Pós-Graduação em Ciência da Computação

João Emanoel Ambrósio Gomes

Caracterização de Grupos Baseada em Informações Relacionais

Universidade Federal de Pernambuco
posgraduacao@cin.ufpe.br
http://cin.ufpe.br/~posgraduacao

Recife
2018

João Emanoel Ambrósio Gomes

Caracterização de Grupos Baseada em Informações Relacionais

Tese apresentada ao Programa de Pós-graduação em Ciência da Computação do Centro de Informática da Universidade Federal de Pernambuco como requisito parcial para obtenção do grau de Doutor em Ciência da Computação.

Área de Concentração: Inteligência Artificial
Orientador: Prof. Dr. Ricardo Bastos Cavalcante Prudêncio

Recife
2018

Catalogação na fonte
Bibliotecária Monick Raquel Silvestre da S. Portes, CRB4-1217

G633c Gomes, João Emanoel Ambrósio
Caracterização de grupos baseada em informações relacionais / João Emanoel Ambrósio Gomes. – 2018.
124 f.: il., fig., tab.

Orientador: Ricardo Bastos Cavalcante Prudêncio.
Tese (Doutorado) – Universidade Federal de Pernambuco. CIn, Ciência da Computação, Recife, 2018.
Inclui referências.

1. Inteligência artificial. 2. Informações relacionais. I. Prudêncio, Ricardo Bastos Cavalcante (orientador). II. Título.

006.3 CDD (23. ed.) UFPE- MEI 2019-021

Tese de Doutorado apresentada por **João Emanoel Ambrósio Gomes** à Pós-Graduação em Ciência da Computação do Centro de Informática da Universidade Federal de Pernambuco, sob o título "**Caracterização de Grupos Baseada em Informações Relacionais**" Orientador: **Orientador: Prof. Dr. Ricardo Bastos Cavalcante Prudêncio** e aprovada pela Banca Examinadora formada pelos professores:

Aprovado em: 13/12/2018

Prof. Orientador: Dr. Ricardo Bastos Cavalcante Prudêncio

BANCA EXAMINADORA

Prof. Dr. Tsang Ing Ren
Centro de Informática/UFPE

Profa. Dr. Luciano de Andrade Barbosa
Centro de Informática/UFPE

Profa. Dr. Rafael Ferreira Leite de Mello
Departamento de Estatística e Informática/UFRPE

Prof. Dra. Solange Oliveira Rezende
Departamento de Ciências da Computação/USP

Prof. Dra. Tatiane Nogueira Rios
Departamento de Ciências da Computação/UFBA

Para minha mãe, Lúcia Ambrósio, pela plenitude do amor.

Para Meu pai, Francisco de Assis Gomes (Tuta), pelas lições de simplicidade e, por isso, de grandeza.

Aos meus irmãos Kahlil e Kallyne, por serem parte de mim e vivermos uma verdadeira fraternidade.

A minha esposa Regina, pois de todos os remédios caseiros, uma boa esposa é o melhor.

AGRADECIMENTOS

Antes de começar o ato de agradecimento, propriamente dito, devo assumir que esta é a última sessão que estou escrevendo na tese de doutoramento, e de fato, considero-a mais importante. A deixei por último com a certeza que seria inevitável vir à tona a emoção, como uma força gigante, e se transformando em felicidade, de auto reconhecimento, de sensação de missão cumprida e acima de tudo, e evidentemente, o sentimento de gratidão às pessoas, aos lugares por onde passei e às instituições que me acolheram.

Antes de tudo, gostaria de agradecer com toda minha energia, a DEUS e a Nossa Senhora Aparecida, os grandes proporcionadores de oportunidades, das possibilidades de se reinventar a cada dia, com novas perspectivas, novas metas, de conhecer novas pessoas e de tentar entender e seguir em frente neste mundo, que é um presente divino. Eu coloco toda minha confiança em ti, Senhor!

Agora, o nó na garganta vem com força, para agradecer a minha FAMÍLIA, com caixa alta e tudo. Meu pai, Seu Tuta, minha mãe, Dona Lúcia, obrigado por tudo, literalmente, todas as minhas conquistas nesses 31 anos, são na verdade, suas. Vocês me deram o dom da vida, me deram asas, me ensinaram a voar, e ainda quando caí, me ajudaram a levantar. Vocês são meus professores da vida, tentarei passar o máximo possível para as minhas gerações futuras (filhos, netos, bisnetos) tudo o que vocês me ensinaram.

Aos meus irmãos Getúlio Kahlil e Kallyne Maria, e as minhas sobrinhas Iara Maria, Ana Júlia, Sarah Maria, Lívia Maria e Alice por todo o apoio proporcionado e por sempre me fazerem sentir que onde quer que eu esteja sempre teria uma casa para voltar. Vocês são uma luz na minha vida, eu nada seria sem vocês.

Tenho uma gratidão imensa à minha esposa, Regina, que abraçou os meus sonhos como se fossem dela também, me deu forças, me acompanhou em todos os momentos, me fez crescer como ser humano. Te amo minha companheira, meu amor!

Quero agradecer de maneira honrosa, ao meu orientador Ricardo Prudêncio, um cara com uma visão ímpar de mundo, da área, e em tudo que se propõe a fazer e a investir, sempre se mostrou diferenciado. Ter tido a oportunidade de ser orientado por este ser, realmente é algo que vou levar para sempre. Ele tira "leite de pedra"! Ricardo, você também é responsável por esse caminho, que hoje eu sigo em paz.

Ao amigo e "coorientador", professor André Câmara, pelo suporte, incentivo e pelas discussões que ajudaram muito o desenvolvimento deste trabalho.

Aos meus amigos ("irmãos") Hilário Tomaz, Renê Gadelha e Jamilson Batista pela grande convivência ao longo desses sete anos (mestrado e doutorado), cujos momentos memoráveis lembrarei pelo resto da minha vida. Grandes aprendizados obtive com vocês, dentre eles, como se comportar em festas e como sair delas.

Aos amigos do Instituto Federal, Paulo, Ícaro, Elenilson, Oto, Jailton, Alexandre,

Vanessa, Gabi, Camila, Suzano. Em especial aos companheiros do IF pra frente, Thiago, Sarah e Andrezza, meu muito obrigado meus amigos.

A todos os amigos de Brejo Santo-CE em especial Pedro Ivo, Daniel, Davi, Armando, Antônio, Novinho, Moises, Joseph e Rodrigo que apesar da distância e do tempo nunca deixaram a amizade acabar.

Aos meus cunhados Ewerton, Luciana e Virgínia pela torcida e todo o apoio dado.

Ao meu sogro Chicão e minha sogra Corrinha (em memória), que também ajudaram de maneira significativa nessa caminhada da vida.

Aos amigos do laboratório da Associação de Pós-graduação (APG), em especial Alex Nery, Francisco Airton e Adriano Ferraz pela convivência durante os quatro anos de doutorado e pelas conversas, sempre interessantes, durante os intervalos do café.

Aos professores da graduação Ryan Azevedo e Rodrigo Rocha por terem sempre me incentivado e encorajado o meu interesse pela vida acadêmica.

Aos membros da banca examinadora pelas contribuições e direcionamentos no intuito de enriquecer este trabalho.

Ao Centro de Informática da Universidade Federal de Pernambuco - CIN-UFPE, pela excelente estrutura física e pessoal proporcionada a todos os seus alunos.

Por fim, mas não menos importante, ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) pelo auxílio financeiro.

RESUMO

Com o crescimento das redes sociais, diversas pesquisas vêm sendo realizadas para entendimento de suas estruturas. A análise e a extração de conhecimento das redes são largamente empregadas, dentre as investigações a compreensão do comportamento e das tendências das comunidades é uma atividade estratégica. Grande parte desse esforço está direcionado à detecção dos agrupamentos implícitos nas redes, detecção de comunidades; entretanto igualmente relevante é a atividade de rotulagem dos grupos, denominada caracterização de comunidades. Essa visa à descrição das comunidades a partir dos atributos individuais dos usuários. Entre as principais características dos métodos atuais de caracterização grupos, temos: (1) caracterização baseada apenas nos atributos dos usuários, (2) níveis de relevâncias equivalentes a todos os usuários e (3) consideração de todos os usuários da comunidade na caracterização. Todavia, em ambientes nos quais haja conexões entre seus usuários, como as redes sociais, uma nova dimensão de informação se apresenta, através da análise dos relacionamentos e afinidades entre os usuários (informação relacional). Presumivelmente, todas as comunidades têm os seus usuários influentes. Esses são os líderes de opinião, e podem desempenhar um papel mais importante para refletir as peculiaridades de uma comunidade. Tratar a escalabilidade das redes tende a ser um dos principais desafios das abordagens de caracterização de grupos, pois essa propriedade reflete diretamente na complexidade de descrição e robustez. Buscando o desenvolvimento de uma abordagem escalável e a incorporação dos benefícios supracitados com o uso das informações relacionais, propomos uma abordagem para caracterização de comunidades sociais baseada em informações relacionais. Assim, foi proposta a adição de uma nova etapa ao processo de caracterização de grupos, essa é responsável por filtrar os principais nós das comunidades a partir das informações relacionais (centralidade), ou seja, selecionar os nós que serão considerados no processo de caracterização dos grupos. O propósito é selecionar os nós, que representem/generalizem as comunidades, produzindo os melhores perfis possíveis, sem perdas de informações relevantes. Definiu-se como estudo de caso para esta tese as redes de coautoria, mais precisamente utilizou-se a biblioteca arXiv. Descrever comunidades acadêmicas é algo fundamental, proporcionando entendimento e acompanhamento das pesquisas, bem como a verificação das mudanças de temas nas comunidades. Os resultados, obtidos em três experimentos, demonstraram a capacidade da abordagem proposta na produção de perfis descritivos para os grupos observados, tanto fazendo uso de métodos de caracterização de grupos como de rotulagem de agrupamentos em documentos, com um custo computacional consideravelmente menor.

Palavras-chave: Inteligência Artificial. Análise de Redes Sociais. Descrição de Comunidades Sociais. Caracterização de Grupos. Informações Relacionais.

ABSTRACT

With the growth of social networks, several types of research have been carried out to understand their structures. Knowledge analysis and extraction of networks are widely used, among investigations understanding the behavior and trends of communities is a strategic activity. Much of this effort is directed to the detection of implicit groupings in networks, community detection; however, equally relevant is the communities labeling task, called group profiling. It aims at describing communities from the individual attributes of users. Among the main characteristics of the current group profiling methods, we have (1) characterization based only on the attributes of the users, (2) levels of relevancy equivalent to all users and (3) consideration of all users of the community in the characterization. However, in environments where there are connections between users, such as social networks, a new dimension of information is presented, through the analysis of relationships and affinities between users (relational information). Presumably, all communities have their influential users. These are opinion leaders, and they can play a more important role in reflecting the peculiarities of a community. Treating network scalability tends to be one of the main challenges of group profiling approaches, as this property directly reflects the complexity of description and robustness. Looking for the development of a scalable approach and incorporating the benefits mentioned above with the use of relational information, we propose an approach for group profiling based on the relational information. Thus, it was proposed to add a new stage to the group profiling process, which is responsible for filtering the main nodes of the communities from the relational information (centrality), that is, to select the nodes that will be considered in the group profiling process. The purpose is to select the nodes, which represent/generalize the communities, producing the best possible profiles, without loss of relevant information. The co-authoring networks were defined as a case study for this thesis, more precisely the arXiv library was used. Describing academic communities is fundamental, providing understanding and monitoring of research, as well as verifying the changes of themes in the communities. The results, obtained in three experiments, demonstrated the ability of the proposed approach to producing descriptive profiles for the observed groups, using group profiling methods and cluster labeling, with a considerably lower computational cost.

Keywords: Artificial intelligence. Social Networks Analysis. Description of Social Communities. Group Profiling. Relational Information.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE SIGLAS

ARS	*Analise de Redes Sociais*
BNS	*Bi-Normal Separation*
CGBA	*Caracterização de Grupos Baseado em Agregação*
CGBD	*Caracterização de Grupos Baseada em Diferenciação*
CGBDE	*Caracterização de Grupos Baseada em Diferenciação Egocêntrica*
CUR	Community-based Users' Representation
LSI	Latent Semantic Indexing
MAM	*Multi-Level Aggregation Method*
MI	*Mutual Information*
PLN	Processamento de Linguagem Natural
PMI	*Pointwise Mutual Information*
RSOs	*Redes Sociais Online*
TF-IDF	*Term Frequency-Inverse Document Frequency*
WRS	*Wilxocon Rank Sum*

SUMÁRIO

1 INTRODUÇÃO

Avanços recentes promovidos pelas tecnologias da informação e comunicação causaram mudanças significativas na sociedade, especialmente na forma como as pessoas interagem. Entre essas tecnologias, as redes sociais virtuais tornaram-se um fenômeno global com profundas influências no cotidiano social humano. As grandes mídias como jornais, revistas e portais, anteriormente os mais importantes provedores de informações para a população, vêm perdendo público. Ou seja, o modelo de disseminação "um para muitos" foi sendo substituído pelo modelo "muito para muitos" Stempel, Hargrove e Bernt (2000). As mídias sociais deram espaço para que usuários gerassem e compartilhassem conteúdo de forma expressiva, deixando de lado o comportamento passivo de absorção da informação e caracterizando uma forma de democratização na geração de conteúdo. Como consequência temos uma produção contínua de dados que podem ser analisados e explorados para diferentes propósitos Getoor e Diehl (2005).

Nesse contexto, a mineração de dados atualmente lida com o problema de investigar estruturas com muitas informações (propriedades), conjuntos de dados heterogêneos (como por exemplo, *Redes Sociais Online* (RSOs). Tais conjuntos de dados são melhores representados por redes ou grafos Getoor e Diehl (2005). Assim, estudos sobre mineração de links vêm sendo realizados para Análise de Redes Sociais (ARS), dentre esses se destacam os trabalhos voltados para análise das estruturas modulares das redes, denominadas comunidades/grupos Baumes et al. (2004), Sun et al. (2007), Tang e Liu (2010a). A homofilia foi uma das primeiras características estudas pelos pesquisadores de redes sociais Almack (1922), Bott (1928), Wellman (1926). Segundo o conceito de homofilia McPherson, Smith-Lovin e Cook (2001a), a probabilidade de pessoas semelhantes se relacionarem é superior a de pessoas com características distintas. A homofilia apresenta grande destaque nos estudos de *Analise de Redes Sociais* (ARS), sendo definida como o conceito base para o estudo das comunidades. Basicamente, uma comunidade é um conjunto de usuários que interagem uns com os outros com frequência Wasserman e Faust (1994). De acordo com (BACK, 1965), as interações sociais diádicas e os grupos formam a base para a instanciação social de qualquer indivíduo.

Uma das características mais tradicionais nas mídias sociais é a aglomeração de usuários em comunidades. Uma grande quantidade de agrupamentos vem surgindo nas RSOs, esses têm como peculiaridades, maiores interações entre os usuários aglomerados e elevada produção de conteúdos de interesses comuns. Esse fenômeno é conhecido, no contexto da ARS, como o processo de formação de comunidades Baumes et al. (2004), Sun et al. (2007), Tang e Liu (2010a). De modo geral as comunidades caracterizam-se pela cooperação de seus participantes, os quais compartilham valores, interesses, metas e posturas de apoio mútuo, através de interações no universo *online*. As comunidades podem ser identificadas

de forma explícita, por exemplo, quando são formadas por assinaturas de usuários; ou implicitamente, a partir da análise dos padrões de relacionamento dos usuários (topologia da rede). Quando implícitas, as comunidades podem ser automaticamente detectadas por técnicas adequadas de agrupamento gráfico Xie, Kelley e Szymanski (2011), algoritmos de detecção de comunidades. Os maiores esforços científicos voltados para o estudo das comunidades sociais online têm sido direcionados ao processo de detecção de comunidades, todavia igualmente relevante e complementar a atividade de detecção, é o processo de entendimento/descrição das comunidades.

1.1 CARACTERIZAÇÃO DE GRUPOS SOCIAIS

É amplamente observado que os indivíduos de uma comunidade compartilham interesses, atividades ou características semelhantes. A geração de conteúdos nas comunidades tem atraído a atenção de analistas e pesquisadores que desejam extrair ou inferir informações sobre os agrupamentos presentes nas redes. Diversas áreas, com diferentes propósitos, têm se desenvolvido na ARS direcionada ao estudo das comunidades, são exemplos: predição de comportamento, marketing, comércio eletrônico, detecção de comunidades, caracterização de grupos, entre outras Tan et al. (2013). O processo de extração de atributos descritivos de um grupo de pessoas é referido como caracterização de grupos (do inglês, *Group Profiling*) Tang et al. (2008). A descrição de comunidades pode ser de grande interesse em diferentes aplicações. No entanto, identificar um conjunto compacto de características relevantes para resumir uma comunidade não é trivial. As análises devem ser eficientes, capazes de lidar com grafos com milhões de nós e arestas que produzem enormes quantidades de informações.

Alguns trabalhos pioneiros tentaram compreender a formação de grupos com base na análise estatística da estrutura da rede. (BACKSTROM et al., 2006) estudaram algumas questões básicas sobre a evolução dos grupos, tais como: quais são as características estruturais que determinam qual grupo um indivíduo vai aderir? Os autores descobriram que a quantidade de amigos no grupo é um fator relevante para determinar que um novo usuário junte-se a ele. Isso proporciona um nível global de análise estrutural, facilitando o entendimento sobre como as comunidade atraem novos usuários. Já (LESKOVEC et al., 2008) observaram que o agrupamento espectral (um método popular usado para a detecção da comunidade) sempre encontra, mesmo que em pequena escala, comunidades triviais, ou seja, comunidades que estão conectadas ao restante da rede através de uma única aresta. Os dois trabalhos focam em uma imagem global (estatística) das comunidades. Dessa forma, mais pesquisas são necessárias para entender a formação de grupos específicos.

A caracterização de grupos sociais pode facilitar a compreensão das comunidades implícitas extraídas com base na estrutura da rede, bem como as comunidades explícitas formadas por assinaturas de usuários. Diversas são as aplicações para caracterização de grupos, dentre essas podem ser citadas: entendimento das estruturas sociais, visualização e

navegação da rede, acompanhamento a mudanças de temas de um grupo, casos alarmantes e *marketing* direto Tang, Wang e Liu (2011).

