. C.; DOMINGUES, M. J. C. S.; CORREIA, R. B. MONDINI, V. E. D. **Redes Sociais Digitais:** uma análise de utilização pelas instituições de ensino superior do sistema ACAFE de Santa Catarina. Revista Eletrônica de Ciência Administrativa, Campo Largo, v. 11, n.1, p. 48-60, jan-jun 2012.

MOKARZEL, F. C.; SOMA, N. Y. **Introdução à Ciência da Computação**. Rio de Janeiro: Elsevier, 2008.

NASCIMENTO, C. S. do. **PANDORA**: uma ferramenta para visualização incremental e análise de redes sociais acadêmicas. Dissertação (Mestrado em Ciência da Computação) – Universidade Federal do Rio Grande do Sul, Porto Alegre, 2013.

NEWMAN, M. E. J. **Fast algorithm for detecting community structure in networks**. Physical Review E, 69:026113, 2004.

_____. **Networks: an Introduction.** Oxford University Press, 2010.

NEWMAN, M. E. J.; GIRVAN, M. **Finding and evaluating community structure in networks.** 2003.

PAIVA, E. R. F. **Detecção de novidade em fluxos contínuos de dados multiclasse**. 2014. Tese (Doutorado em Ciências da Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação de São Carlos, São Carlos, 2014.

PATRÍCIO, M. R. **Tecnologias Web 2.0:** Recursos Pedagógicos na Formação Inicial de Professores. Dissertação (Mestrado em Multimédia) Faculdade de Engenharia da Universidade do Porto, Porto, 2009. Disponível em: < http://hdl.handle.net/10198/1971>. Acesso em: 29 jan. 2014.

PEW INTERNET. How people get local news and information in different communities. Disponível em: <http://www.pewinternet.org/Reports/2012/Communities-and-Local-News.aspx> . Acesso em: 21 fev. 2013.

PIMENTEL, M.; FUKS, H. **Sistemas Colaborativos**. Rio de Janeiro: Campus, 2011.

PINHEIRO, C. A. R. **Web Warehousing**. Extração e Gerenciamento de dados na Internet. Rio de Janeiro: Editora Axcel Books do Brasil, 2003.

PISANI, F.; PIOTET D. **Como a Web transforma o mundo:** a alquimia das multidões. São Paulo: Editora Senac São Paulo, 2010.

POINT TOPIC. Global broadband statistics Q3 2012. 2013. Disponível em: <http://point-topic.com/dslanalysis.php>. Acesso em: 18 fev. 2013.

QUEIROZ, E. M. G. **Redes ópticas multidomínio: métodos de escolha de nós de borda e algoritmo de roteamento de tráfego**. Tese (Doutorado em Ciências) – Escola de Engenharia de São Carlos da Universidade de São Paulo. São Carlos, 2012.

QUEIROZ, T. R. **Esboço de uma rede de cooperação em um arranjo produtivo local na indústria calçadista paulista.** Tese (Doutorado em Engenharia de Produção) – Universidade Federal de São Carlos. São Carlos, 2012.

QUINTEIRO, J. A. T. **Segmentação de indivíduos no Facebook que gostam de música: abordagem exploratória, recorrendo à comparação entre dois algoritmos, k-means e fuzzy c-means**. Dissertação (Mestrado em Gestão)- Universidade de Técnica de Lisboa, Lisboa, 2011.

RECUERO, R. C.. Comunidades em redes sociais na Internet: um estudo de caso dos fotologs brasileiros. Liinc em Revista, Rio de Janeiro, v.4, n.1, p.63-83, mar. 2008.

_____. **Teoria das redes e redes sociais na internet:** considerações sobre o Orkut, os weblogs e os fotologs. In: Congresso Brasileiro de Ciências da Comunicaç ao. 27. INTERCOM. 2004.