Analisando a aplicação em *marketing* direto, é possível que os consumidores (clientes) *online* formem diversos grupos, e que cada grupos possua um conjunto de diferentes mensagens, comentários e opiniões sobre os produtos. Uma vez que um perfil pode ser gerado para cada grupo, as empresas podem conceder novos produtos de acordo com as características dos grupos, sugerir produtos direcionados aos grupos, entre outras oportunidades Wang et al. (2010).

Dada uma rede particionada em comunidades, onde cada nó é associado a um conjunto de recursos (por exemplo, tópicos de interesse individual), a tarefa de caracterização de grupos é a seleção automática dos recursos mais descritivos para cada grupo, características que os generalizam/descrevam. De acordo com (TANG; WANG; LIU, 2011), a caracterização de grupos convencionalmente pode ser realizado usando duas abordagens distintas: *Caracterização de Grupos Baseado em Agregação* (CGBA) e *Caracterização de Grupos Baseada em Diferenciação* (CGBD). Na CGBA, o objetivo é encontrar características frequentes entre os indivíduos de um grupo, ignorando os demais usuários da rede. Essa tem por limitação a perda das informações globais, que refletem o contexto pelo qual a rede está inserida, uma vez que se limita a análise de cada comunidade isoladamente. Já na CGBD, o objetivo é identificar características que diferenciem um grupo dos demais indivíduos da rede, características que identifiquem o que torna a comunidade analisada divergente das outras. Essa trás consigo uma análise global das informações da rede no processo de caracterização das comunidades, todavia se faz necessária à consideração de todos os nós, o que pode ser impraticável em alguns casos (dependendo das dimensões da rede).

De acordo com (TANG; WANG; LIU, 2011) uma abordagem geral de caracterização de grupos pode ser definida em duas etapas básicas: inicialmente, é realizado o pré-processamento sobre a base de dados, identificando os atributos individuais dos usuários. Após o pré-processamento, os principais atributos dos usuários são selecionado; (1) fase de **Representação dos Usuários**, tendo como saída a matriz de dados pronta para execução de métodos descritores. Por fim, é iniciada a (2) fase de **Caracterização dos Grupos**, com a aplicação de métodos de caraterização sobre a matriz de dados pré-processada na fase anterior. Ao final dessa etapa, temos um perfil descritor para cada comunidade da rede.

1.2 DEFINIÇÃO DO PROBLEMA

Redes Sociais *Online* como Facebook, YouTube e Twitter conectam pessoas que estão produzindo grandes quantidades de dados em suas interações Tan et al. (2013). O volume, a velocidade de geração e processamento dos dados de diferentes fontes criam grandes desafios isolados ou combinados a serem superados, tais como: armazenamento, processamento, visualização e, principalmente análise dos dados. Quando se trata de gerência dos dados,

o volume varia de acordo com a capacidade das ferramentas utilizadas em cada área de aplicação. Porém percebe-se que, apesar do tamanha ser a parte mais evidente do problema, no contexto de ARS deve-se observar outras características, que requerem diferentes técnicas de processamentos e análises, as quais podem não estar diretamente associados ao tamanho absoluto dos dados Kurka, Godoy e Zuben (2015). Ou seja, tratar a escalabilidade das redes tende a ser um dos principais desafios da ARS.

Em um processo de caracterização de comunidades, foco desta tese, obviamente o ideal é que todas as informações e usuários sejam considerados, incluindo assim, mais conhecimento para ser propagado ao longo da rede. Certamente o desejo dos analistas é estudar as bases de dados por completo, não apenas uma amostra, ou ao menos aumentar as amostras o máximo possível. Fato esse, não trivial ou, ainda, inadequado para o contexto atual. Extrair conhecimento a partir de grandes massas de dados é de fato desafiador como discutido até aqui, pois além de serem heterogêneos em sua representação, os dados das RSOs são de conteúdo multidisciplinar Sorić et al. (2017).

Analisar grandes volumes de dados extraídos de comunidades sociais *online* permite que novas informações sejam obtidas, as quais não eram possíveis de serem verificadas devido às amostras desses tipos de dados serem menor. Porém, o aumento dos dados a serem analisados somam novos desafios aos já existentes na área de caracterização de grupos de redes sociais. O aumento das massas de dados das redes sociais está fazendo com que as técnicas, metodologias e ferramentas de mineração de dados e análise de grafos sejam adaptadas, melhoradas ou soluções novas sejam criadas Yaqoob et al. (2016), Ridgway (2016), Sorić et al. (2017), Olshannikova et al. (2017).

No contexto do estudo da caracterização de grupos sociais, todas as dificuldades de processamento das informações são absorvidas. Dessa forma, o processamento dos dados dos usuários analisados, suas representações, tende a ser uma tarefa muito onerosa no processo de geração dos perfis. O módulo de Representação de Usuários é essencial para a obtenção de um resultado efetivo, pois apresenta como saída os atributos considerados úteis para a caracterização dos grupos. Essencialmente são selecionados a partir do conjunto de dados, os atributos/características a serem analisados durante o processo de geração dos perfis. Basicamente é realizado um pré-processamento sobre os dados coletados dos usuários, selecionando apenas os atributos relevantes (com importância descritiva) para representação do domínio da rede em estudo. Uma representação incorreta dos usuários tende a resultar em uma caracterização também errónea, evidenciando a importância da realização de um pré-processamentos efetivo e robusto aos problemas supracitados.

Com a ampla adoção dessas redes, esses dados aumentaram em volume e precisam de processamento rápido, exigindo, por esse motivo, que abordagens no tratamento sejam empregadas. Em contextos já bastante explorados como sites e páginas da Web, técnicas de áreas como Recuperação de Informação, Mineração de Dados e Processamento de Linguagem Natural têm sido aplicadas com muito sucesso para extrair informação e

derivar conhecimento de corpus textuais Hanna (2004), Khribi, Jemni e Nasraoui (2008), Manning, Raghavan e Schütze (2008). Sabendo-se que a realização da filtragem dos atributos representativos da rede não é uma tarefa trivial, se faz necessário a consideração dessas técnicas de processamento de dados, bem como o acompanhamento e validação por especialistas do domínio da rede em estudo. Do contrário, poderá implicar em resultados incompletos ou até mesmo incorretos. Como por exemplo, a desconsideração de um atributo representativo.

Dada toda a problemática delineada, métodos de caracterização de grupos necessitam de um elevado tempo de processamento para recuperação e tratamento de todas as informações. Paralelamente a todas as dificuldades supracitadas, de acordo com (TANG; WANG; LIU, 2011), é desejável que todo método de caracterização de grupos satisfaça as seguintes propriedades:

- Descritivo: Os atributos selecionados para caracterizar um grupo devem refletir o perfil do grupo, ou seja, o interesse comum ou afiliação;
- Robusto: Grandes quantidades de dados são produzidas a cada dia em mídias sociais. Esses dados tendem a ser muito ruidosos. O método de caracterização de grupo deve ser robusto aos ruídos;
- Escalável: Em mídias sociais, uma rede com grandes dimensões é normal. A cada dia, novos usuários são inseridos na rede e novas conexões ocorrem entre os nós existentes. Dessa forma, os usuários envolvem-se em várias atividades e interações, produzindo uma grande quantidade de conteúdos ricos em informações relevantes sobre os usuários. Indiretamente isso também representa um grande desafio para os estudos de caracterização de grupos;

Os três critérios apresentados por (TANG; WANG; LIU, 2011), trazem consigo uma série de desafios e lacunas a serem superadas, todavia o tratamento da escalabilidade das redes tende a ser o maior gargalo das abordagens de caracterização de comunidades, pois essa propriedade reflete diretamente na complexidade das anteriores (descrição e robustez). Sabendo-se da necessidade de abordagens escaláveis, a caracterização de grupos em uma visão global, considerando-se todos os nós de uma rede, pode ser considerada inviável a depender da dimensão e evolução da rede analisada. Em tais casos seria necessário um elevado tempo de processamento para recuperação e processamento de todas as informações. Devido tais dificuldades, normalmente se tem a intenção de reduzir o tamanho do conjunto de dados utilizados; seja por dificuldades na coleta (dados ruidosos, duração, entre outros) ou pelos elevados custos computacionais na rotulagem das comunidades. Vislumbrando isso (TANG; WANG; LIU, 2011), buscaram a redução dos dados com base em uma visão egocêntrica das comunidades, *Caracterização de Grupos Baseada em Diferenciação Egocêntrica* (CGBDE). Essa, diferencia o grupo analisado apenas dos nós vizinhos

externos (margem), ou seja, considera como classe negativa apenas a visão egocêntrica da comunidade.

Embora pioneira no tratamento da escalabilidade para caracterização de grupos, a abordagem CGBDE apresenta algumas limitações, tais como: perca do contexto global no qual a rede está inserida, inaplicabilidade em casos de comunidades isoladas e em situações de "poucos" nós vizinhos. De fato algumas comunidades podem apresentar uma visão egocêntrica limitada, com poucos vizinhos ou com nós de baixa representatividade, irrelevantes à diferenciação da comunidade analisada. Outros aspectos das abordagens e métodos atuais que dificultam a escalabilidade na geração dos perfis de grupos, são: (1) caracterização baseada apenas nos atributos dos usuários, (2) níveis de relevâncias equivalentes a todos os usuários e (3) consideração de todos os usuários da comunidade analisada na caracterização.

Porém, as RSOs são ambientes ricos em interações entre os usuários, relacionamentos que representam/impulsionam grande parte dos conteúdos gerados. A análise desses dados se apresenta como uma nova dimensão de informação, as informações relacionais. Sabendo da disponibilidade das informações relacionais nas redes, considera-las na caracterização de grupos sociais pode enriquecer e melhorar todo o processo de descrição.

Sendo assim, podemos definir o problema abordado nesta tese da seguinte maneira: dada uma rede social particionada em comunidades, o objetivo é construir uma abordagem que considere as informações relacionais durante o processo de rotulagem e seja capaz de gerar perfis descritores para as comunidades reduzindo consideravelmente a complexidade computacional. Nesse sentido, podemos definir em termos objetivos a principal hipótese deste trabalho como:

Hipótese: A consideração das informações relacionais durante o processo de caracterização de grupos sociais possibilita melhorias na geração de perfis descritores reduzindo consideravelmente a complexidade computacional?

1.3 OBJETIVOS E ABORDAGEM PROPOSTA

Com base nos problemas e possibilidades apontadas, a presente tese de doutorado teve como principal objetivo desenvolver uma metodologia baseada em informações relacionais para a caracterização de grupos em redes sociais.

As informações relacionais podem promover grandes benefícios ao estudo e compreensão de comunidades em redes sociais, apresentando-se como o grande diferencial em relação às abordagens de rotulagem de agrupamentos em documentos. De modo geral, independente da abordagem de caracterização de comunidades adotadas, agregação, diferenciação global ou egocêntrica, os métodos atuais somente fazem uso de atributos dos usuários para a caracterização das comunidades. Sabendo-se que a rotulagem de comunidades sociais é realizada em um cenário que há informação relacional disponível, abordagens que considerem essa nova dimensão de informação são necessárias.

Ao se analisar uma comunidade, um aspecto importante é identificar quais são os nós ou os links mais importantes (ou centrais), pois esses podem revelar algumas peculiaridades relevantes sobre a comunidade. Normalmente utilizam-se as medidas de centralidade como forma de quantificar essa importância. A noção de centralidade, em várias aplicações, é associada à importância do elemento na estrutura, no problema apresentado, na comunidade. Espera-se, por exemplo, elevados[1] índices de centralidade de nó para uma pessoa influente em um grupo social. Essa informação tende a apresentar relevante atuação na análise dos dados, ou seja, nós centrais podem: generalizar conteúdos, apresentar maior influência na comunidade, ponderar a caracterização, entre outras possibilidades. Medidas de centralidade têm sido aplicadas com sucesso para rotulagem de agrupamentos em documentos[2] Role e Nadif (2014), Ienco e Meo (2008), Role e Nadif (2014) e obtenção dos autores mais proeminentes em redes de coautoria Liu et al. (2005).

Nesta tese, propomos uma abordagem que considere as informações relacionais durante o processo de caracterização dos grupos. Assim, é adicionado um novo módulo ao processo de caracterização de grupos, esse é responsável por filtrar os principais nós das comunidades a partir das informações relacionais, ou seja, selecionar os nós que serão considerados no processo de caracterização dos grupos. O propósito é selecionar os nós, que representem/generalizem as comunidades, produzindo os melhores perfis possíveis para as comunidades, sem perdas de informações relevantes. Vislumbramos com isso a redução dos custos computacionais, mantendo os perfis em bom nível de descrição.

Para alcançar o objetivo principal de maneira satisfatória, alguns objetivos específicos foram estabelecidos:

- Avaliar os impactos na adaptação de técnicas tradicionais de rotulagem de agrupamentos em documentos ao problema de comunidades, bem como adaptação do método *Wilxocon Rank Sum* (WRS) à caracterização em dados textuais;
- Implementar e avaliar a abordagem baseada em informações relacionais para caracterização de comunidades;
- Implementar e avaliar a representação de usuários baseada em comunidades.

Paralelamente, como solução adicional aos problemas de processamento dos dados dos usuários (representação), também propomos a Representação dos Usuários Baseada em Comunidades. Para tanto, aplicou-se as técnicas de seleção de atributos sobre os conjuntos de características de cada comunidade individualmente. Após termos a execução para todas as comunidades, realizamos a integração de todas as representações em um único subconjunto de características. Essa abordagem trata o desbalanceamento entre as

[1] Considerando, obviamente, que quanto maior o valor da centralidade, mais central o nó.

[2] A partir dos grafos de termos alguns trabalhos exploram o uso de métodos baseados em grafo, para quantificar a relevância dos tópicos/termos.

amostras, garantindo que todas as comunidades possuam um universo de características relevantes que às representem. Evitando, assim, que os usuários de uma comunidade sejam bem representados, enquanto os de outra comunidade sejam representados superficialmente.

Buscando afirmações sobre a adaptação de métodos e abordagens de rotulagem de agrupamentos em documentos ao problema de comunidades sociais. Como primeiro experimento, confrontamos os métodos CGBD já propostos (*Bi-Normal Separation* (BNS) e WRS), propondo uma variação do WRS para atributos textuais, bem como verificamos a eficácia da adaptação de métodos tradicionais de rotulagem de agrupamentos em documentos (*Term Frequency-Inverse Document Frequency* (TF-IDF) e Qui-quadrado). Os resultados desse experimento não apresentaram diferenças estatísticas entre as performances dos quatro métodos, o que consolida a expectativa inicial e levanta indícios positivos para a adaptação de abordagens de rotulagem em documento ao problema de caracterização de comunidades. No segundo experimento investigou-se o uso das informações relacionais na caracterização das comunidades. Para tanto, o filtro baseado em informações relacionais foi avaliado, esse seleciona com base nas medidas de centralidade os usuários a serem considerados no processo de caracterização; ou seja, nós que generalizam o conteúdo das comunidades. Verificou-se que de forma geral foi obtido uma considerável redução da complexidade temporal, obtendo resultados equivalentes (perfis bem similares ao global); bem como, em alguns casos, perfis melhores. Por fim, no Experimento III, verificou-se o tratamento inicial dos dados, buscando um balanceamento adequado entre os atributos selecionados como representativos para cada comunidade. Dessa forma, foi avaliada a representação de usuários baseada em comunidades. Pode-se notar que em todas as comunidades, independentemente da abordagem caracterização de grupos, global ou baseada em informações relacionais, os perfis obtidos após a aplicação da representação baseada em comunidades foram melhores, atingindo uma média global de 76,54% das avaliações. Ou seja, a desconsideração de atributos relevantes às comunidades foi evitada durante o processo de representação dos usuários.

1.4 ESTRUTURA DO DOCUMENTO

Este trabalho de doutorado está organizado em cinco capítulos. Uma breve descrição dos capítulos é apresentado a seguir:

- **Capítulo 2 - Caracterização de Grupos em Redes Sociais:** detalha-se os conceitos sobre as redes sociais relacionados ao estudo de caracterização de grupos, delineando também as abordagens e métodos para rotulagem de agrupamentos em dados textuais;
- **Capítulo 3 - Caracterização de Grupos Baseada em Informações Relacionais:** baseando-se no levantamento bibliográfico, a metodologia proposta é descrita;

- **Capítulo 4 - Estudo de Caso:** buscando evidências da viabilidade da metodologia proposta, realizamos experimentos comparativos entre métodos de caracterização de grupos e rotulagem de agrupamentos em documentos; bem como, verifica-se a aplicabilidade da representação de grupos baseada em comunidades com o uso das informações relacionais no processo de caracterização;
- **Capítulo 5 - Conclusão:** o capítulo final desta tese apresenta as conclusões a respeito do trabalho, discutindo os resultados, limitações e trabalhos futuros.

2 CARACTERIZAÇÃO DE GRUPOS EM REDES SOCIAIS

Este capítulo tem como objetivo explanar os principais conceitos sobre redes sociais relacionados à pesquisa. Será apresentada a evolução histórica, principais tipos, análise de redes complexas, detecção de comunidades, e de forma mais detalhada, as principais abordagens e métodos explorados na literatura da caracterização de comunidades sociais (*group profling*). A revisão de literatura foi essencial para a compreensão da área, bem como para a identificação das perguntas ainda em aberto a serem preenchidas. Sabendo-se que a caracterização de grupos sociais é uma das subáreas da Análise de Redes Sociais, inicia-se o levantamento bibliográfico tratando da teoria dos grafos, detalhado a seguir.

2.1 GRAFOS

A 'redescoberta' da importância da análise de redes sociais se deu pelo uso intensivo das mídias sociais Getoor e Diehl (2005). Nesta seção são descritos os conceitos básicos da teoria dos grafos e as aplicações através da análise de seus dados.

Em 1736, ao propor uma rigorosa demonstração matemática para o problema das pontes de Kögnisberg[1], o matemático suíço Leonhard Euler deu início ao ramo da matemática conhecido como Teoria dos Grafos, que constitui a base para estudos acerca das redes em geral Biggs (1976), Barabási (2009). Segundo (BOLLOBAS, 1998), Euler teria criado o primeiro grafo da história.

Em teoria dos grafos, um grafo é uma representação abstrata de um conjunto de objetos, no qual alguns deles estão conectados uns aos outros por meio de um conjunto finito de links (ver Figura 1). A representação de um grafo é dada por:

$$G = (V, A) \tag{2.1}$$

onde V é um conjunto de vértices ou nós não vazio, enquanto A é o conjunto de pares não ordenados de V, denominados arestas ou links Biggs (1976).

Redes geralmente são modeladas por meio de grafos, visto que esses possuem muitas características que podem ser associadas a diferentes propriedades estruturais de uma rede, além de possuírem um arcabouço matemático com o qual muitas dessas propriedades podem ser mensuradas e provadas Wasserman e Faust (1994).

Quando representadas por meio de grafos, os nós representam as entidades (membros ou atores) da rede e as arestas indicam como essas entidades se relacionam entre si Deo (1974). Por exemplo, em uma rede de co-autoria, os nós são os autores/pesquisadores que

[1] O problema é baseado na cidade de Kögnisberg (território da Prússia até 1945, atual Kaliningrado). No início do século XVII, o rio Pregel - que se estendia por toda a cidade - possuía sete pontes que o cruzavam. Essas serviram de base para a formulação do seguinte questionamento: seria possível atravessar todas as pontes sem passar mais de uma vez em cada ponte?

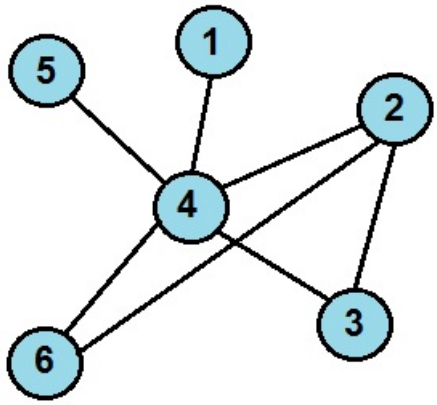

Figura 1 – Representação gráfica de um grafo. Os círculos representam os nós e as linhas que os conectam, as arestas.

fazem parte dela e as arestas são os laços gerados a partir da interações de co-autoria entre eles.

Nos grafos, diversas propriedades podem ser associadas aos vértices e as arestas, tais como postagens, publicações, peso, sentido das ligações e tipos de vértices. Dessa forma, as arestas podem ou não ser direcionadas, onde grafos com conexões direcionadas são chamadas de dígrafos (por exemplo, uma rede que representa citações de trabalhos científicos). Existem também grafos com mais um tipo de vértice, como exemplo nas Redes Sociais Educacionais, onde podemos ter alunos e professores. Por fim, as arestas podem ter valoração ou peso. Relações que consideram apenas a ausência ou presença de relacionamento são chamadas de dicotômicas, já quando se associa valores discretos ou contínuos as arestas, são denominadas valoradas Silva et al. (2007).

Outro ponto de estudo da teoria dos grafos é o particionamento de grafos. Seja $G = (V, A)$ um grafo, $P = (V', A')$ é dito um sub-grafo/partição de G se $V' \subset V$ e $A' \subset A$; ou seja, $P \subset G$. O particionamento de um grafo, é o conjunto de subgrafos distintos que unidos integram o grafo, ou seja, $G = \{P_1, P_2, ..., P_i, ..., P_n\}$. Os subgrafos que formam o particionamento são chamados de partições, e o valor P_i refere-se uma partição que contem um conjunto de vértices. São denominadas arestas de *corte*, as arestas que ligam vértices de diferentes partições. O corte de um partição, $EC(P_i)$, é igual a soma dos pesos das arestas de corte que ligam vértices de P_i a outras partições Almeida (2009).

A representação gráfica de um grafo é útil para entendimento e visualização de algumas de suas propriedades. Todavia, quando o grafo é muito grande e complexo, esse tipo de representação não é muito adequado. Salientando, ainda, a necessidade de outras representações para que o processo de análise de redes seja realizado automaticamente e corretamente por um computador.

Dentre as principais representações utilizadas para armazenar dados de um grafo, tem-se: matriz adjacência e lista de adjacências. Em uma matriz de adjacência, o valor de uma célula a_{ij} representa a ligação entre os vértices i e j. Para redes sem peso, o valor

de $a_{ij} = 1$, quando existe um relacionamento (aresta) entre os vértices i e j, ou $a_{ij} = 0$ caso contrário. Em caso de redes ponderadas os valores dos pesos são armazenados em uma matriz de pesos. Quando o grafo apresenta um número muito elevado de vértices em relação às arestas, a matriz adjacência tende a ser esparsa (muitos elementos a_{ij} têm valor 0), resultando em uma alocação de memória desnecessária. Como alternativa para essas situações utiliza-se listas de adjacências; essas permitem uma economia considerável de espaço, pois só são armazenadas as ligações existentes entre os pares de vértices. Todavia, o custo de acesso é maior devido a necessidade de busca-las em uma estrutura de lista.

Recentemente uma nova área de pesquisa surgiu para analisar propriedades estatísticas de "grandes" grafos (milhares e até milhões de vértices). Essa nova área é conhecida como análise de redes complexas, e é delineada a seguir.

2.2 ANÁLISE DE REDES COMPLEXAS

A teoria das redes complexas não só estende o formalismo da teoria dos grafos, mas principalmente, propõe medidas e métodos fundamentados em propriedades reais dos sistemas Costa et al. (2007). As redes complexas apresentam um comportamento multidisciplinar e estão presentes em aplicações de áreas distintas, utilizando a computação com ferramenta para modelagem, tratamento e análise dos dados.

A construção de redes complexas pode se basear em dois métodos Watts e Strogatz (1998): empírico e teórico. O método empírico associa cada nó da rede a uma determinada entidade, por exemplo, um computador, uma empresa ou um aeroporto, e as arestas conectam essa entidades conforme alguma relação que elas tenham. Esse método representa as redes baseadas em mundos reais (i.e., em sistemas físicos reais). O método teórico baseia-se em modelos de crescimento das redes para reproduzir as propriedades das redes reais, ou seja, modelos matemáticos. Essas são as redes construídas a partir de modelos matemáticos, que objetivam reproduzir aspectos quantitativos que emergem das redes complexas reais. Esses modelos evoluíram tanto ao longo do tempo, que se tornaram capazes de modelar o crescimento de muitos sistemas complexos, como a Internet, descrito por (COHEN et al., 2001). Essa seção descreve os principais tipos de redes complexas, conforme (ALBERT; BARABáSI, 2002), e apresenta com mais detalhes nos tópicos a seguir as redes de relacionamento, devido a sua extrema importância, e as redes de coautoria, por ser o estudo de caso desta tese. Os modelos de redes que serão apresentados a seguir têm como objetivo explicar como as redes descritas abaixo são geradas, suas propriedades, crescimento e evolução:

- **Sociais**: Redes que representam relações entre pessoas. Onde as pessoas são representadas por vértices e as relações entre elas por arestas. A Figura 2 apresenta um grafo referenciado as relações entre as pessoas nas diversas redes sociais exis-

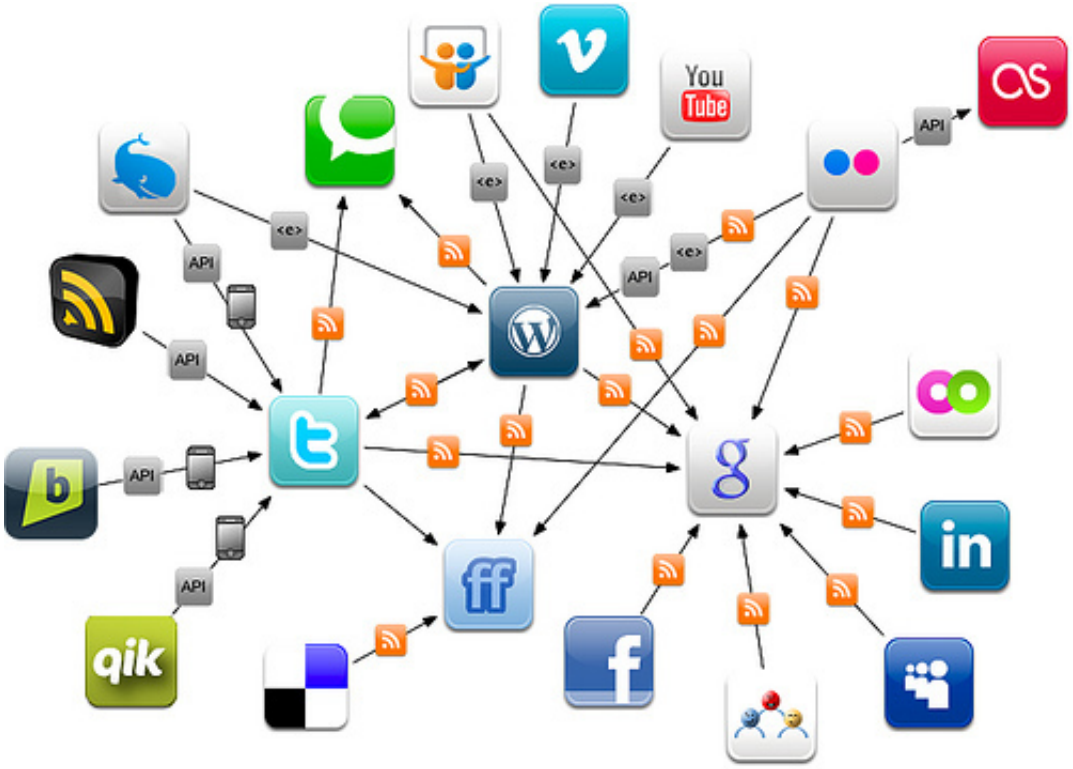

Figura 2 – Exemplos de redes sociais existentes e suas interações.

tes, caracterizando o aumento das interações e a elevação do fluxo de informações proporcionado pelas redes sociais;

- **Tecnológicas**: Redes desenvolvidas para distribuição de algum bem ou recurso, tais com redes de distribuição de água, eletricidade e telefonia;

- **Biológicas**: São redes construídas para modelar sistemas biológicos. Um exemplo é a rede metabólica, que modela o mecanismo de interação entre os substratos do metabolismo celular, tais como ATP, ADP e H_2O (nós) por meio de reações químicas das quais eles participam (arestas);

- **de Informação:** Redes obtidas a partir de bases de conhecimento formal, tais como as citações de artigos científicos e a *World Wide Web*. Como exemplo, tem-se a rede de citação, onde os nós representam trabalhos científicos publicados e as arestas evidenciam a referência de um artigo a outro previamente publicado.

- **de Coautoria:** Nas redes de coautoria, os autores de trabalhos publicados em veículos de difusão de conhecimento científico como revistas, jornais e eventos são considerados atores. Segundo (SILVA; PEREIRA, 2008), um ator para análise de redes sociais, é uma unidade discreta que pode ser de diferentes tipos: uma pessoa, ou conjunto discreto de pessoas agregados em uma unidade social coletiva, como subgrupo, organizações e outras coletividades. Esses atores são ligados por laço

relacional, conhecido também como ligação (linkage) ou simplesmente laço. Neste trabalho os atores são chamados de nós ou vértices e os laços de links ou arestas.

As relações entre pesquisadores podes ser representado com duas propriedades: (1) direção, podendo a relação ser direcional, caso tenha um ator como transmissor e outro como receptor (redes de citação), ou não-direcionais, em que a relação é recíproca (rede de coautoria); (2) valoração, podendo ser dicotômicas (quando se leva em consideração somente a presença ou ausência de uma relação, ou seja, um link de coautoria entre dois pesquisadores), ou valoradas quando atribui-se peso (de valores discretos) a relação (link) ou conforme o número de artigos publicados em coautoria pelos pesquisadores (SILVA et al., 2007). Nesta tese considerou-se uma rede de co-autoria (não-direcional) e dicotômicas, já que tem-se como interesse a caracterização dos grupos, ou seja, a reciprocidade entre seus membros.

Diversas redes de coautoria vêm sendo estudadas ao longo dos anos Newman (2001). No Brasil a RedeCI oferece um importante arcabouço bibliográfico para realização de estudos de coautoria dentro do campo de ciência da informação Silva et al. (2007), uma vez que mantém registros de diversos veículos de difusão, possibilitando a realização de estudos bibliométricos por parte de pesquisadores interessados Brandão, Parreiras e Silva (2007).

(SILVA et al., 2007) fizeram um estudo na RedeCI e apontaram a existência de uma baixa colaboração global entre os autores. Eles concluíram que em relação à colaboração global as sub-redes de autores colaboram pouco entre si, sendo que o maior grau de colaboração é local, dentro de uma sub-rede/comunidade. A presença de um alto número de comunidades e de autores individuais (que não colaboram com outros autores) faz com que a colaboração global, medida pela densidade da rede, seja baixa. Com isso conclui-se que estudos sobre as comunidades em redes de co-autoria são necessários e relevantes, principalmente com foco na justificativa e compreensão das suas formações. No entanto, não se pode afirmar que este comportamento esteja variando ao longo do tempo e que a tendência é o aumento ou a diminuição da colaboração global, apesar de haver indícios de que a topologia da rede se altera ao longo do tempo Brandão, Parreiras e Silva (2007).

A seguir descrevem-se brevemente as principais bases (mais acessadas e com maior número de publicações) para formação de redes de coautoria, atualmente. A base utilizada nesse trabalho é a arXiv. Ela será descrita brevemente, pois posteriormente será apresentada com maiores detalhes.

- ArXiv[2]: A arXiv pertence e é operada pela Cornell University e apresenta cerca de 1.447.659 documentos cadastrados, envolvendo as seguintes áreas:

[2] https://arxiv.org/

Física, Matemática, Ciência da Computação, Biologia Quantitativa, Finanças Quantitativas, Estatística, Engenharia Elétrica e Ciência de Sistemas, e Economia.

- DBLP[3]: A DBLP (*Digital Bibliography and Library Project*) é uma base de dados bibliográficos que provê uma vasta quantidade de informações sobre trabalhos científicos na área das Ciências da Computação. Atualmente a base é constituída de mais de 4.327.515 de publicações e cresce continuamente Sá e Prudencio (2011).
- ACM[4] A ACM (Association for Computing Machinery) Digital Library (DL) é a mais completa coleção de artigos e registros bibliográficos existentes hoje sobre os campos da tecnologia de computação e informação. O banco de dados inclui a coleção completa das publicações da ACM, incluindo revistas, anais de conferências, revistas, boletins informativos e títulos multimídia.
- AMiner[5]: A rede é uma das bases de dados disponíveis do ArnetMiner, um engenho de busca e mineração de redes sociais acadêmicas. As principais funcionalidades do engenho são: busca de perfis; encontrar especialistas; análise de conferência; busca de cursos; busca de subgrafos; navegação em tópicos; ranking de pesquisadores; e gerenciamento de usuário. A base de dados completa do sistema relaciona mais de 233.127.915 publicações.

Ao se analisar uma rede, um aspecto importante é decidir quais são os nós ou os links mais importantes (ou centrais). As medidas de centralidade são tradicionalmente as formas mais convencionais para quantificar essa importância, ou seja, identificar a relevância dos nós para a rede. A seguir damos seguimento a revisão bibliográfica desta tese, direcionando o foco para as tradicionais medidas de análise de redes sociais. O foco principal são as medidas que quantificam os nós das redes, essas têm aplicação direta com a abordagem proposta nesta tese.

2.2.1 Medidas de Centralidade

Alguns nós têm uma posição relativamente privilegiada devido à sua localização no grafo/rede. Um exemplo óbvio são os pontos de articulação. Modelando-se uma rede profissional como um grafo, pessoas em pontos de articulação atuam como *mediadores*, podendo obter vantagens devido a essa posição. A existência, em diversas redes, de pontos importantes motiva a busca por formas capazes de quantifica-los em relação as suas relevâncias. As redes do mundo real são geralmente estruturas complexas, isto é, não são governadas por fórmulas determinísticas (redes regulares). Essas podem apresentar

[3] https://dblp.uni-trier.de/
[4] https://dl.acm.org/
[5] https://aminer.org/

diversas topologias e características conforme o domínio estudado, aplicando-se algumas medidas, as redes podem ser analisadas, caracterizadas e modeladas Costa et al. (2007). Nesta seção, serão apresentadas medidas utilizadas para quantificar algumas das principais propriedades locais e globais de uma rede/grafo.

A análise de centralidade facilita o entendimento sobre o fluxo de informação dentro de um grafo, pois através dela é possível verificar quais os vértices têm maior importância dentro do grafo, ou seja, os nós mais importantes para o fluxo de informação. (FREEMAN, 1979), abordou o conceito de centralidade revisando um grande número de medidas até então publicadas e reduziu-as a três definições clássicas, a centralidade de grau (Degree Centrality), centralidade de proximidade (Closeness Centrality) e centralidade de intermediação (Betweenness Centrality). A seguir será apresentada cada medida básica individualmente, bem como outras medidas tradicionais derivadas delas.

- Grau: Também chamado de valência, o grau é uma medida da influência direta que um vértice tem em relação a seus contatos. Representando o número de arestas incidentes a ele; "Um nó importante está conectado com muitos nós". Na Figura 1, por exemplo, o vértice 4 apresenta o maior grau (igual a 5), pois ele está diretamente conectado a todos os outros vértices da rede, nós 1,2,3,5 e 6. Em outras palavras, o grau de um nó nada mais é do que o número de nós diretamente conectados a ele (nós vizinhos), conforme a equação:

$$grau(v_i) = |\Gamma(v_i)| \tag{2.2}$$

  Onde $\Gamma(v_i)$ representa o conjunto de nós vizinhos ao nó v_i e $|\Gamma(v_i)|$ mede a cardinalidade desse conjunto, ou seja, representa o número de vizinhos do nó v_i. Para análise de grau de toda a rede tem-se, *grau médio*, essa trata-se de uma medida simples para caracterização de redes, onde é determinado o número de médio de ligações entre o vértices de uma rede (ver Equação 2.3).

$$\langle deg \rangle = \frac{1}{n} \sum_{i=1}^{n} grau(v_i) \tag{2.3}$$

  A análise do grau e de sua distribuição permite determinar se a construção da rede obedece à alguma lei de formação ou tem caráter aleatório. Um exemplos de aplicação da centralidade de grau é mostrado em (JEONG et al., 2001): analisando as interações (arestas) entre as proteínas (nós) da bactéria *Helicobacter pylori*, é verificada uma correlação entre o grau e a letalidade da proteína se retirada do sistema.