REZENDE, S. O. **Mineração de Dados.** In: XVIII Encontro Nacional de Inteligência Artificial, 2005, São Leopoldo. Disponível em: <http://www.lbd.dcc.ufmg.br/colecoes/enia/2005/0102.pdf> Acesso em: 31 out. 2013.

RODRIGUES, D. M. S. **Detecção de comunidades no sistema de correio electrónico universitário.** 2009. Dissertação (Mestrado em Ciência da Complexidade) Instituto Superior de Ciências do Trabalho e da Empresa, 2009.

ROSSONI L.; GUARIDO FILHO, E. R. **Cooperação Interinstitucional no campo da pesquisa em estratégia**. Revista de Administração de Empresa, 2007, v. 47, n.4. Disponível em: <http://www.scielo.br/pdf/rae/v47n4/v47n4a07.pdf>. Acesso em: 24 jan. 2014.

RUSSEL, M. A. **Mineração de Dados na Web Social.** São Paulo: Novatec Editora, 2011.

SMOLKA, R. B. **Redes de cooperação entre EBTs do setor médico-hospitalar da região de São Carlos, Araraquara e Ribeirão Preto**. Dissertação (Mestrado em Engenharia de Produção) – Universidade Federal de São Carlos, São Carlos, 2006

STAUFFER, D.; AHARONY, A.; ADLER, J. **Efficient Hopfield Pattern Recognition on a Scale-Free Neural Network**. The European Physical Journal B, 32:395–399, 2003.

SALDANHA, M.; FREITAS, C. **Segmentação de Imagens Digitais**: Uma Revisão. Divisão de Processamento de Imagens-Instituto Nacional de Pesquisas Espaciais (INPE), São Paulo, 2009.

SANTOS, A. M. **Smart Marketing na TV Digital Interativa através de um sistema de recomendação de anúncios**. Dissertação (Mestrado em Engenharia Elétrica) – Pontifícia Universidade Católica de Campinas, Campinas, 2012.

SANTOS, D. M. B. **Seleção de Modelos de Classificação Através de Heurísticas**. 2005. Dissertação (Mestrado em Informática) – Universidade Federal de Campina Grande, Campina Grande, 2005.

SARTORI, R. **Mineração de dados da Polícia Militar de Santa Catarina no município de Balneário Camboriú para geração de Informação e Conhecimento na área de Segurança Pública.** Trabalho de Conclusão de Curso de Ciência da Computação. Universidade do Vale do Itajaí, Itajaí, 2012.

SIQUEIRA, M. C. **Gestão Estratégica da Informação**. Rio de Janeiro: Brasport, 2005.

STRÖELE, V.; ZIMBRÃO, G.; SOUZA, J. M. **Análise de redes sociais científicas**: modelagem multi-relacional. In: I Brazilian Workshop Soc Netw Anal Min. 2012.

TAN, Pang-Ning; STEINBACH, M.; KUMAR, V. **Introdução ao Datamining Mineração de Dados**. Rio de Janeiro: Editora Ciência Moderna, 2009.

TRINDADE, A. R.; OCHI, L. S. **Um algoritmo evolutivo híbrido para a formação de células de manufatura em sistemas de produção**. Pesquisa Operacional. Rio de Janeiro, v.26, n.2, 255-294, 2006.

TURBAN, E.; MCLEAN, E.; WETHERBE, J. **Tecnologia da Informação para Gestão**. Transformando os negócios na economia digital. 3 ed. Porto Alegre: Bookman, 2004.

VIEIRA, G. **Modelagem matemática-computacional da conectividade cerebral em ressonância magnética funcional para o estudo do estado de repouso.** Dissertação (Mestrado em Ciências) Universidade de São Paulo, São Paulo, 2011.

VELLOSO, F. C. **Informática:** conceitos básicos. 8 ed. Rio de Janeiro: Elsevier, 2011.

WASSERMAN, S.; FAUST, K. **Social network analysis: Method and applications.** Cambridge: Cambridge University Press, 1994.