- Proximidade: Em muitas aplicações, a centralidade de um nó depende principalmente de sua distância para outros nós. O grau de proximidade (do inglês, *closeness centrality*) de um vértice v é a média dos caminhos mínimos desse vértice para qualquer outro vértice da rede Clauset, Moore e Newman (2008); "Um nó importante

está próximos de outros nós". Assim, essa medida indica quão próximo um nó está do restante da rede, estando relacionada com o tempo que uma informação leva para ser compartilhada por todos os vértices na rede. As proximidades mais baixas são observadas nos vértices centrais, pois esses possuem menor distância em média para os demais. É calculado pela soma dos inversos das Distâncias Geodésicas (*dist*) de um nó em relação aos demais nós do grafo. A centralidade de proximidade, portanto pode ser descrita pela equação:

$$proximidade(v_i)^{-1} = \sum_{v_j \in V, v_j \neq v_i} dist(v_i, v_j) \tag{2.4}$$

Onde $dist(v_i, v_j)$ representa a distância geodésica entre os nós v_i e v_j. Ou seja, o cálculo da centralidade de proximidade envolve o cálculo das distâncias entre todos os pares de nós, que pode ser feita em tempo $O(n^3)$ usando o algoritmo de Floyd-Warshall; em redes esparsas, é possível efetuar o cálculo em $O(mn + n^2 \log n)$ usando o algoritmo de Johnson Borgatti e Everett (2006).

- Intermediação: O grau de intermediação (do inglês, *betweenness centrality*) de um vértice v é dado pela quantidade de caminhos mínimos, entre qualquer par de vértice da rede, que passa por ele; "Um nó importante faz parte de muitos caminhos". Portanto, essa medida analisa a frequência com que um nó aparece no caminho geodésico entre dois outros nós quaisquer, ou seja, procura identificar nós com grande potencial de controle do fluxo de informação na rede. Pode ser descrito pela equação:

$$intermediacao(v_i) = \sum_{j<k} \frac{g_{jk}(v_i)}{g_{jk}} \tag{2.5}$$

Onde g_{jk} é o número de caminhos geodésicos que ligam os nós v_j e v_k e $g_{jk}(v_i)$ é o número de caminhos geodésicos do total g_{jk} que unem os nós v_j e v_k passando pelo nó v_i em algum ponto.

(BRANDES, 2001) descreve um algoritmo para calcular a centralidade de intermediação de todos os nós de um grafo, com uma complexidade de tempo de $O(mn + n^2 \log n)$, para grafos com pesos, e $O(mn)$, para grafos sem pesos.

Das três medidas básicas, algumas outras podem ser derivadas. Apresentamos aqui 3 medidas de destaque na literatura e aplicadas nessa tese, seguindo a classificação de centralidade sugerida por (BORGATTI; EVERETT, 2006).

- Autovetor: Ao contrário da Centralidade de Grau (Equação 2.2.1), outra maneira de interpretar a centralidade de um nó é considera-la em função da centralidade dos seus vizinhos, ou seja: "um nó importante tem vizinhos importantes". Baseada em autovalores e autovetores de matrizes simétricas, cujo objetivo é medir a importância de um vértice em função da importância de seus vizinhos. Isto quer dizer que,

mesmo se um vértice está conectado somente a alguns outros vértices da rede (tendo assim uma baixa centralidade de grau), estes vizinhos podem ser importantes e, conseqüentemente, o vértice será também importante, obtendo uma centralidade de autovetor elevada. Seja w_{ji} o peso da aresta entre nós v_j e v_i e λ uma constante, a centralidade de autovetor de um nó v_i é dada por:

$$C_E(v_i) = \frac{1}{\lambda} \sum_{v_j \in N(v_i)} w_{ji} \times C_E(v_i) \tag{2.6}$$

(BONACICH, 1972) foi quem primeiro recomendou que o autovetor principal v_i associado a λ fosse usado como uma medida de centralidade. Uma vantagem dessa medida é a possibilidade de uso de métodos numéricos (como o método das potências) para a determinação do autovetor.

Um problema da medida de autovetor está no reflexo global de nós importantes geram sobre os seus vizinhos, ou seja, se um nó importante é vizinho de um grande número de nós, todos eles ganham importância. Seria razoável sugerir que a importância do nó fosse "diluída" entre seus links. Essa é a principal ideia do algoritmo *PageRank* descrito a seguir.

- *PageRank*:

  O *PageRank* Baseia-se na medida de centralidade do autovetor ($C_E(v_i)$) e implementa o conceito de 'votação', diferenciando-se da centralidade de autovetor pela introdução de um fator de amortecimento e de escala. Seja d um fator de amortecimento (definido nessa tese como 0,85, idem Mihalcea e Tarau (2004)), a pontuação de PageRank $S(v_i)$ de um nó v_i é inicializada para um valor padrão e computada iterativamente até a convergência usando a seguinte equação:

$$C_{PR}(v_i) = (1-d) + \left(d \times \sum_{v_j \in N(v_i)} \frac{w_{ji} \times C_{PR}(v_j)}{\sum_{v_k \in N(v_j)} w_{jk}} \right) \tag{2.7}$$

  Conforme visto na centralidade de proximidade, outra possibilidade de se atribuir a centralidade é primeiro designar o nó (ou grupo de nós) mais central, e a partir desse fazer uso da distância dos outros nós em relação ao escolhido (ou grupo de nós). Essa outra variação é conhecida com centralidade de excentricidade, descrita a seguir.

- Excentricidade: Em Pesquisa Operacional, alguns problemas de localização consistem em se determinar um local de modo que minimize o tempo máximo de viagem entre o mesmo e todas as demais localizações. Estes problemas possuem diversas aplicações práticas, como por exemplo, a instalação de um hospital, cujo objetivo é minimizar o tempo máximo de atendimento de uma ambulância a uma possível emergência.

É neste sentido que (HAGE; HARARY, 1995) propuseram uma medida chamada centralidade de eficiência (comumente chamada nas aplicações de centralidade de excentricidade), baseada no conceito de excentricidade[6] de um vértice. Essa tem o objetivo de determinar uma localização que minimize as distâncias a todas as demais localizações da rede. Seja G um grafo conexo com n vértices e seja v_i um vértice de G. Então o menor caminho percorrido para a maior distância (seja (v_i, v_j) onde v_j é o nó mais distante do v_i). De fato, se a excentricidade do nó v_i é alta, isso significa que todos os outros nós estão próximos. Em contraste, se a excentricidade é baixa, isso significa que há pelo menos um nó (e todos os seus vizinhos) que está longe do nó v_i. A excentricidade de um nó v_i é dada por:

$$C_{ECC}(v_i) = \frac{1}{dist(v_i, v_j)} \tag{2.8}$$

onde $dist(v_i, v_j) = max\{dist(v_i, v_j) : v_i \in V\}$.

Ou seja, a centralidade de excentricidade/eficiência de v_i é dada pelo inverso da excentricidade de v_i; assim, esta medida indica que um vértice é mais eficiente quanto menor for sua excentricidade. Para o grafo G dado na Figura 1, temos que as distâncias entre v_1 e todos os demais vértices são dadas por:

$dist(v_1, v_2) = 2, dist(v_1, v_3) = 2, dist(v_1, v_4) = 1, dist(v_1, v_5) = 2, dist(v_1, v_6) = 2.$

Logo, a excentricidade é dada pela distância de v_1 a v_2, ou seja, $dist(v_1, v_2) = 2$. Ou seja, a centralidade de eficiência de v_1 é:

$C_{ECC}(v_i) = \frac{1}{dist(v_i, v_j)} = \frac{1}{2} = 0,5$

Além das medidas locais, baseadas na centralidade, temos medidas globais, tais como: Diâmetro, Coeficiente de Agrupamento e Densidade. Todas essas medidas estão consolidadas na literatura podem ser encontradas em detalhes nos trabalhos de (NEWMAN, 2003) e (BARABáSI, 2009).

2.2.2 Detecção de Comunidades

Uma propriedade bastante comum em redes complexas é a presença de estruturas modulares chamadas de grupos/comunidades. As comunidades são aglomerados formados por grupos de nós altamente relacionados com outros nós pertencentes ao mesmo aglomerado, mas pouco densos com relação aos relacionamentos que possuem com nós de aglomerados

[6] Dado um vértice $v_i \in V$, afastamento, ou excentricidade $e(v_i)$ é a maior distância entre v_i e algum outro vértice em G

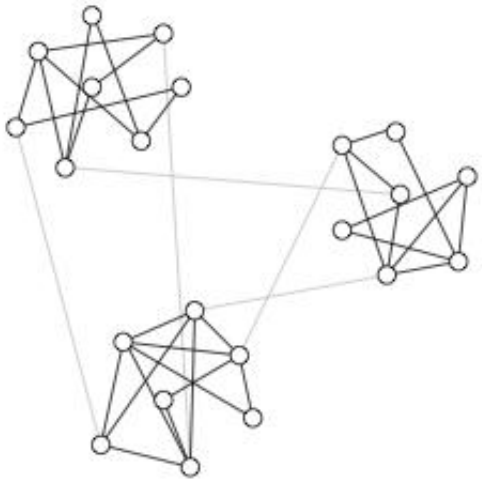

Figura 3 – Formação de comunidades. Os traços mais escuros indicam as ligações entre membros da comunidade (conexões entre nós de um mesmo aglomerado). Os traços claros indicam como as comunidades se relacionam entre si. Fonte: Girvan e Newman (2002)

vizinhos Girvan e Newman (2002) (ver Figura 3). Hoje, sabe-se que a clusterização/aglomeração está presente na Web, nas redes físicas de computadores na Internet, nas cadeias alimentares, nas redes moleculares, nas redes sociais, entre outras; ou seja, trata-se de uma propriedade genérica das redes complexas Barabási (2009).

Embora as comunidades possam estar presentes em distintas redes, a sua conotação varia com o tipo da rede em questão. Em uma rede social, podem conotar subgrupos sociais com interesses em comum, por exemplo, Girvan e Newman (2002); enquanto que em uma rede metabólica, está mais relacionada a formação de grupos funcionais Yuruk, Xu e Schweiger (2007).

De acordo com (WASSERMAN; FAUST, 1994), um grupo (comunidade) é um conjunto de usuários que interagem uns com os outros com frequência. Diversos estudos e aplicações já foram desenvolvidas para a identificação de grupos, incluindo: visualização de rede, análise de inteligência Baumes et al. (2004), Compreensão de rede Sun et al. (2007),estudo comportamental Tang e Liu (2010a), segmentação e recomendação Wang et al. (2010), e filtragem colaborativa Chen et al. (2009).

Paralelamente também vem sendo desenvolvidos diversos métodos para detecção de comunidades em redes complexas Wasserman e Faust (1994), dentre eles podemos citar: Algoritmo de (GIRVAN; NEWMAN, 2002), *Fast greedy modularity optimization* de (CLAUSET; MOORE; NEWMAN, 2008), *Fast modularity optimization* de (BLONDEL et al., 2008), entre outros. A seguir o algoritmo *Fast modularity optimization* de (BLONDEL et al., 2008), também conhecido como algoritmo *Multi-Level Aggregation Method*, é detalhado. Esse foi método utilizado no estudo para identificação das comunidades.

Nesta tese considerou-se como algoritmo para detecção das comunidades, o *Multi-*

Level Aggregation Method (MAM). Esse é um método de passos múltiplos, baseado na modularidade local otimizada Girvan e Newman (2002). De acordo com (Newman, 2006), a modularidade é uma função benefício utilizada na análise de redes ou grafos, tais como as redes de computadores ou as redes sociais. Ela quantifica a qualidade das partições de uma rede em módulos ou comunidades. Boas partições, com valores elevados de modularidade (a modularidade de uma partição é um valor escalar entre -1 e 1), são aqueles que apresentam conexões internas densas entre os nós dentro dos módulos, mas apenas ligações esparsas entre diferentes módulos. Segundo (NEWMAN, 2003), a modularidade também pode ser usada como uma função de otimização.

O algoritmo de detecção de comunidade MAM é dividido em duas fases, que são repetidas iterativamente. Na primeira fase a partir de uma rede de N nós, como primeiro passo, cada nó da rede recebe uma comunidade diferente. Logo, na partição inicial a quantidade de comunidades é igual ao número de nós existentes. Então, para cada nó i são considerados os seus j nós vizinhos, para avaliação do ganho de modularidade que ocorreria na remoção de i da sua comunidade, e colocando-o na comunidade de j. Dessa forma o nó i é colocado na comunidade que obteve o maior ganho positivo. Todavia, se nenhum ganho positivo é possível, i fica em sua comunidade de origem. Esse processo é aplicado repetidamente e sequencialmente para todos os nós até que nenhuma melhoria possa ser alcançada e a primeira fase é, então, concluída.

A segunda fase do método MAM consiste na construção de uma nova rede, cujos nós agora são as comunidades encontradas durante a primeira fase. Para geração dessa rede, novos pesos são definidos para as ligações entre os nós, dado pela soma dos pesos das ligações entre os nós correspondentes de ambas as comunidades Blondel et al. (2008). Após a conclusão da segunda fase, é possível reaplicar a primeira fase do algoritmo para a rede ponderada resultante, e iterar novamente. A combinação dessas duas fases será denominada “por passagem”. Por construção, a quantidade de comunidades diminui em cada passagem, e como consequência a maior parte do tempo de processamento é usada na primeira fase. As passagens são iteradas (ver Figura 4) até que não haja mais mudanças e a máxima modularidade é atingida. O algoritmo de MAM, ainda, resulta em uma hierarquia de comunidades. A altura da hierarquia gerada é determinada pelo número de passagens (geralmente é um número pequeno).

Quando o valor máximo de modularidade local é alcançado durante a optimização, isso corresponde ao número de agrupamentos a partir das comunidades formadas. Uma vez que geralmente existem vários valores de máxima modularidade local disponíveis durante a fase de optimização, esses números de fragmentação sob diferentes valores de modularidade. Dessa forma, podem ser considerados níveis de agrupamento (resoluções) diferentes. Portanto, o número aproximado de comunidade pode ser encontrado a partir dos diferentes níveis de agrupamento, pois o valor máximo de modularidade global pode ser encontrado entre os valores máximos de modularidades locais. Dessa forma, o MAM

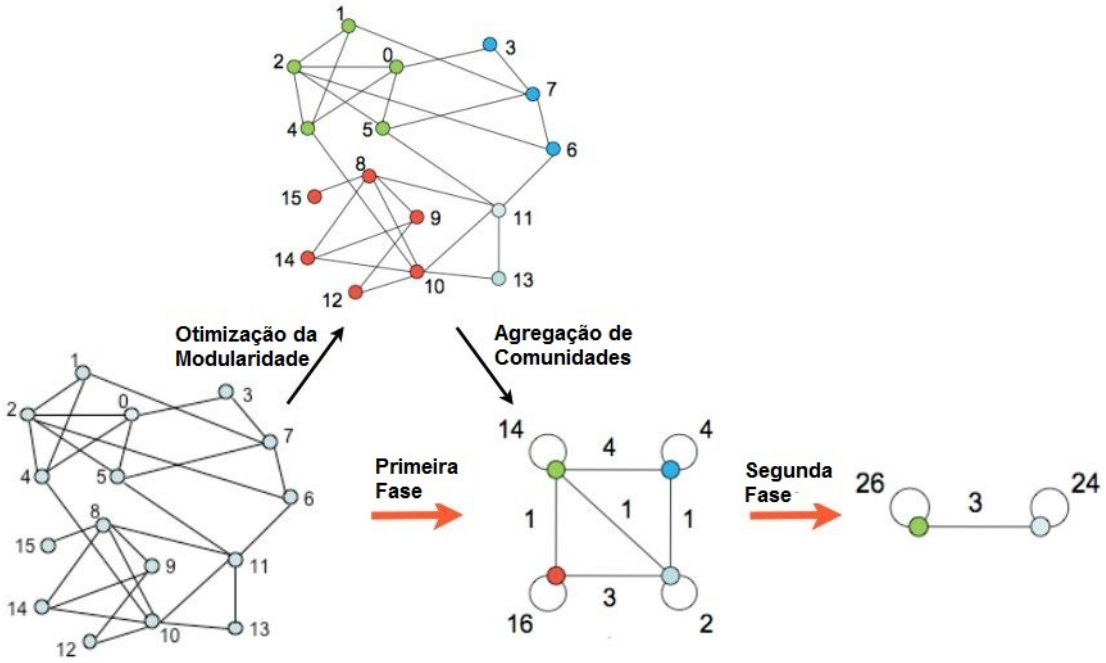

Figura 4 – Visualização dos passos do Método MAM. Cada passo é realizado em duas fases: na primeira a modularidade é otimizada, permitindo somente alterações locais das comunidades; já na segunda as comunidades encontradas são agregadas para a construção de uma nova rede de comunidades. Esses passos são repetidos iterativamente até que nenhuma aumento de modularidade seja possível e uma hierarquia de comunidades seja produzida. Fonte: Blondel et al. (2008)

fornece um esquema de heurística que possibilita a identificação do número de comunidades automaticamente.

Em um estudo comparativo realizado por (FORTUNATO; LANCICHINETTI, 2009), foram analisados diversos métodos de detecção de comunidades. De acordo com autor, o método de MAM oferece um compromisso equilibrado entre a precisão da estimativa da máxima modularidade, que é melhor que técnicas gananciosos como *Fast greedy modularity optimization* Clauset, Moore e Newman (2008), e uma complexidade computacional, que é basicamente linear ao número de Links do grafo. O autor ainda conclui que o MAM está entre os métodos que apresentaram melhor desempenho sobre redes não direcionadas e não ponderadas (como por exemplo, as redes de coautoria, estudo de caso do presente trabalho). Além de ser muito rápido, essencialmente em redes de tamanho linear. De fato, apesar de simples esse algoritmo tem várias vantagens, dentre elas podem ser referenciadas, os passos intuitivos e de fácil implementação, e o resultado ser não supervisionado Blondel et al. (2008).

2.3 CARACTERIZAÇÃO DE GRUPOS SOCIAIS

Nas redes sociais uma comunidade (grupo) é um conjunto de usuários que interagem uns com os outros com frequência Wasserman e Faust (1994), sendo umas das características

mais relevantes no estudo das redes. Um número considerável de propostas têm explorado a identificação de comunidades nas redes Chintalapudi e Prasad (2015), Malliaros e Vazirgiannis (2013), mas um aspecto igualmente importante é o processo de rotular (gerar perfis) os grupos obtidos, caracterização de comunidades Tang, Wang e Liu (2011). O usuário final pode decidir explorar uma comunidade geralmente guiado pelo seu perfil; porém, resumir uma comunidade com base em um pequeno conjunto de características não é trivial.

De acordo com (TANG; WANG; LIU, 2011), o estudo de caracterização de grupos sociais (do inglês *Group Profiling*) visa à construção de um perfil descritivo para um dado grupo fornecido. Em relação a objetivos, a caracterização de comunidades é bastante semelhante a uma classe mais tradicional de algoritmos, conhecida como rotulagem em agrupamentos de documentos textuais (do inglês, *cluster labeling*).

As principais diferenças entre elas são a fonte de informação (instâncias em rotulagem de agrupamentos e propriedades de rede, nós e links, para caracterização de comunidades) e o processo de identificação dos grupos (aprendizagem não supervisionada tradicional (por exemplo, *clustering*) e detecção de comunidades em redes, respectivamente). Todavia, o principal propósito é o mesmo, obter descrições concisas e significativas para os agrupamento/comunidades.

Alguns trabalhos exploram especialmente a geração de rótulos para a comunidade considerando as informações textuais produzidas pelos usuários, que visam à identificação de palavras-chaves para indicar os possíveis tópicos aos quais os usuários agrupados se refiram Tang, Wang e Liu (2011). Considerando a semelhança entre rotulagem em documento e caracterização de comunidades, técnicas convencionais de rotulagem em agrupamentos de documentos podem ser usadas para caracterizar comunidades, em casos que os indivíduos produzam informação textual. Diversas abordagens vêm sendo desenvolvidas para realização da caracterização de comunidades em redes socais, e essas caracterizações chegam a ser tão relevantes quanto à própria comunidade.

Alguns trabalhos pioneiros tentam compreender a formação de grupos com base na análise estatística da estrutura da rede. (BACKSTROM et al., 2006) estudaram algumas questões básicas sobre a evolução dos grupos, tais como: quais são as características estruturais que determinam qual grupo um indivíduo vai aderir? Os autores descobriram que a quantidade de amigos no grupo é um fator relevante para determinar que um novo usuário junte-se ao grupo. Isso proporciona um nível global de análise estrutural, facilitando o entendimento sobre como as comunidades atraem novos usuários. Já (LESKOVEC et al., 2008) observaram que o agrupamento espectral (um método popular usado para a detecção da comunidade) sempre encontra, mesmo que em pequena escala, comunidades triviais, ou seja, comunidades que estão conectadas ao restante da rede através de uma única aresta. Os dois trabalhos focam em uma imagem global (estatística) das comunidades. Dessa forma, mais pesquisas são necessárias para entender a formação de grupos específicos.

A caracterização de grupos sociais pode facilitar a compreensão das comunidades implícitas extraídas com base na estrutura da rede, bem como as comunidades explícitas formadas por assinaturas de usuários. Diversas são as aplicações para caracterização de grupos, dentre essas podem ser citadas: entendimento das estruturas sociais, visualização e navegação da rede, acompanhamento a mudanças de temas de um grupo, casos alarmantes e marketing direto Tang, Wang e Liu (2011).

Analisando a aplicação em *marketing* direto, é possível que os consumidores (clientes) *online* formem diversos grupos, e que cada grupo possua um conjunto de diferentes mensagens, comentários e opiniões sobre os produtos. Uma vez que um perfil pode ser gerado para cada grupo, as empresas podem conceder novos produtos de acordo com as características dos grupos, sugerir produtos direcionados aos grupos, entre outras oportunidades Wang et al. (2010).

De acordo com (KUMAR; NOVAK; TOMKINS, 2006) uma rede pode ser dividida em três regiões: *singletons* (nós que não interagem com outras pessoas), comunidades isoladas, e um componente gigante conectado. Comunidades isoladas realmente ocupam uma porção muito estável em toda a rede e, consequentemente, a probabilidade de duas comunidades isoladas se unirem com a evolução da rede é baixíssima. Todavia, com a caracterização dos grupos disponíveis, é possível que uma comunidade isolada ou um *singleton* identifique outros grupos semelhantes, e realizem interação (conexões) de comunidades isoladas de interesses semelhantes. Baseando-se nesses possibilidades algumas abordagens vêm sendo desenvolvidas para a realização da caracterização de grupos socais. De acordo com (TANG; WANG; LIU, 2011), é desejável que todo método de caracterização de grupos satisfaça as seguintes propriedades:

- Descritivo: Os atributos selecionados para caracterizar um grupo devem refletir o perfil do grupo, ou seja, o interesse comum ou afiliação.
- Robusto: Grandes quantidades de dados são produzidas a cada dia em mídias sociais. Estes dados tendem a ser muito ruidosos. O método de caracterização de grupo deve ser robusto aos ruídos.
- Escalável: Em mídias sociais, uma rede com grandes dimensões é normal. A cada dia, novos usuários são inseridos na rede, e novas conexões ocorrem entre os nós existentes. Dessa forma, os usuários envolvem-se em várias atividades e interações, produzindo uma grande quantidade de conteúdos ricos em informações relevantes sobre os usuários. Indiretamente, isso também representa um grande desafio para os estudos de caracterização de grupos.

(BARBIER; TANG; LIU, 2011) destacam alguns esforços de investigação em curso para compreensão de grupos através das mídias sociais, apresentando: as motivações, desafios e

tarefas. É um trabalho de cunho conceitual, apresentando a base para uma pesquisa de caracterização de grupos em redes sociais.

- Motivações

  Com a caracterização de grupos sociais, temos grandes oportunidades comerciais, tais como: marketing direcionado; oportunidades para entidades comerciais entenderem melhor os clientes; e através da mídia social, as empresas ganham acesso a amostras maiores ou até mesmo toda a população de consumidores ou potenciais consumidores, bem como: a possibilidade de melhorar a comunicação com os grupos; muitos sites e aplicações de mídia social que podem beneficiar os próprios grupos, prestando assistência na predição de grupos semelhantes, e, assim, sugerir/identificar potenciais ligações.

- Desafios

  - Compreender e recuperar grupos de tamanhos variados de forma eficaz;

  - Dados de mídias sociais online apresentam ruídos (grande quantidade de spam e textos despadronizados);

  - Os grupos evoluem em mídias sociais. A cada dia novos usuários se juntam a novos grupos, formando novas conexões e, consequentemente, interações adicionais ocorrem entre os grupos existentes.

- Tarefas

  São apontadas quatro tarefas principais para a compreensão de um grupo através de mídias sociais online.

  - Extração do Grupo: o próprio grupo deve ser detectado, ou extraído, entre os dados da rede social, a fim de ser estudado. Os grupos podem está implícitos ou explícitos (assinaturas de usuários). Os grupos implícitos na rede, precisam ser detectados via algoritmos. As duas principais abordagens são: a estática e a dinâmica (considerando as informações temporais).

  - Caracterização do Grupo: definir os descritores e características dos grupos, essas serão utilizadas para geração do seu perfil.

  Uma vez que um grupo é extraído, uma questão fundamental é: "dado um grupo de usuários, podemos descobrir por que eles se conectaram?". Isto resume a tarefa de caracterização de grupos, que visa obter características que descrevam o grupo. Aqui os autores ainda apontam as principais abordagens de caracterização de grupos sociais, citadas nas próximas subseções desta proposta.

  - Sentimento e Influência do grupo: O Sentimento de um grupo e sua capacidade de influenciar outros indivíduos ou grupos, precisam ser entendidos.

Até o momento, vimos como identificar o grupo e descobrir suas principais características representativas, mas o que sabemos sobre essas características? Por exemplo, em uma rede social de jogos foi selecionado como atributo caracterizador o termo "PES"; o grupo tem afinidade para o jogo "PES"? O grupo prefere outro estilo de jogo de Futebol? Para ganhar outros elementos que retratem a forma como o grupo se sente sobre jogos de futebol, ou como ele pode influenciar os outros sobre o tema, precisamos obter informações adicionais sobre o sentimento do grupo.

Analisando o sentimento do grupo temos *insights* das opiniões no grupo sobre um tema ou assunto particular (atributo selecionado como caracterizador). As opiniões de um grupo também podem influenciar outros grupos ou indivíduos. Dessa forma, a compreensão do sentimento do grupo pode revelar: atitudes de um grupo, emoções, opiniões e crenças; aspectos importantes da catalogação de grupos de mídia social. Como por exemplo: Os *insights* sobre as atitudes e opiniões de um grupo podem ajudar as empresas a entender como os clientes vêm produtos e ofertas, bem como, compreender o efeito de campanhas e publicidades baseadas em mídia sociais.

Dessa forma, a análise de sentimento pode ser vista como uma atividade sequente à caracterização de grupo, buscando o entendimento das opiniões sobre cada característica selecionada para o perfil do grupo. (BARBIER; TANG; LIU, 2011) estão particularmente interessados em saber se um determinado sentimento irá ou não se propagar dentro de um grupo/rede. As primeiras pesquisas revelam que os níveis de participação entre usuários podem ser inversamente correlacionados com a propensão para propagação de emoções entre os usuários em uma rede Zafarani, Cole e Liu (2010). Em outras palavras, os usuários gregários são menos ativos online, todavia são mais propensos a ajudar a replicar o sentimento de outros usuários.

Outra questão relevante é a identificação de membros influentes no grupo, esses podem mudar as atitudes do grupo, motivações e opiniões. Identificar os usuários influentes em um grupo/comunidade pode ser útil para a compreensão do sentimento que conduz/representa o grupo, além de, aprender com a informação tende a ser propagada em um dado grupo particular.

- Composição: a possibilidade de estudar a composição do grupo além de um único meio social online (rede), pode complementar as avaliações referentes aos membros do grupo. Ou seja, quem são as pessoas que compõem o grupo? Em outras palavras, qual é a composição do grupo? Verificar a influencia de nós além da própria estrutura da rede, considerando a participação dos usuários em outros ambientes virtuais.

Os membros de grupos online não estão restritos a apenas um site de mídia social online, e isso é fundamental. Por exemplo, embora os membros de um grupo foram originalmente caracterizados com o perfil de jogos de esporte em uma rede, é muito

provável que muitos dos membros do grupo usem outras mídias sociais, tendo ligações também nessas. A capacidade de identificar indivíduos através de sites de mídia social podem fornecer informações valiosas sobre o grupo que o indivíduo pertence. Essas informações pode ajudar os pesquisadores a entender melhor como o grupo é formado e como ele pode crescer. Comerciantes podem utilizar estratégias adicionais para encorajar os consumidores a adentrarem em um dado grupo. Os cientistas sociais, podem obter uma melhor compreensão de como a informação é propagada de grupo em grupo, através de sites de mídia social heterogêneos.

De acordo com (BARBIER; TANG; LIU, 2011), as quatro tarefas se complementam; por exemplo, a compreensão da estrutura do grupo (perfil) sem o entendimento do sentimento do grupo (sentimento sobre os rótulos), não vai lhe dar uma avaliação precisa do seu status. Ter um perfil representativo para um grupo, mas não compreender sua estrutura, pode ser uma descrição equivocada. Ignorar os detalhes da composição do grupo, pode levar a falsas confianças e decisões mal informadas sobre o grupo. Além disso, a colaboração de cientistas sociais e outros especialistas podem fornecer perspectivas adicionais necessárias para uma compreensão exata e abrangente de um grupo particular. A seguir, serão apresentadas as principais abordagens de caracterização de grupos relacionadas a esta proposta.

2.3.1 Caracterização de Grupos Baseado em Agregação

Sabendo-se que o estudo de caracterização de grupos tem como objetivo encontrar as características que são compartilhadas por todo o grupo, uma abordagem natural e simples seria encontrar as características (atributos) que são mais prováveis de ocorrer dentro do grupo. Esse é o objetivo essencialmente da abordagem de caracterização de grupos baseada em agregação (CGBA) Barbier, Tang e Liu (2011), Tang e Liu (2010b), Tang, Wang e Liu (2010).

A abordagem de CGBA é simples, basicamente atributos individuais do grupo são agregados e um *ranking* desses é gerado, escolhendo os top-k (k trata-se do número limite de atributos, ou seja, k=10 seria selecionados os 10 principais atributos) atributos mais frequentes Tang, Wang e Liu (2011). Observando os atuais sistemas de etiquetagem, é verificado que a abordagem de CGBA é amplamente utilizada, na forma de nuvens de *tags*. Nuvens de *tags* são amplamente utilizados em mídias sociais para mostrar a popularidade de uma *tag* (atributos, marca, produto, palavras mais frequentes em fóruns, entre outros) pelo seu tamanho da fonte. Considerando-se toda a rede, como um grupo, em seguida, uma nuvem de marcação é produzida com base na agregação das características do grupo.

Todavia, esse método pode ser sensível em algumas situações. Normalmente em um fórum de rede social algumas palavras que não contribuem para a caracterização de um grupo podem vir a ser muito utilizadas. Isso resultaria em uma caracterização equivocada para o grupo analisado, não satisfazendo a propriedade em que os atributos selecionados

Tabela 2 – Estatísticas baseadas em Grupo e Atributo (dados binários)

	Grupo = +	Grupo = −
A=1	verdadeiro positivo (vp)	falso positivo (fp)
A=0	falso negativo (fn)	verdadeiro negativo (vn)

para caracterização dos grupos devem ser "Descritivos". A seguir, são apresentados alguns métodos da abordagem CGBA, dividindo-os em: critérios internos gerais, já aplicados à problemas de caracterização de comunidades; e para dados textuais, delineando técnicas convencionais de rotulagem de agrupamentos em documentos textuais.

2.3.1.1 Critérios Internos Gerais

Método Baseado em Agregação Tang, Wang e Liu (2011)

Suponha que existam n vértices em uma rede social G, e d atributos $\{A_1, A_2, ..., A_d\}$. Para um dado grupo g, estamos interessados nos atributos mais descritivos para sua caracterização. Dessa forma, podemos tratar os vértices do grupo como uma classe positiva (indicada como "+") e os demais pertencentes a classe negativa ("−"). Dessa forma, vértices pertencentes a classe positiva são chamados de casos positivos; idem para os casos negativos. Para um dado atributo A (composto pelo par (atributo, valor)) temos as seguintes estatísticas:

onde:

- Verdadeiros positivos (vp) é o número de casos positivos (vértices do grupo) que contêm um dado atributo A.
- Verdadeiro negativo (vn) é o número de casos negativos (vértices não presentes no grupo) que não contêm o atributo A.
- Falso positivo (fp) é o número de casos negativos contendo o atributo A.
- Falso negativo (fn) é o número de casos positivos não contendo o atributo A.

De posse dessas medidas, podemos calcular a probabilidade condicional de um atributo ocorrer em um dado grupo. Assim, a taxa de verdadeiro positivo (tvp), é a probabilidade condicional de um atributo ocorrer em um grupo. Em particular:

$$tvp = P(A|+) = \frac{vp}{vp + fn} \tag{2.9}$$

De posse dessas informações, (TANG; WANG; LIU, 2011) propuseram um método que calcula a relevância (*score*) de um atributo no grupo baseado em sua tvp, Equação 2.10. Basicamente são selecionados os top-k atributos mais frequentes no grupo analisado, ou

seja, os k atributos de maior $P(A_i|+)$. Ressaltando, que nesse método só são considerados os vértices do grupo, e que (TANG; WANG; LIU, 2011) consideraram apenas atributos binários.

$$\max_{\{A_i\}_{i=1}^{k}} \sum_{i=1}^{k} P(A_i|+) \tag{2.10}$$

2.3.1.2 Critérios Internos para Dados Textuais

Term Frequency (TF)

Datado em 1957 por (LUHN, 1957), o *Term Frequency* (frequência de termos) é um dos métodos mais antigos e simples de seleção de atributos em textos. Analisando um corpus de textos, os documentos que pertencem ao mesmo tópico apresentam maiores probabilidades de utilizarem palavras semelhantes. Portanto, os termos frequentes serão bons indicadores para um determinado tópico (grupo).

O método TF pondera o número de ocorrências do termo t em um cluster c como seu score de relevância, denotado por $tf_{t,c}$. Dessa forma:

$$tf_{t,c} = \begin{cases} 1 + \log(tf_{t,c}) & se\ tf_{t,c} > 0 \\ 0 & caso\ contrario \end{cases} \tag{2.11}$$

Os termos com elevados valores de TF são definidos como descritores candidatos para o cluster, os demais são desconsiderados. Ou seja, O peso de um termo que ocorre em um cluster é diretamente proporcional à sua frequência. Outra possibilidade é a seleção dos Top-K temos mais relevantes, sendo k um número inteiro previamente definido.