WATTENBERG, M. **Visual Exploration of Multivariate Graphs**. In Proceedings of Conference Human Factors in Computing Systems (SIGCHI 2006). Montreal, Canada, 2006.

WATTS, D. J.; STROGATZ, S. H. **Collective dynamics of small-world networks**. Nature 393, 1998.

APÊNDICE A
CÓDIGO DO PROGRAMA utilizando algoritmo K-MÉDIAS

```
import eka.clusterers.ClusterEvaluation; import
weka.clusterers.SimpleKMeans; import weka.core.Instance;
import weka.core.Instances;
import weka.core.converters.ConverterUtils.DataSource;

public class CFurlan {

    public Instances dados; public String path;
    public String[] options = new String[2];

    public CFurlan(String caminho, int nclusters, int seed ){
        this.path = caminho;
        this.options[0] = String.valueOf(nclusters); this.options[1] =
        String.valueOf(seed);
    }

    public void ledados() throws Exception{

        DataSource source = new DataSource(path); dados =
        source.getDataSet();
    }


    public void imprimedados(){

        for(int i=0; i<dados.numInstances();i++)
        {
            Instance actual = dados.instance(i);

            for(int y=0; y<1; y++)
            {
                System.out.println((i+1) + " : "+ actual.toString(y));
            }
        } }

    public void clustering() throws Exception{

        SimpleKMeans cluster = new SimpleKMeans();
        cluster.setOptions(options);
        cluster.setNumClusters(this.options[0]);
        cluster.setSeed(this.options[1]);
        cluster.setDisplayStdDevs(true); cluster.getMaxIterations();
        cluster.buildClusterer(dados);
```

```
        Instances ClusterCenter = cluster.getClusterCentroids();
        Instances SDev = cluster.getClusterStandardDevs(); int[]
        ClusterSize = cluster.getClusterSizes(); ClusterEvaluation
        eval = new ClusterEvaluation(); eval.setClusterer(cluster);
        eval.evaluateClusterer(dados);

        for(int i=0;i<ClusterCenter.numInstances();i++){
     System.out.println("Cluster#"+( i +1)+ ": "+ClusterSize[i]+" dados
                                                                    .");
      System.out.println("Centróide:"+ ClusterCenter.instance(i));
            System.out.println("STDDEV:" + SDev.instance(i));
                                          System.out.println("Cluster
                          Evaluation:"+eval.clusterResultsToString());

        } }


    public static void main(String[] args) throws Exception {

        CFurlan cf = new CFurlan("C:\\cliente_produto1.csv",10,10);
        cf.ledados();


        cf.clustering(); }


}
```

APÊNDICE B
RESULTADO DA APLICAÇÃO DO ALGORITMO k-MÉDIAS

```
=== Run information ===
Scheme:weka.clusterers.SimpleKMeans -N 5 -A "weka.core.EuclideanDistance -R first-last" -I 500 -S 10
Relation:     cliente_produto1
Instances:    72
Attributes:   241
[list of attributes omitted]
Test mode:evaluate on training data
=== Model and evaluation on training set ===
kMeans
======
Number of iterations: 3
Within cluster sum of squared errors: 3.7432296892360455
Missing values globally replaced with mean/mode
Cluster centroids:
                                Cluster#
Attribute         Full Data             0              1              2              3              4
                       (72)           (1)           (32)           (26)           (12)           (1)
=============================================================================================================
actor_id    1390772685.6806       665273067 1708349146.7188 1350352653.3077     628877914.5      2147483647
p1                   5.4167               0          0.125              0              0             386
p2                   1.5833               0            0.5              0            0.5              92
p3                        1               0              0              0              0              72
p4                   2.8056               0          0.125         0.0769              0             196
p5                   0.5278               0              0              0              0              38
p6                   0.2778               0           0.25              0              0              12
p7                   0.6667               0              0              0              0              48
p8                   1.4444               0          0.125         0.1538              0              96
p9                        0               0              0              0              0               0
p10                  0.0278               0              0              0              0               2
p11                       0               0              0              0              0               0
p12                       0               0              0              0              0               0
p13                       0               0              0              0              0               0
p14                 11.4167               0         0.1875         0.0769         1.3333             798
p15                  0.6111               0              0              0              0              44
p16                  0.0833               0              0              0              0               6
p17                       0               0              0              0              0               0
p18                       0               0              0              0              0               0
```