Centroid Labels

Um modelo frequentemente utilizado no campo da recuperação de informação é o modelo de espaço vetorial, esse representa documentos como vetores Salton (1989). As entradas no vetor correspondem aos termos no vocabulário. Muitos vetores fazem uso de pesos que refletem a importância de um termo em um documento, e/ou a importância do termo em uma coleção de documentos. Existem várias maneiras de calcular o centroide durante a fase de treinamento, e várias propostas de métodos baseados em centroides estão disponíveis na literatura. Uma das mais comum é a média e de acordo com (HAN; KARYPIS, 2000) cada centroide, $\vec{c_j}$, é representado pela média de todos os vetores dos documentos do grupo, como segue:

$$\vec{c_j} = \frac{1}{Dc_j} \cdot \sum_{\vec{d_i} \in Dc_j} \vec{d_i} \tag{2.12}$$

Se um termo no vetor do centroide tem um alto valor, subentendesse que o termo correspondente ocorre com frequência dentro do cluster. Esse termo pode ser utilizado como

um descritor (rótulo) para o cluster. Uma desvantagem desses dois métodos supracitados, *TF* e *Centroid Labels*, está na possibilidade de seleção de palavras frequentes, mas com pouca relevância semântica para o conteúdo do cluster.

Title labels

Uma alternativa para a rotulagem baseada no centroide é rotulagem pelo título dos documentos. Nesse método, para documentos intitulados, encontra-se no cluster o documento de menor distância euclidiana para o centroide, utilizando o seu título como um rótulo para o cluster Ramage et al. (2009). Uma vantagem de usar títulos de documentos é que eles fornecem informações adicionais que não estariam presente em uma lista de termos. No entanto, eles também têm o potencial de induzir o usuário final a erros, uma vez que um único documento dificilmente é suficientemente representativo para todo o cluster.

External knowledge labels

A Rotulagem de um agrupamento também pode ser feita indiretamente, utilizando conhecimentos externos pré-classificados. Dentre esses métodos, destacam-se os baseados em ontologias. Esses fazem uso de grandes fontes de dados estruturadas, como por exemplo, a Wikipédia Carmel, Roitman e Zwerdling (2009) ou WordNet Wei et al. (2015). Uma ontologia pode ser definida com uma representação formal do conhecimento, comumente apresentando conceitos que apresenta relações. Geralmente uma ontologia é utilizada para ligar palavras que representam o mesmo conceito (domínio). Dessa forma, tem-se a possibilidade de utilizar essa estrutura para identificar as palavras que pertencem ao mesmo conceito tópico. Diferentes métodos são usados para avaliar quais conceitos tópicos se encaixam melhor no texto de entrada Chahine et al. (2009).

Primeiramente, em tais métodos, um conjunto de termos representativos (importantes) para o agrupamento são extraídos de seus documentos. Esses termos, em seguida, podem ser usados para recuperar documentos classificados ou explorar seus conceitos em ontologias, selecionando-se os k documentos ou conceitos mais próximos, onde a partir desses os candidatos a rótulos do agrupamento podem ser extraídos. A etapa final envolve a classificação dos candidatos, podendo ser realizada por métodos baseados em votação ou fusão; o segundo, integra os documentos classificados e as características originais do cluster Carmel, Roitman e Zwerdling (2009). Ou ainda, a aplicação de algoritmos de busca hiperonímia para descrever os grupos através de um título genérico Wei et al. (2015).

Métodos Baseados em Grafos

Alguns trabalhos visam a construção de um grafo de termos a partir de agrupamentos em documentos. Conforme supracitado, algumas pesquisas exploram o uso de ontologias para caracterização de agrupamentos em documentos. A construção do grafos de termos

baseados em ontologias fazem uso da estrutura das ontologias (grafo de conhecimento) Kouznetsov e Zouaq (2014). Outros trabalhos geram os grafos definidos pelas coocorrências dos termos no texto Role e Nadif (2014). Já em (IENCO; MEO, 2008), é considerado como grafo de termos a estrutura hierárquica dos agrupamentos.

A partir dos grafos de termos alguns trabalhos exploram o uso de métodos baseados em grafo para quantificar a relevância dos tópicos/termos Kouznetsov e Zouaq (2014). Essas medidas são amplamente utilizadas na análise de redes complexas, podendo quantificar algumas das principais propriedades locais e globais de uma grafo Freeman (1979). No contexto da seleção de tópicos em documentos, destacam-se as medidas de centralidade. Essas facilitam o entendimento sobre os fluxos de informações no grafo, possibilitando verificar quais os vértices/tópicos têm maior importância para o grafo. Dessa forma, é possível determinar conceitos centrais em documentos/agrupamentos.

Uma vez que o gráfico de palavras é construído, o propósito da abordagem baseada em grafos é utilizar as medidas de centralidade para determinar os nós (nesse caso termos) mais relevantes. As medidas mais aplicadas são: Grau, Proximidade, Intermediação, *Eigenvector Centrality*, *TextRank*, *PageRank* e *HITS* Kouznetsov e Zouaq (2014), Boudin (2013).

Outra abordagem, também baseada em grafos, que vem surgindo é a *Contextualized centroid labels*. Essa, trata-se de uma maneira simples e econômica para superar a limitação da abordagem de *Centroid Labels* (subseção 2.3.1.2), ponderando os termos centroides, com maior peso, na estrutura de um grafo que forneça um contexto para a sua interpretação e seleção Role e Nadif (2014). Nesta abordagem, uma matriz de coocorrência termo-termo denominada T_k é construída para cada cluster S_k. Cada célula representa o número de vezes que termo o i coocorre com o termo j dentro de uma determinada janela de texto (uma frase, um parágrafo, etc.). Em uma segunda etapa, uma matriz de similaridade T_k^{sim} é obtida multiplicando T_k pela sua transposta. Temos $T_k^{sim} = T_k' T_k = (t_{sim_{ij}})$. Sendo o produto escalar de dois vetores normalizados $\tilde{t}_i$ e $\tilde{t}_j$, $t_{sim_{ij}}$ denota a similaridade do cosseno entre os termos i e j. Assim a matriz T_k^{sim} obtida pode ser utilizada como a matriz de adjacência ponderada de um grafo de similaridade de termo. Os termos centroides fazem parte deste grafo, portanto, podem ser interpretados e marcados para inspecionar os termos que os rodeiam no grafo.

O principal propósito do trabalho de (ROLE; NADIF, 2014) é apresentar métodos que vão além da rotulagem plana, permitindo que os usuários explorem as relações entre os termos nos clusters e, assim, melhor interpretar o conteúdo dos clusters.

Alguns estudos comparativos já foram realizados, avaliando as medidas baseadas em grafo para determinação de relevância de tópicos em documentos. Em (BOUDIN, 2013) realizou-se um estudo comparativo de cinco medidas de centralidade, para extração de frases chaves. O *Grau*, apesar de ser conceitualmente a medida mais simples, alcançou resultados compatíveis ao algoritmo *TextRank*, amplamente utilizado. Além disso, os resultados mostram que a *Proximidade* supera significativamente as outras medidas de

centralidade em documentos curtos. (KOUZNETSOV; ZOUAQ, 2014), apresentaram uma comparação de métodos não supervisionados e supervisionados para extração de frases chaves em corpus de domínio. As medidas baseadas em grafos são aplicadas em um grafo que conecta termos e expressões compostas por meio de relações conceituais, representando um corpus inteiro sobre um domínio, em vez de um único documento. Ambos os trabalhos, usaram conjuntos de dados com diferentes domínios.

2.3.2 Caracterização de Grupos Baseado em Diferenciação

Em vez de agregar, podemos selecionar as características que diferenciam um grupo de outras pessoas na rede (restante da rede). Assim, a abordagem de caracterização de grupos baseada em diferenciação (CGBD) seleciona etiquetas para as comunidades comparando a distribuição dos atributos em uma comunidade em relação aos outros nós da rede. Dessa forma, os nós são divididos em duas amostras, os que fazem parte do grupo e os demais.

Contudo, o objetivo da abordagem CGBD é descobrir as principais (top-k) características discriminativas que são representativas para um grupo, os diferenciando do restante da rede. Ou seja, essencialmente a abordagem seleciona atributos presentes em um grupo e que dificilmente aparecem nos outros. A seguir são apresentados os principais métodos da abordagem CGBD, dividindo-as de forma semelhante a CGBA.

2.3.2.1 Critérios de Gerais de Diferenciação

Bi-Normal Separation

Utilizando as mesmas estatísticas da Tabela 2, (TANG; WANG; LIU, 2011) propõem a aplicação do método Bi-Normal Separation (BNS) Forman (2003) para a abordagem baseada em diferenciação. Esse considera além da taxa de verdadeiro positivo (tvp, detalhes Seção 2.3.1.1), a taxa de falso positivo (tfp). A tfp é a probabilidade condicional de um atributo associado aos vértices que não fazem parte do grupo. Especificamente:

$$tfp = P(A|-) = \frac{fp}{fp + vn} \tag{2.13}$$

É um método eficaz que supera outros de seleção de atributos, tais como *information gain* e X^2 *statistic* Tang e Liu (2005) Forman (2003). A relevância (*score*) de um atributos usando o BNS é definido como:

$$BNS = |F^{-1}(tvp_A) - F^{-1}(tfp_A)|, \tag{2.14}$$

onde tvp_A e tfp_A são respectivamente, as taxas de verdadeiro e falso positivo (descrito nas equações 2.9 e 2.13), e F^{-1} é a função de probabilidade cumulativa inversa de uma distribuição normal padrão. Como a diferença de perfis de grupo discriminativo e a seleção

de atributos só se preocupam com as características que são descritivas para o grupo (a classe positiva), foi aplicada a seguinte restrição para os atributos selecionados.

$$tvp_A > tfp_A \tag{2.15}$$

Isso deve-se ao fato de o atributo A ter que ser mais descritivo em relação à classe positiva. Dessa forma, para a aplicação do BNS como CGBD combina-se a equação 2.14 e a restrição da 2.15. O objetivo mantêm-se, muda-se apenas o método que apontam a relevância do atributo, ou seja, seleciona-se os top-k atributos com maiores *scores* calculados pelo BNS, conforme descrito na Equação 2.16.

$$\max_{\{A_i\}_{i=1}^{k}} \sum_{i=1}^{k} |F^{-1}(tvp_{A_i}) - F^{-1}(tfp_{A_i})| \tag{2.16}$$

Teste Wilcoxon Rank Sum

Diferentemente de (TANG; WANG; LIU, 2011), onde os autores verificam a diferença entre as probabilidades condicional de um atributo dado o grupo e dado o restante da rede, (GOMES et al., 2013) trataram o problema de caracterização de grupos sociais, verificando as diferenças entre as duas distribuições, ou seja, o quão diferente é a distribuição de um atributo no grupo em relação ao restante da rede; selecionado os atributos com maiores diferenças.

Essa estratégia trata-se de um estudo estatístico que verifica a igualdade entre as distribuições dos atributos, no grupo e restante da rede. Em um estudo estatístico, a escolha de qual método utilizar está diretamente relacionada a forma pela qual os dados encontram-se distribuídos. Em caso dos dados não estarem em uma distribuição normal, métodos paramétricos, como o *t-student*, não são viáveis Kimball (1954). Logo, um método não-paramétrico é frequentemente uma escolha mais segura, sendo também mais robusto.

Partindo dos problemas apontados na utilização dos dados coletados de uma rede social, dados não normalizados e presença de atributos ruidosos, (GOMES et al., 2013) propuseram uma adaptação do teste estatístico não-paramétrico *Wilcoxon Rank Sum* (WRS) a uma abordagem baseada em diferenciação para geração de perfis caracterizadores de grupos. A seguir, são apresentados os passos realizados na aplicação do teste WRS para caracterização dos grupos:

1. Especifique um nível de significância limite α (por exemplo, 0.05), para indicar os atributos caracterizadores de um grupo;

2. Divide-se os dados coletados da rede social em duas amostras: grupo e restante da rede. Sobre as amostras, calcule o método estatístico WRS para cada atributo representativo dos usuários (todos os atributos selecionados na fase de *representação dos usuário*);

3. Use as estatísticas para calcular o valor correspondente de p;

4. Selecionar os atributos representativos cujos valores de p são inferiores ao α estabelecido, o que significa que a diferença entre o grupo e o restante da rede para o atributo em estudo é estatisticamente significante, apontando dessa forma, o atributo como um caracterizador do grupo. Ou ainda, seleciona-se os k atributos com menores valores de p.

As vantagens da aplicação do método estatístico WRS para construção de perfis de grupos sociais, está no fato de conseguir tornar a abordagem independente da distribuição dos dados. Ou seja, o método desenvolvido independente da forma como os dados estão distribuídos (em redes sociais os dados tendem a não estarem distribuídos segundo uma distribuição normal). Salientando, ainda, que em casos de distribuições normais, o teste WRS é quase tão eficiente quanto o teste *t-Student* Mei et al. (2008).

No trabalho de (GOMES et al., 2013), foi realizado um estudo de caracterização de grupos em uma rede social educacional. Os atributos analisando eram quantitativos (idade, quantidade de acessos, quantidade de acertos, etc). O estudo apresentou resultados bastante interessantes, todavia não foram confrontados com o estado da arte.

2.3.2.2 Critérios de Diferenciação para Dados Textuais

Inverse Document Frequency (IDF)

Como já visto, a métrica de TF (Seçao 2.3.1.2) pode prejudicar cálculos de similaridade entre documentos, pois termos com frequência de ocorrência menor do que outros podem possuir uma capacidade maior para discriminar documentos. Uma das maneiras de evitar esse problema é atribuir um peso alto quando um termo ocorre em poucos documentos e um peso baixo, caso contrário. Esse é o propósito da métrica IDF, calculada pela Equação 2.17:

$$IDF_t = \log \frac{N}{1 + df_t} \tag{2.17}$$

Onde N é o número total de nós em uma rede, e df_t é a quantidade de nós da rede N onde o atributo t aparece. Os termos com valores elevados IDF podem discriminar uma comunidade das outras, e, portanto, serem utilizados como candidatos a descritores da comunidade Maqbool e Babri (2005).

Ganho de Informação

O ganho de informação (do inglês Information Gain), também conhecido como informação mútua média (do inglês average mutual information) Yang e Pedersen (1997) ou perda esperada na entropia (do inglês expected entropy loss) Glover et al. (2002) é uma medida baseada na avaliação da capacidade de uma característica (termo) separar documentos em

categorias. Sendo uma medida de informação teórica que quantifica o grau de dependência de duas variáveis aleatórias.

Para avaliar a relevância de um termo, o ganho de informação mensura a redução esperada na pureza de uma coleção de documentos de treino (redução da entropia) causada pela divisão dos documentos de treino de acordo com uma característica (termo).

A entropia do conjunto de treino S é calculada pela Equação 2.18:

$$Entropia(S) \equiv \sum_{j=1}^{|C|} -Pr(c_j) \log Pr(c_j) \tag{2.18}$$

onde C é o conjunto de clusters (classes) e $Pr(c_j)$ é a proporção de documentos de treino no cluster c_j sobre o total de documentos de treino Mitchell (1997).

O ganho de informação do termo t é calculado pela Equação 2.19:

$$IG(S,t) \equiv Entropy(S) - \sum_{v \in (t,\bar{t})} \frac{S_v}{S} Entropy(S_v) \tag{2.19}$$

onde S é um conjunto de treino, S_t é um subconjunto de documentos de treino que possuem o termo t e $S_{\bar{t}}$ é um subconjunto de documentos de treino que não possuem o termo t. Durante rotulagem baseada em diferenciação, calcula-se o IG de cada termo no cluster e seleciona os k-termos com maiores valores de IG, sendo $k \in \mathbb{N}$.

Pointwise mutual information

Pointwise Mutual Information (PMI), ou ponto de informação mútua, é uma medida de associação usada na teoria da informação e estatística. Em contraste com a *Mutual Information* (MI), o PMI refere-se a eventos singulares, enquanto que IM refere-se à média de todos os eventos possíveis Evert (2004).

Nos campos da teoria da probabilidade e teoria da informação, a informação mútua mede o grau de dependência de duas variáveis aleatórias. A informação mútua de duas variáveis X e Y é definido como:

$$I(X,Y) = \sum_{x \in X} \sum_{y \in Y} p(x,y) \log_2 \left(\frac{p(x,y)}{p_1(x) p_2(y)} \right) \tag{2.20}$$

onde $p(x,y)$ é a distribuição de probabilidade conjunta das duas variáveis, $p_1(x)$ representa a distribuição de probabilidade de X, e $p_1(y)$ representa a distribuição de probabilidade de Y.

Na caso da rotulagem de agrupamento, a variável X está associada com a adesão de um cluster (C), e a variável Y é associada com a presença de um termo (T). Ambas

as variáveis podem ter valores de 0 ou 1, de modo que a equação pode ser reescrita do seguinte modo:

$$I(C,T) = \sum_{c\in 0,1}\sum_{t\in 0,1} p(C=c,T=t)\log_2\left(\frac{p(C=c,T=t)}{p(C=c)p(T=t)}\right) \tag{2.21}$$

Neste caso, $p(C=1)$ representa a probabilidade de que um documento selecionado aleatoriamente, é um membro de um cluster específico, e $p(C=0)$ representa a probabilidade de que ele não é. Do mesmo modo, $p(T=1)$ representa a probabilidade de que um documento selecionado aleatoriamente, contém um dado termo, e $p(t=0)$ representa a probabilidade de que isso não acontece. A função de distribuição de probabilidade conjunta $p(C,T)$, representa a probabilidade de que os dois eventos ocorrem simultaneamente. Por exemplo, $p(0,0)$ é a probabilidade de que um documento não é um membro do conjunto de C e não contém termo t; $p(0,1)$ é a probabilidade de que um documento não é um membro do cluster C e contém o termo t; e assim por diante.

Chi-Squared Selection (χ^2)

Teste do qui-quadrado de Pearson pode ser usado para calcular a probabilidade de que a ocorrência de um evento coincide com as expectativas iniciais. Em particular, ele pode ser usado para determinar se os dois eventos, A e B, são estatisticamente independentes[7] Witten e Frank (2005). O valor da estatística é:

$$\chi^2 = \sum_{a\in A}\sum_{b\in B}\frac{(O_{a,b}-E_{a,b})^2}{E_{a,b}} \tag{2.22}$$

onde $O_{a,b}$ é a frequência observada de coocorrência de a e b, e $E_{a,b}$ é a frequência esperada de coocorrência.

Na caso da rotulagem de cluster, a variável A está associado com a adesão de um cluster (C), e a variável B está associada com a presença de um termo (T). Ambas as variáveis podem ter valores de 0 ou 1, de modo que a equação pode ser reescrita como se segue:

$$\chi^2 = \sum_{a\in 0,1}\sum_{b\in 0,1}\frac{(O_{a,b}-E_{a,b})^2}{E_{a,b}} \tag{2.23}$$

Por exemplo, $O_{1,0}$ é o número observado de documentos que estão em um cluster particular, mas não contêm um certo termo, e $E_{1,0}$ é o número esperado de documentos que estão em um cluster particular, mas não contêm um certo termo. O nosso pressuposto

[7] Dois eventos são independentes se: $P(A\cap B) = P(A)P(B)$; o que equivale a: $P(A|B) = P(A)$ e $P(B|A) = P(B)$

inicial é que os dois eventos são independentes, de modo que as probabilidades esperadas de coocorrência podem ser calculadas multiplicando as probabilidades individuais:

$$E_{1,0} = N * p(C = 1) * p(T = 0) \tag{2.24}$$

onde N é o número total de documentos na base de dados. Se os dois eventos são dependentes, consequentemente a ocorrência do termo torna a ocorrência da classe mais provável (ou menos provavelmente), por isso deve ser útil como um atributo caracterizador. Essa é a lógica da utilização do χ^2 para seleção de atributos.

2.3.2.3 Critérios de Diferenciação Híbridos para Dados Textuais

Como já vimos, podemos distinguir duas abordagens básicas para rotulagem de agrupamentos em documentos: critérios internos, que selecionam etiquetas, baseando-se exclusivamente no conteúdo do próprio cluster, ao passo que a abordagem baseada em critérios de diferenciação, rotula um cluster comparando os seus termos com os termos que ocorrem em outros clusters. Combinando as características das duas abordagens básicas, temos o critério híbrido para dados textuais (*Hybrid Cluster Labeling*) (MANNING; RAGHAVAN; SCHüTZE, 2008). As próximas subseções descrevem a abordagem híbrida, apontando seus principais métodos.

Term Frequency-Inverse Document Frequency (TF-IDF)

O método IDF (Seção 2.3.2.2) pode atribuir pesos iguais para termos com alta frequência de ocorrência por não considerar a frequência de ocorrência de um termo no documento. Para contornar esse problema, é necessário que os termos que ocorram muito em determinado documento (frequência local alta) e, simultaneamente, que ocorram em poucos documentos (frequência global baixa) recebam pesos maiores. Esse é o propósito da métrica TF-IDF.

A TF-IDF é uma métrica que combina a TF com a IDF. Dessa forma, o peso total do termo t no cluster c se torna a combinação do seu peso local (a métrica TF) e global (a métrica IDF) Treeratpituk e Callan (2006). A $TF - IDF$ é calculada pela Equação 2.25.

$$tf\,idf_{t,c} = tf_{t,c} * idf_t \tag{2.25}$$

onde $tf_{t,c}$ é a TF do termo t no cluster c, calculada pela Equação 2.11, e idf_t é a IDF do termo t, calculada pela Equação 2.17.

Frequent and Predictive Words

Para o método Frequent and Predictive Words (em português, Palavras Frequentes e Preditivas), os termos caracterizadores são selecionados com base no produto da frequência

local e da predição Popescul e Ungar (2000), como segue:

$$p(t|C) \times \frac{p(t|C)}{p(t)} \tag{2.26}$$

A fórmula consiste de duas partes, cada uma com um significado bem definido: a predição $\frac{p(t|C)}{p(t)}$, é semelhante a um estimador de informação mútua e TF-IDF, uma medida que atribui maiores pesos aos termos que ocorrem com frequência em um determinado cluster e pesos menores aos termos que ocorrem com frequência em todos os clusters; $p(t|C)$ é a frequência da palavra em um determinado cluster e $p(t)$ é a frequência de um termo t em toda a coleção de documentos. Palavras que receberam valores elevados de predição são bons discriminadores para diferenciar um cluster de outro.

Latent Semantic Indexing (LSI)

O método Latent Semantic Indexing (LSI) tem sido desenvolvido para superar problemas como sinonímia e polisemia, recorrentes em abordagens vetoriais. Esse visa melhorias sobre o modelo de espaço vectorial base, substituindo a matriz de termo-documento original por uma aproximação. Isso é feito usando a decomposição singular valor (*Singular Value Decomposition*) e análise de componentes principais (*Principal Components Analysis*), técnicas originalmente usadas em processamento de sinal para redução de ruido, preservando o sinal original. Partindo do princípio de que a matriz termo-documento original é ruidosa (presença de sinonímia e polissemia), a aproximação é interpretada como uma redução de ruído, portanto, melhor que a matriz de termos original.

O método LSI foi adaptado por (KUHN; DUCASSE; GÍRBA, 2007) em uma nova maneira de rotular grupos de código fonte. A ideia-chave é reverter o processo de pesquisa padrão, onde os termos são utilizados na consulta para localizar documentos, e em vez disso, utilizam-se os documentos presentes no cluster como consulta para identificar os termos mais semelhantes, considerando-os os mais relevantes para caracterização. Dessa forma, para rotular um cluster, analisa-se o conjunto resultante final de termos selecionando os k-termos mais semelhantes, ressaltando que $k \in \mathbb{N}$.

Baseado em Conhecimentos Externos

Baseando-se em pesquisa sobre desambiguação e exploração das estruturas fornecidas pelo *DBpedia*, (HULPUS et al., 2013) propuseram uma nova abordagem para a rotulagem de tópico em textos, utilizam os dados estruturados expostos pelo *DBpedia*. Partindo da hipótese de que, palavras que co-ocorrem no texto provavelmente se referem a conceitos que estão altamente conectados (próximos) no gráfico DBpedia. A partir rede semântica, (HULPUS et al., 2013) aplicaram medidas de centralidade em grafo, para identificação dos conceitos que melhor representam os tópicos.

Uma hipótese importante desse trabalho é que os grafos de conceitos do tópico analisado, são mais propensos a tornar-se conectados do que a busca de conceitos aleatórios no *DBpedia*. Para termos com diferentes conceitos, foi aplicado o método de desambiguação, proposto em Hulpus et al. (2012). De posse do grafo de conceitos do termo analisado, aplicou-se as medidas de grafo baseada em centralidade para identificar os conceitos mais centrais no que diz respeito aos conceitos fundamentais. Algumas medidas foram melhores adaptadas para o problema, renomadas como "*focused centralities*". As medidas utilizadas foram: Grau, *Closeness Centrality on All Graph*, *Focused Closeness Centrality*, *Focused Information Centrality*, *Betweenness Centrality on all graph*, *Focused Betweenness Centrality* e *Focused Random Walk Betweenness*.

Combinação de Medidas Estatísticas e Conhecimentos Externos

Sabendo-se que as abordagens atuais concentram-se sobre a utilização de recursos estatísticos dos documentos agrupados ou fontes externas, como a Wikipedia. (HENNIG et al., 2014) combinaram essas ideias para lucrar com ambas as vantagens e criaram um algoritmo, que pode lidar com documentos em cluster, bem como termos.

Para tanto, os autores criaram critérios para definir rótulos únicos, adequados para cada agrupamento (nó) na hierarquia, esses são:

- O rótulo deve ser geral o suficiente para descrever todos os documentos ou termos do agrupamento (em uma hierarquia de cluster de documentos um nó pai contém todos os documentos de seus filhos).
- O rótulo deve ser específico o suficiente para distinguir o agrupamento de seus nós filhos e irmãos.
- Por fim, o rótulo em um cluster pai deve ser uma generalização para os clusters filho e, consequentemente, os rótulos dos clusters filhos, uma especialização do cluster pai.

Dessa forma, foram propostos dois algoritmos que automaticamente encontram rótulos para clusters hierárquicos binários, onde cada cluster pode ter no máximo dois filhos na hierarquia. O primeiro é uma otimização ao algoritmo *Descriptive Score* (*DScore*) de (TREERATPITUK; CALLAN, 2006) (visto em Treeratpituk e Callan (2006)). Esse depende das características estatísticas dos termos na hierarquia dos agrupamentos, visando adaptar o *DScore* à hierarquias de clusters dispersos. Para rotular um cluster, considera-se as características estatísticas dos termos no cluster em si, os seus filhos e seu cluster irmão. O segundo algoritmo utiliza a rede categoria DBpedia Lehmann et al. (2015) para encontrar rótulos apropriados. Enquanto o primeiro algoritmo pode ser aplicado apenas a clusters de documentos; o segundo algoritmo também pode ser aplicado a cluster de termos. Assim, os algoritmos se complementam, uma vez que o segundo algoritmo utiliza apenas um

conjunto de termos relevantes para cada cluster, que pode ser gerado aplicando o primeiro algoritmo.

2.3.3 Caracterização de grupos baseada em diferenciação egocêntrica

Na abordagem de diferenciação anterior, todos os nós de fora de um grupo são considerados como pertencendo à classe negativa. No entanto, o cálculo baseado em diferenciação de grupos aplicado a uma visão global (considerando-se todos os nós de uma rede), pode ser considerado um desperdício de processamento. A escalabilidade também pode ser uma preocupação. Redes sociais online mais populares são enormes, com centenas de milhões de nós. Dessa forma, é necessário um elevado tempo de processamento para recuperação de todas as informações de uma rede do mundo real. Ressaltando que em algumas aplicações, somente uma vista egocêntrica está disponível. Em outras palavras, pessoas têm informações sobre os seus amigos, mas pouco conhecimento sobre as pessoas que não conhecem.

Nesse contexto, (TANG; WANG; LIU, 2011) propuseram a abordagem caracterização de grupos baseado em diferenciação egocêntrica (CGBDE), que em vez de diferenciar um grupo de toda a rede, o diferencia dos vizinhos[8](*fringe*) dos membros do grupo, ou seja, perfis de grupo com base na visão egocêntrica. A Figura 5, representa as três abordagens supracitadas.

Para a abordagem egocêntrica, a única mudança em relação a diferenciação "global", é que a classe negativa passa a ser o *fringe*, e não todos os demais nós (global). Dessa forma, todos os métodos baseados em diferenciação global, podem ser adaptados/aplicados em uma abordagem egocêntrica.

(TANG; WANG; LIU, 2011) realizaram um estudo comparando as três abordagens de caracterização de grupos supracitadas. Para realização do estudo comparativo, selecionaram como bases de dados os sites de mídia social: BlogCatalog[9] e LiveJournal[10]. Ao final do estudo, foi descoberto que as abordagem baseadas em diferenciação (CGBD e CGBDE) sempre superavam a baseada em agregação (CGBA). Isso tornou-se mais perceptível quando os atributos individuais utilizados para caracterização apresentavam ruídos, como por exemplo: atributos coletados postagens de usuários, registros de atividades e relatos de usuários. Ou seja, a aplicação da abordagem de CGBA aplicada a dados ruidosos, selecionou uma grande quantidade de atributos que não descreviam o grupo corretamente.

Dessa forma, (TANG; WANG; LIU, 2011) concluíram que a abordagem de CGBA somente é aplicável em um ambiente relativamente livre de ruído. Pois, se os perfis são construídos a partir de atributos ruidosos, como posts de usuários, registro de atividades de usuários

[8] Vizinhos de um grupo se referem aos nós fora de um grupo, que são conectados a pelo menos um membro do grupo.
[9] http://www.blogcatalog.com/
[10] http://www.livejournal.com/

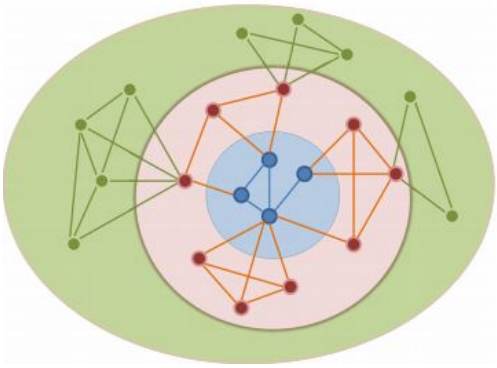

Figura 5 – Representação das três principais abordagens de caracterização de grupos sociais: CGBA, CGBD e CGBDE. Nós azuis no centro formam um grupo. Nós vermelhos na área rosa são os vizinhos do grupo (*fringe*/visão egocêntrica). Fonte: Tang, Wang e Liu (2011).

ou relatos de interesses, abordagens baseadas em diferenciação, apresentam melhores perfis para caracterização dos grupos.

Como visto muitos dos estudos sobre comunidades em redes sociais têm sido direcionados a compreensão e, principalmente, detecção dessas em ambientes sociais. Tradicionalmente as comunidades são detectadas com base na topologia das redes, na busca por melhorias uma das principais tendências recentes tem sido a consideração dos atributos na tarefa de detecção. Martin-Borregon et al. (2014) consideram os diferentes processos de agregação entre usuários em ambientes sociais, esse trabalho explora a detecção e formação de grupos a partir de três características: distribuição geográfica (espaço), temporal (evolução) e social. Já Quercia, Aiello e Schifanella (2016), têm por foco a identificação de comunidades que têm "desviantes", essas possuem usuários que têm por características condutas que são comumente consideradas inadequadas em relação às normas ou padrões morais da sociedade. Acredita-se comumente que tais comunidades desviantes são grupos sociais isolados, de nicho, cuja atividade está bem separada da vida social da grande mídia. De acordo com essa suposição, os estudos de pesquisa os consideraram em isolamento. Com base nisso, Quercia, Aiello e Schifanella (2016), concentraram-se em redes de consumo de conteúdo adulto, que estão presentes em muitas mídias sociais on-line e na Web em geral.

Outros trabalhos têm focado na detecção temporal das comunidades, investigando a evolução da estrutura de redes temporais, que consistem em nós e arestas que variam no tempo. Apesar de fugir ao foco principal desta tese, esse problema é de importância crucial no campo da análise de redes sociais. A estrutura da comunidade na rede temporal contribui para a compreensão do processo de evolução das entidades no sistema complexo. A abordagem tradicional de detecção dinâmica da comunidade desconsidera os diversos instantes no seu processo de identificação. Isso tem por consequência uma baixa eficiência por ignorar informações históricas da comunidade. Neste contexto, He e Chen (2015) apresentaram um algoritmo rápido para detecção de comunidade dinâmica em redes temporais, que propõe aproveitar as informações da comunidade no momento anterior

buscando melhorias na eficiência, mantendo a qualidade. Já em He et al. (2017), é sugerido um método mais simples que aproveita as informações históricas da comunidade para detectar comunidades dinâmicas. Basicamente, depois de dividir cada comunidade no passo anterior em alguns módulos, não se usa esses módulos apenas para detectar as comunidades no momento atual, mas também mapear as comunidades nas etapas do tempo.

Já em Appel et al. (2019), busca-se a detecção e predição de comunidades em redes sócias considerando e combinando múltiplas informações, links, conteúdo e análise temporal. Essa abordagem funciona extraindo a estrutura semântica latente da rede de forma multidimensional, considerando ainda, a continuidade temporal dessas incorporações. O objetivo central da abordagem é simplifica a análise temporal da rede subjacente usando a incorporação como um substituto. Uma consequência dessa simplificação é possibilitar, também, o uso dessa sequência temporal de incorporações para prever futuras comunidades. Maiores detalhes sobre estudos de detecção de comunidades temporais em Dakiche et al. (2018). Esse trabalho apresenta um levantamento de estudos anteriores realizados sobre o problema de rastrear a evolução das comunidades ao longo do tempo em redes sociais dinâmicas.

Outra vertente de estudo das comunidades, têm levado em consideração as suas tipificações. Ou seja, antes de estudar adentro a natureza de cada comunidade, afirma-se que se faz necessário, verificar qual o seu tipo sociológico. O propósito é verificar se os resultados computacionais são equivalentes aos consolidados conceitos da sociologia S. (1981), Walther (1997). Apesar do crescimento continuo das pesquisas sobre redes sociais on-line nos últimos anos, pouca atenção tem sido dedicada à exploração da natureza dos links sociais. Interações on-line têm sido interpretadas como indicativas de um processo social ou outro, por exemplo, troca de status ou confiança. Todavia, essa análise em muitas vezes justificam de forma superficial as relações entre os dados observados e os conceitos teóricos. Nesse contexto, Aiello, Schifanella e State (2014) procuram romper essa lacuna na ciência social computacional, propondo um método não supervisionado e livre de parâmetros, na busca pela descoberta dos domínios fundamentais de interação, consolidados nas redes sociais não virtuais. Como continuidade, em Aiello (2017), Aiello (2015), além da busca pelo entendimento das comunidades on-line ao longo de suas características espaço-temporais e de atividades, buscam fomentar uma nova visão para estudar as agregações sociais em suas dimensões sociais e tópicas.

Como se pode observar, a grande maioria dos trabalhos relacionais recentes não está diretamente idealizada à caracterização/descrição das comunidades. Esses têm como inovação a consideração dos atributos individuais dos usuários, mas com objetivos diferentes da rotulagem das comunidades. Pode-se citar como exemplo Appel et al. (2019), esse inova na consideração dos atributos, mas tem por foco melhorias na detecção das comunidades; bem como Aiello (2017), Aiello (2015), Aiello, Schifanella e State (2014), que direciona o

foco para o estudo da natureza das comunidades. Entender a natureza das comunidades nos remete a forte indicadores de melhorias nas descrições, caracterização dos grupos, o que nos direciona a uma abordagem híbrida a ser testada como um trabalho futuro. Bem como, Wang et al. (2014), direciona a possibilidades de abordagens conjuntas de detecção e rotulagem para comunidades.

2.4 CONSIDERAÇÕES FINAIS

Neste capítulo, foram apresentados os principais conceitos sobre as redes complexas. A análise de redes complexas é uma das áreas que surgiram recentemente com o intuito de estudar grandes volumes de dados modelados como redes (grafos). Algumas de suas principais medidas de centralidade foram discutidas nesse capítulo, propriedades utilizadas diretamente na abordagem proposta nesta tese.

Uma propriedade relevante em diversas redes complexas é a existência de grupos de nós densamente conectados, denominados comunidades. O estudo de métodos para caracterização desses grupos tem despertados interesses da comunidade científica. Neste capítulo foram descritas as principais abordagens de caracterização de grupos, entre elas a baseada em diferenciação (CGBD), a qual é um dos focos de estudo desse trabalho.