p19	0	0	0	0	0	0
p20	0.0556	0	0	0	0	4
p21	0	0	0	0	0	0
p22	2.1389	0	0.125	0.0769	0.1667	146
p23	0.5556	0	0	0	0	40
p24	0.9722	0	0.125	0	0	66
p25	0	0	0	0	0	0
p26	0.1111	0	0	0	0	8
p27	0	0	0	0	0	0
p28	0.4722	0	0	0.0769	0	32
p29	0.1389	0	0.125	0	0	6
p30	0	0	0	0	0	0
p31	0	0	0	0	0	0
p32	3.6944	0	0.625	0.0769	0.3333	240
p33	7.5278	6	0.375	1	0.1667	496
p34	7.25	0	0.1875	0.9231	0.3333	488
p35	6.5833	0	0	0.0769	0	472
p36	5.9722	0	0	0	0	430
p37	0.5	0	0	0.0769	0.1667	32
p38	0	0	0	0	0	0
p39	0.4167	0	0.0625	0	0	28
p40	0.3889	0	0	0	0	28
p41	0.0278	0	0	0.0769	0	0
p42	0.0556	0	0	0	0	4
p43	1.9167	0	0.0625	0	0	136
p44	0.0556	0	0	0	0	4
p45	0	0	0	0	0	0
p46	0	0	0	0	0	0
p47	0.5	0	0	0	1.1667	22
p48	2.5278	0	0.0625	0	0	180
p49	2.2778	0	0.25	0	0.6667	148
p50	0	0	0	0	0	0
p51	0	0	0	0	0	0
p52	0	0	0	0	0	0
p53	0	0	0	0	0	0
p54	0	0	0	0	0	0
p55	0	0	0	0	0	0
p56	0.0278	0	0	0	0	2
p57	0.3889	0	0	0	0	28
p58	0	0	0	0	0	0

p59	0	0	0	0	0	0
p60	0	0	0	0	0	0
p61	0	0	0	0	0	0
p62	0.3056	0	0	0.6923	0	4
p63	0.2222	0	0	0	0	16
p64	0	0	0	0	0	0
p65	0	0	0	0	0	0
p66	0.0556	0	0	0	0	4
p67	0	0	0	0	0	0
p68	0	0	0	0	0	0
p69	0.0556	0	0	0	0	4
p70	0.75	0	0	0	0	54
p71	0	0	0	0	0	0
p72	0	0	0	0	0	0
p73	0.2222	0	0	0	0	16
p74	0	0	0	0	0	0
p75	0	0	0	0	0	0
p76	0	0	0	0	0	0
p77	0.4722	0	0	0	0	34
p78	0	0	0	0	0	0
p79	0	0	0	0	0	0
p80	0	0	0	0	0	0
p81	0	0	0	0	0	0
p82	0	0	0	0	0	0
p83	0	0	0	0	0	0
p84	0	0	0	0	0	0
p85	0	0	0	0	0	0
p86	0.1111	0	0	0	0	8
p87	0	0	0	0	0	0
p88	0	0	0	0	0	0
p89	0	0	0	0	0	0
p90	0.0278	2	0	0	0	0
p91	0	0	0	0	0	0
p92	0	0	0	0	0	0
p93	0	0	0	0	0	0
p94	0	0	0	0	0	0
p95	0	0	0	0	0	0
p96	0	0	0	0	0	0
p97	0	0	0	0	0	0
p98	0	0	0	0	0	0

p99	0	0	0	0	0	0
p100	0	0	0	0	0	0
p101	0	0	0	0	0	0
p102	0	0	0	0	0	0
p103	0	0	0	0	0	0
p104	0	0	0	0	0	0
p105	0	0	0	0	0	0
p106	0	0	0	0	0	0
p107	0	0	0	0	0	0
p108	0	0	0	0	0	0
p109	0	0	0	0	0	0
p110	0	0	0	0	0	0
p111	0	0	0	0	0	0
p112	0	0	0	0	0	0
p113	0	0	0	0	0	0
p114	0	0	0	0	0	0
p115	0	0	0	0	0	0
p116	0	0	0	0	0	0
p117	0	0	0	0	0	0
p118	0	0	0	0	0	0
p119	0	0	0	0	0	0
p120	0	0	0	0	0	0
p121	0	0	0	0	0	0
p122	0	0	0	0	0	0
p123	0	0	0	0	0	0
p124	0	0	0	0	0	0
p125	0	0	0	0	0	0
p126	0	0	0	0	0	0
p127	0	0	0	0	0	0
p128	0	0	0	0	0	0
p129	0	0	0	0	0	0
p130	0	0	0	0	0	0
p131	0	0	0	0	0	0
p132	0	0	0	0	0	0
p133	0	0	0	0	0	0
p134	0	0	0	0	0	0
p135	0	0	0	0	0	0
p136	0	0	0	0	0	0
p137	0	0	0	0	0	0
p138	0	0	0	0	0	0

p139	0	0	0	0	0	0
p140	0	0	0	0	0	0
p141	0	0	0	0	0	0
p142	0	0	0	0	0	0
p143	0	0	0	0	0	0
p144	0	0	0	0	0	0
p145	0	0	0	0	0	0
p146	0	0	0	0	0	0
p147	0	0	0	0	0	0
p148	0	0	0	0	0	0
p149	0	0	0	0	0	0
p150	0	0	0	0	0	0
p151	0	0	0	0	0	0
p152	0	0	0	0	0	0
p153	0	0	0	0	0	0
p154	0	0	0	0	0	0
p155	0	0	0	0	0	0
p156	0	0	0	0	0	0
p157	0	0	0	0	0	0
p158	0	0	0	0	0	0
p159	0	0	0	0	0	0
p160	0	0	0	0	0	0
p161	0	0	0	0	0	0
p162	0	0	0	0	0	0
p163	0	0	0	0	0	0
p164	0	0	0	0	0	0
p165	0	0	0	0	0	0
p166	0	0	0	0	0	0
p167	0	0	0	0	0	0
p168	0	0	0	0	0	0
p169	0	0	0	0	0	0
p170	0	0	0	0	0	0
p171	0	0	0	0	0	0
p172	0	0	0	0	0	0
p173	0	0	0	0	0	0
p174	0	0	0	0	0	0
p175	0	0	0	0	0	0
p176	0	0	0	0	0	0
p177	0	0	0	0	0	0
p178	0	0	0	0	0	0

p179	0	0	0	0	0	0
p180	0	0	0	0	0	0
p181	0	0	0	0	0	0
p182	0	0	0	0	0	0
p183	0	0	0	0	0	0
p184	0	0	0	0	0	0
p185	0	0	0	0	0	0
p186	0	0	0	0	0	0
p187	0	0	0	0	0	0
p188	0	0	0	0	0	0
p189	0	0	0	0	0	0
p190	0	0	0	0	0	0
p191	0	0	0	0	0	0
p192	0	0	0	0	0	0
p193	0	0	0	0	0	0
p194	0	0	0	0	0	0
p195	0	0	0	0	0	0
p196	0	0	0	0	0	0
p197	0	0	0	0	0	0
p198	0	0	0	0	0	0
p199	0	0	0	0	0	0
p200	0	0	0	0	0	0
p201	0	0	0	0	0	0
p202	0	0	0	0	0	0
p203	0	0	0	0	0	0
p204	0	0	0	0	0	0
p205	0	0	0	0	0	0
p206	0	0	0	0	0	0
p207	0	0	0	0	0	0
p208	0	0	0	0	0	0
p209	0	0	0	0	0	0
p210	0	0	0	0	0	0
p211	0	0	0	0	0	0
p212	0.4167	0	0	0	0	30
p213	0	0	0	0	0	0
p214	0.0278	0	0	0	0	2
p215	0	0	0	0	0	0
p216	0	0	0	0	0	0
p217	0	0	0	0	0	0
p218	0	0	0	0	0	0

p219	0.0278	0	0	0	0	2
p220	0	0	0	0	0	0
p221	0	0	0	0	0	0
p222	0.0833	0	0	0	0.1667	4
p223	0	0	0	0	0	0
p224	0.1667	0	0	0	0	12
p225	0	0	0	0	0	0
p226	0.4444	0	0	0	0	32
p227	0	0	0	0	0	0
p228	0	0	0	0	0	0
p229	1.6667	0	0	0	0	120
p230	0	0	0	0	0	0
p231	0	0	0	0	0	0
p232	0.5556	0	0	0.1538	0.1667	34
p233	0	0	0	0	0	0
p234	0	0	0	0	0	0
p235	0	0	0	0	0	0
p236	0	0	0	0	0	0
p237	0.1944	0	0	0	0.1667	12
p238	0.0556	0	0	0	0	4
p239	0	0	0	0	0	0
p240	0	0	0	0	0	0

Time taken to build model (full training data) : 0.07 seconds

=== Model and evaluation on training set ===

Clustered Instances

0	1	(1%)
1	32	(44%)
2	26	(36%)
3	12	(17%)
4	1	(1%)

APÊNDICE C
CÓDIGO DO PROGRAMA DA CENTRALIDADE DE INTERMEDIAÇÃO

```
import java.awt.Color;
import java.awt.Dimension;
import java.awt.Shape;
import java.awt.geom.Ellipse2D;
import ava.awt.geom.Rectangle2D;
import java.awt.geom.Point2D;
import java.util.Collection;
import java.util.Hashtable;
import java.util.Iterator;
import java.util.LinkedList;
import javax.swing.JFrame;
import org.apache.commons.collections15.Transformer;
import weka.core.Instance;
import weka.core.Instances;
import weka.core.converters.ConverterUtils.DataSource;
import edu.uci.ics.jung.algorithms.layout.ISOMLayout;
import edu.uci.ics.jung.algorithms.layout.Layout;
import edu.uci.ics.jung.graph.UndirectedSparseGraph;
import edu.uci.ics.jung.graph.Graph;
import edu.uci.ics.jung.graph.util.EdgeType;
import edu.uci.ics.jung.visualization.RenderContext;
import edu.uci.ics.jung.visualization.VisualizationViewer;
import du.uci.ics.jung.visualization.control.DefaultModalGraphMouse;
import edu.uci.ics.jung.visualization.control.ModalGraphMouse;
import edu.uci.ics.jung.visualization.renderers.Renderer;
import edu.uci.ics.jung.visualization.transform.shape.GraphicsDecorator;

public class Graph_Algos {

    public Instances dados; public String path;
    public String[] options = new String[2];
    private Graph<MyNode, MyLink> g = new
    UndirectedSparseGraph<Graph_Algos.MyNode, Graph_Algos.MyLink>();

    static int edgeCount_Directed = 0; // This works with the inner
    MyEdge class

    public Graph_Algos(String caminho) {
      this.path = caminho; }
```

```
class MyNode {
    // static int edgeCount = 0; // This works with the inner
    MyEdge class

    String id;
    public MyNode(String id) {this.id = id;}

    public String toString()
    {return "V" + id;}

}

class MyLink
{
    double weight;
    int id;

    public MyLink(double weight)
    { this.id = edgeCount_Directed++;

        this.weight = weight;
    }

    public String toString()
    {return "E" + id;}

}

//used to construct graph and call graph algorithm used in JUNG
public void BetweenNess_Centrality_Score(LinkedList<String>
Distinct_nodes, LinkedList<
String> source_vertex, LinkedList<String> target_vertex,
LinkedList<Double> Edge_Weight) {

    //CREATING weighted directed graph

    Graph<MyNode, MyLink> g = new
    UndirectedSparseGraph<Graph_Algos.MyNode, Graph_Algos.
    MyLink>();

    //create node objects
    Hashtable<String, MyNode> Graph_Nodes = new Hashtable<String,
    Graph_Algos.MyNode>();

    LinkedList<MyNode> Source_Node = new
    LinkedList<Graph_Algos.MyNode>();

    LinkedList<MyNode> Target_Node = new
    LinkedList<Graph_Algos.MyNode>();
```

```
LinkedList<MyNode> Graph_Nodes_Only = new
LinkedList<Graph_Algos.MyNode>();


//create graph nodes
for(int i=0;i<Distinct_nodes.size();i++)
{
    String node_name = Distinct_nodes.get(i);
    MyNode data = new MyNode(node_name);
    Graph_Nodes.put(node_name, data);
    Graph_Nodes_Only.add(data);
}

//Now convert all source and target nodes into objects
System.out.println(source_vertex.size()+ " - " +
target_vertex.size() + " - " + Edge_Weight.size() + " - "+
Graph_Nodes_Only.size());
for(int t=0;t<source_vertex.size();t++)
{
    Source_Node.add(Graph_Nodes.get(source_vertex.get(t)));
    Target_Node.add(Graph_Nodes.get(target_vertex.get(t)));
}

//Now add nodes and edges to the graph for(int
i=0;i<Edge_Weight.size();i++)
{
    g.addEdge(new
    MyLink(Edge_Weight.get(i)),Source_Node.get(i),
    Target_Node.get(i), EdgeType.UNDIRECTED);
}

Transformer<MyLink, Double> wtTransformer = new
Transformer<MyLink,Double>()
{ public Double transform(MyLink link) { return link.weight;
    } };

edu.uci.ics.jung.algorithms.scoring.BetweennessCentrality<MyN
ode,MyLink> BC1 = new
edu.uci.ics.jung.algorithms.scoring.BetweennessCentrality<MyN
ode, MyLink>(g, wtTransformer);

//Calculating Betweenness Centrality score of nodes

for(int i=0;i<Graph_Nodes_Only.size();i++)
{
    System.out.println("Graph Node
    "+Graph_Nodes_Only.get(i)+" Betweenness
```

```
            Centrality"
            +BC1.getVertexScore(Graph_Nodes_Only.get(i)));
        }

        //Calculating Betweenness centrality score of edges:
        Collection<MyLink> link1 = g.getEdges(); Iterator<MyLink>
        keys1 = link1.iterator(); while(keys1.hasNext())
        {
            MyLink link2 = keys1.next();
            System.out.println("Graph Edge "+ link2+" Betweenness
            centrality score "+BC1.
            getEdgeScore(link2));
        }

        Layout<MyNode, MyLink> layout = new ISOMLayout<MyNode,
        MyLink>(g); layout.setSize(new Dimension(700, 700));

        VisualizationViewer<MyNode, MyLink> vv = new
        VisualizationViewer<MyNode, MyLink>( layout);

        vv.getRenderContext().setVertexLabelTransformer(new
        Transformer<MyNode, String>() {
            @Override
            public String transform(MyNode arg0) { return
                arg0.toString();
            } });

        vv.getRenderer().setVertexRenderer(new MyRenderer());
         final DefaultModalGraphMouse<String, Number> graphMouse =
         new DefaultModalGraphMouse<String, Number>();
         graphMouse.setMode(ModalGraphMouse.Mode.PICKING);
         vv.setGraphMouse(graphMouse);


        JFrame frame = new JFrame(); frame.getContentPane().add(vv);
        frame.setDefaultCloseOperation(JFrame.EXIT_ON_CLOSE);
        frame.pack(); frame.setVisible(true);

    }

    static class MyRenderer implements Renderer.Vertex<MyNode,
    MyLink> {

        @Override
        public void paintVertex(RenderContext<MyNode, MyLink> rc,
        Layout<MyNode, MyLink> layout, MyNode vertex) {
```

```
        GraphicsDecorator graphicsContext =
        rc.getGraphicsContext(); Point2D center =
        layout.transform(vertex);

        char[] ss = vertex.toString().toCharArray();

        System.out.println("VMyrender: "+ss[0]+ss[1]);
        Shape shape = null; Color color = null;

        if(ss[1]=='p')
        { shape = new Ellipse2D.Double(center.getX()-7,
            center.getY()-7, 15, 15); color = new
            Color(235,28,34);
        } else
        { shape = new Rectangle2D.Double(center.getX()-7,
            center.getY()-7, 15, 15); color = new
            Color(0,128,255);
        }
        graphicsContext.setPaint(color);
        graphicsContext.fill(shape);
    }
}//close static class


public void ledados() throws Exception
{
    DataSource source = new DataSource(path); dados =
    source.getDataSet();
}

public void addVertex()
{
    int aux = 0;

    LinkedList<String> Distinct_Vertex = new
    LinkedList<String>();
    LinkedList<String> Source_Vertex = new LinkedList<String>();
    LinkedList<String> Target_Vertex = new LinkedList<String>();
    LinkedList<Double> Edge_Weight = new LinkedList<Double>();
    // add the distinct vertexes (actor_id) for(int i=0;
    i<dados.numInstances();i++)
    {
        Instance actual = dados.instance(i);

        Distinct_Vertex.add(actual.toString(0));
    }
```

```
// add the distinct vertexes (produtos) for(int a = 1; a<241;
a++)
{Distinct_Vertex.add(("p"+a).toString());}



System.out.println(Distinct_Vertex.size());

// add source, target and edge for(int i=0;
i<dados.numInstances();i++)
{
    Instance actual = dados.instance(i);

    aux =0;
    for(int z=1; z<actual.numAttributes(); z++)
    {

        if(Integer.parseInt(actual.toString(z))!=0)
        {
            Source_Vertex.add(actual.toString(0));
            Target_Vertex.add(("p"+z));
            Edge_Weight.add(Double.parseDouble(actual.toStrin
            g(z))); aux++;
        } }

    if(aux==0)
    { for(int yy=0; yy<Distinct_Vertex.size(); yy++)
        {
            if(Distinct_Vertex.get(yy).equals(actual.toString
            (0)))
            {
                System.out.println("Remove:
                "+Distinct_Vertex.get(yy));
                Distinct_Vertex.remove(yy);
            }

        }
    }
}

System.out.println(Distinct_Vertex);
System.out.println(Source_Vertex);
System.out.println(Target_Vertex);
System.out.println(Edge_Weight);

System.out.println("Betweenness calculation ");

this.BetweenNess_Centrality_Score(Distinct_Vertex,
Source_Vertex, Target_Vertex,
Edge_Weight);
```

```
        System.out.println("Graph \n " + this.g);

        System.out.println("---: "+Edge_Weight.size());

    }

    public static void main(String[] args) throws Exception {
        Graph_Algos GA1 = new
        Graph_Algos("C:\\cliente_produto1.csv");
        GA1.ledados();

        GA1.addVertex();
    }
}
```