Este estudo bibliográfico foi de grande importância para o embasamento da proposta aqui levantada. Uma vez compreendidos e explorados esses métodos e abordagens, podemos adaptá-los de forma adequada ao problema em questão obtendo assim melhores resultados e ganhos. No Capítulo 3, apresentamos nossa proposta e delineamos cada um dos seus módulos. Como estudo de caso inicial, verificamos a viabilidade da aplicação de técnicas convencionais de rotulagem de agrupamentos em documento textuais ao contexto de caracterização de grupos sociais. Confrontando, posteriormente, os resultados obtidos a uma caracterização baseada em informações relacionais. Por fim ainda avaliamos o pré-processamento dos dados iniciais através da representação de usuários baseada em comunidades. Foram aplicados os métodos: TF-IDF, IDF, WRS, BNS e Qui-Quadrado, os experimentos são descritos em detalhes no Capítulo 4.

3 CARACTERIZAÇÃO DE GRUPOS BASEADA EM INFORMAÇÕES RELACIONAIS

Como já visto no Capítulo 2, a ARS é um amplo campo de pesquisa que relaciona técnicas, estratégias e métricas para o estudo das redes. A análise e a extração de conhecimento das redes são largamente empregadas. Dentre essas, a compreensão do comportamento e das tendências das comunidades é uma atividade estratégica Wasserman e Faust (1994). De acordo com (TANG; WANG; LIU, 2011), é desejável que todo método de caracterização de grupos satisfaça as seguintes propriedades: descritivo, robusto e escalável.

As redes sociais online mais populares são enormes, com centenas de milhões de nós. Dessa forma, é necessário um elevado tempo de processamento para recuperação de todas as informações. Obviamente o ideal é que todos os nós autores sejam considerados para a caracterização de grupos durante o processo de rotulagem. Isto incluirá mais conhecimento para ser propagado ao longo da rede. Entretanto, como já apontado, um dos principais pilares de um bom método de caracterização é a escalabilidade.

Normalmente se tem a intenção de reduzir o tamanho do conjunto de dados utilizado ao mínimo possível; seja por dificuldades na coleta (dados ruidosos, duração, entre outros) ou pelo elevado custo computacionais na rotulagem das comunidades. Conforme já visto na seção REF a abordagem CGBDE proposta por (TANG; WANG; LIU, 2011), busca isso com base em uma visão egocêntrica das comunidades. Todavia, essa abordagem apresenta algumas limitações, tais como: perda do contexto/visão global da rede, inaplicabilidade em casos de comunidades isoladas e em situações de “poucos” nós vizinhos, uma vez que algumas comunidades podem apresentar uma visão egocêntrica limitada, com poucos vizinhos ou com nós de baixa representatividade (não relevantes para a caracterização).

É visível a carência por abordagens eficazes na filtragem de nós e conteúdos, reduzindo os dados de entrada e mantendo as informações relevantes para a caracterização das comunidades. Uma abordagem imediata seria realizar uma seleção aleatória entre os nós das comunidades. Se o conjunto resultante representar o todo, a caracterização pode apresentar um bom desempenho. Todavia, a variabilidade dos resultados pode ser grande e dependeríamos totalmente da sorte na obtenção de uma amostra representativa.

Outros aspectos das abordagens atuais de caracterização de grupos sociais são: (1) caracterização baseada apenas nos atributos dos usuários, (2) níveis de relevâncias equivalentes a todos os usuários e (3) consideração de todos os usuários da comunidade na caracterização. Todavia, em ambientes nos quais haja conexões entre seus usuários (como as redes sociais), uma nova dimensão de informação se apresenta, através da análise dos relacionamentos e afinidades entre os usuários (informação relacional).

As informações relacionais podem promover grandes benefícios ao estudo e compreensão de comunidades em redes sociais, apresentando-se como o grande diferencial em relação às

abordagens de rotulagem de agrupamentos em documentos. De modo geral, independente da abordagem de caracterização de comunidades adotada, agregação, diferenciação global ou egocêntrica, os métodos atuais somente fazem uso de atributos dos usuários para a caracterização das comunidades. Sabendo-se que a rotulagem de comunidades sociais é realizada em um cenário que há informação relacional disponível, abordagens que considerem essa nova dimensão de informação são necessárias.

Ao se analisar uma comunidade, um aspecto importante é identificar quais são os nós ou os links mais importantes (ou centrais), pois esses podem revelar algumas peculiaridades relevantes sobre a comunidade. Normalmente utilizam-se as medidas de centralidade como forma de quantificar essa importância. A noção de centralidade, em várias aplicações, é associada à importância do elemento na estrutura, no problema apresentado, na comunidade. Espera-se, por exemplo, elevados[1] índices de centralidade de nó para uma pessoa influente em um grupo social. Essa informação tende a apresentar relevante atuação na análise dos dados, ou seja, nós centrais podem: generalizar conteúdos, apresentar maior influência na comunidade, ponderar a caracterização, entre outras possibilidades.

Buscando melhorias sobre os três pilares definidos por (TANG; WANG; LIU, 2011) e a incorporação dos benefícios supracitados com a exploração das informações relacionais, esta tese propõe uma abordagem para caracterização de grupos sociais baseada em informações relacionais. Nas próximas seções são apresentados, em detalhes, a representação do problema e a metodologia proposta.

3.1 REPRESENTAÇÃO DO PROBLEMA

Nesta seção, é apresentada uma descrição formal para o problema de caracterização de grupos sociais. Para tanto, consideramos que os dados (rede) são representados na forma de grafo. Formalmente temos como entrada um grafo $G = (V, E)$, onde $V = \{v_1, v_1, ..., v_n\}$ é o conjunto de vértices e $E \subseteq V \times V$ são as arestas (um grafo com n vértices). Para simplificar, assumimos um grafo não direcionado sem auto-relacionamento, ou seja, $(v_i, v_j) \in E \iff (v_j, v_i) \in E$ e $(v_i, v_i) \notin E$. Adicionalmente, cada vértice é associado a um vetor de dimensão d, $\mathbf{a} \in \mathbb{D}^d, \mathbf{a} = (a_1, a_2, ..., a_d)$ em que $\mathbb{D}$ corresponde ao domínio do atributo (por exemplo, $\{0, 1\}$ ou $\mathbb{R}$). Por exemplo, um nó pode ser associado a um conjunto de tópicos de interesse, que pode ser modelado por atributos binários: $a_j \in \{0, 1\}$, em que $a_j = 1$ indica o interesse do usuário no j-ésimo tópico.

Um grupo/comunidade é representado por um subgrafo $P = (V_P, E_P)$, onde $V_P \subseteq V$, $E_P \subseteq V_P \times V_P$ e $E_P \subseteq E$. Em uma caracterização de grupos, um perfil de um dado grupo é definido por um vetor de atributos $\mathbf{c}_P \in \mathbb{D}^k, k \leq d$. Assim, há um total de $\binom{d}{k}$ possibilidades distintas de caracterização para cada grupo. O objetivo é selecionar os melhores atributos descritivos k para cada grupo $\mathbf{c}_P$. Para tal, pode-se definir uma função

[1] Considerando, obviamente, que quanto maior o valor da centralidade, mais central o nó.

de pontuação $f(a, P)$ para atribuir a importância (ou seja, pontuação descritiva) para cada atributo em uma determinada partição e, em seguida, selecionar os k atributos com maiores pontuações (*scores*).

Dessa forma, tradicionalmente a representação de usuários, subconjunto de atributos $a' \subseteq a$, seleciona um número $y \in \mathbb{N}$ de atributos considerados semanticamente relevantes[2] ao domínio da rede. Esse modelo, devido o desbalanceamento entre as comunidades, pode apresentar tratamentos desiguais nas representações dos usuários, onde comunidades maiores tende a ser mais beneficiadas em relação às menores. Visando tratar essa lacuna, foi proposta uma representação baseada em comunidades, essa direciona um olhar singular para cada comunidade durante o processo de representação, evitando que alguma comunidade seja tratada de forma superficial. Para tanto, são selecionados $w \in \mathbb{N}$ atributos para cada comunidade P_i, ou seja, $a'_{P_i} \subseteq a_{P_i}$, onde a_{P_i} é o conjunto de atributos da comunidade P_i. Para garantir um tratamento equivalente/equilibrado a todas as comunidades, considera-se $w \leq y \div p$, sendo p a quantidade de comunidades do grafo G e sabendo-se que alguns atributos a_i podem estar presentes em mais de uma comunidade P_i. Posteriormente, realiza-se a integração de todos os subconjunto de atributos das comunidades em um único conjunto de dados, $a' = a'_{P_1} \cup a'_{P_2} \cup ... \cup a'_{P_i} \cup ... \cup a'_{P_p}$, finalizando a representação de usuários baseada em comunidades. A representação de usuários baseada em comunidades é detalhada na Seção 3.2.2.

A matriz de dados de entrada para a caracterização de grupos pode ser considerada como M_{nxd}, rede de n nós com d características. Verifica-se que o tamanho dos dados de entrada (M_{nxd}) estão diretamente relacionados a quantidade dos nós (V) e ao conjunto de atributos ($\mathbf{a}$). Como já visto anteriormente, os ambientes de mídias sociais tendem a apresentar crescimentos contínuos em ambos sentido; surgimento de novos usuários e relacionamentos, que resultam consequentemente, em uma produção exponencial de informações. Diante disso, nota-se que uma abordagem efetiva de caracterização de grupos deve ser capaz de reduzir as dimensões da matriz de entrada (M_{nxd}) ao limite mínimo possível, garantindo simultaneamente, a não exclusão das informações relevantes à caracterização. Dessa forma, temos três possibilidades de redução: quantidade de nós (V), de atributos ($\mathbf{a}$) ou ambos.

Ao se analisar uma comunidade, um aspecto importante é decidir quais são os nós mais importantes ou centrais. A forma mais tradicional e usada para quantificar a relevância de nó na rede é através das medidas de centralidade. A centralidade de nó é uma função $c_G : V \to \mathbb{R}$ que possui certas propriedades, descritas a seguir. Primeiro, c_G deve ser um índice estrutural de G, ou seja, se G e H são grafos isomorfos por ϕ, então $c_G(\phi(v)) = c_H(v), \forall\ v \subseteq V$. Outra propriedade, matematicamente menos precisa e mais dependente do contexto, é que a relação de ordem entre $c_G(v_i)$ e $c_G(v_j)$ deve refletir a percepção de que v_i é mais central que v_j, em algum sentido; para a maioria das medidas, quanto mais

[2] Ressaltando que y é um número consideravelmente superior a k (atributos selecionados como caracterizadores).

central for o vértice v, maior o valor de $c_G(v)$, porém pode valer o oposto. Usaremos a notação c_G para referenciar o vetor contendo a centralidade de todos os vértices, ou seja, $c_G = (c_G(v_1), c_G(v_2), ..., c_G(v_n))$.

A abordagem baseada em informações relacionais, proposta nesta tese, visa a redução de n, linhas (nós) da matriz de dados M_{nxd}, a partir da centralidade do conjunto de nós (c_G). Dessa forma, para cada comunidade $P_i = (V_{P_i}, E_{P_i}) \in G$, calcula-se a centralidade de todos os seus nós ($V_P \in G$), obtendo a centralidade de todos os vértices na comunidade analisada (c_P). De posse dos vetores de centralidade c_P de cada comunidade P, seleciona-se o subconjunto de nós mais centrais. Apenas o subconjunto de nós $V_{P_c} \subseteq V_P$ é considerado no processo de caracterização; ou seja, para geração da matriz de dados M, somente serão coletadas e processadas informações dos vértices V_{P_c} (Detalhes do filtro de informações relacionais são apresentados na Seção 3.2.1). Com a redução das linhas n (vértices) da matriz de dados, também tendemos a reduzir as colunas d (atributos considerados para geração de perfis), uma vez que os atributos dos vértices desconsiderados na filtragem não serão coletados. O propósito da abordagem é selecionar os vértices centrais (mais relevantes) que generalizem o conteúdo do conjunto de comunidades P, almejando a redução do custo computacional no processo de geração dos perfis ($\mathbf{c}_P$), sem perdas na qualidade dos perfis gerados (descrição). Dessa forma, considerando-se o filtro de informações relacionais com a representação de usuários baseada em comunidades, tem-se a redução da dimensão da matriz resultante $M_{n \times d}$ para $M_{t \times w}$, com $t \leq n$ e $w \leq d$.

3.2 METODOLOGIA PROPOSTA

Nesta seção, é apresentada a metodologia proposta para caracterização de grupos sociais baseada em informações relacionais.

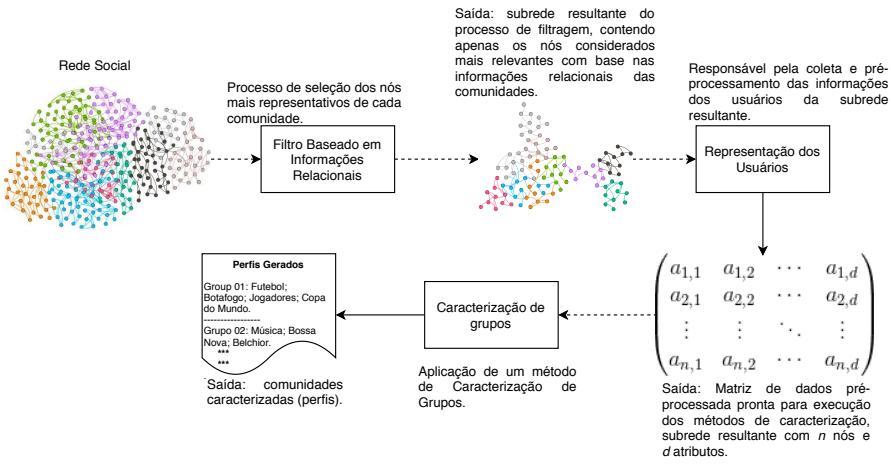

Figura 6 – Visão geral da metodologia de caracterização de grupos baseada em informações relacionais.

A metodologia é definida em três etapas: inicialmente é realizado o (1) **Filtro Baseado em Informações Relacionais**, processo de seleção dos nós mais representativos de cada comunidade. Posteriormente inicia-se o pré-processamento sobre a base de dados, identificando os principais atributos individuais dos usuários selecionados, (2) fase de **Representação dos Usuários**. Finalmente, na etapa (3) **Caracterização das Comunidades**, são aplicados métodos de caracterização de grupos para extração dos perfis. Sabendo-se que o propósito da caracterização é descrever as comunidades e não detectá-las, considerou-se as comunidades como estando explicitas na rede. Em casos de estarem implícitas, se faz necessário as suas identificações via algoritmo de detecção de comunidades. Na Figura 6, a arquitetura desenvolvida é apresentada. Nas subseções seguintes, serão apresentados e detalhados cada um dos seus módulos.

3.2.1 Filtro Baseado em Informações Relacionais

Na busca por abordagens escaláveis, normalmente, propõe-se a redução do conjunto de dados considerados para a caracterização de grupos Tang, Wang e Liu (2011). Ou seja, são necessárias abordagens capazes de selecionar apenas informações relevantes ao domínio, desconsiderando as sem valores semânticos do processo de caracterização. Buscando esse objetivo propomos o Filtro Baseado em Informações Relacionais. O módulo tem como propósito o desenvolvimento de um novo filtro ao processo de caracterização de grupos. Para tanto, considera-se a aplicação de métricas relacionadas aos nós e suas posições na rede, identificando o quão central é um determinado nó em uma determinada comunidade Wasserman e Faust (1994), Scott e Carrington (2011). Tais informações proporcionam a identificação dos nós mais relevantes/influentes nas comunidades. Esses apresentam um grande potencial em um possível processo de filtragem de informações.

Presumivelmente, todas as comunidades têm os seus usuários influentes. Estes são os líderes de opinião e podem desempenhar um papel mais importante para refletir as peculiaridades de uma comunidade. Outro fator que impulsiona o uso das informações relacionais é a facilidade da sua aplicação, uma vez que são propriedades presentes nas redes. Independente da abordagem de caracterização de comunidades adotada, baseada em agregação, diferenciação global ou egocêntrica, os métodos atuais somente fazem uso de atributos dos usuários para a caracterização das comunidades. Alguns trabalhos exploram as informações relacionais para inferir as principais características de usuários a partir das características de seus contatos, obtendo bons resultados e reduzindo consideravelmente a complexidade Yuan e Gay (2006), McPherson, Smith-Lovin e Cook (2001b), Bollen et al. (2011); bem como, (LIU et al., 2005) que fez uso de medidas de centralidade na verificação dos autores mais proeminentes em uma rede de coautoria. Ou seja, a consideração das informações relacionais na caracterização dos grupos é pertinente. Como a rotulagem de grupos sociais é realizada em um cenário com informações relacionais disponíveis, métodos que façam uso dessa nova dimensão de informação são necessários, pois os tradicionais as

negligenciam.

O problema aqui descrito é semelhante ao tratado por métodos baseados em grafo para rotulagem de agrupamentos em documentos Role e Nadif (2014), Ienco e Meo (2008), Role e Nadif (2014). A partir dos grafos de termos alguns trabalhos exploram o uso de métodos baseados em grafo, para quantificar a relevância dos tópicos/termos (selecionando os termos que generalizam o agrupamento). Nesta tese, propomos uma abordagem que considere as informações relacionais durante o processo de caracterização dos grupos. Assim, é adicionado um novo módulo ao processo de caracterização de grupos, esse é responsável por filtrar os principais nós das comunidades a partir das informações relacionais, ou seja, selecionar os nós que serão considerados no processo de caracterização dos grupos.

Dessa forma, é interessante que a subrede resultante do filtro, possua nós bem conectados e ricos em conteúdos relevantes para caracterização. Ou seja, o propósito é selecionar os nós, que representem/generalizem as comunidades, produzindo os melhores perfis possíveis para as comunidades, sem perdas de informações relevantes. Vislumbramos com isso a redução dos custos computacionais, mantendo os perfis em bom nível de descrição. A Figura 8 apresenta uma modelagem do processo de seleção de nós realizado pelo Filtro de Baseado em Informações Relacionais.

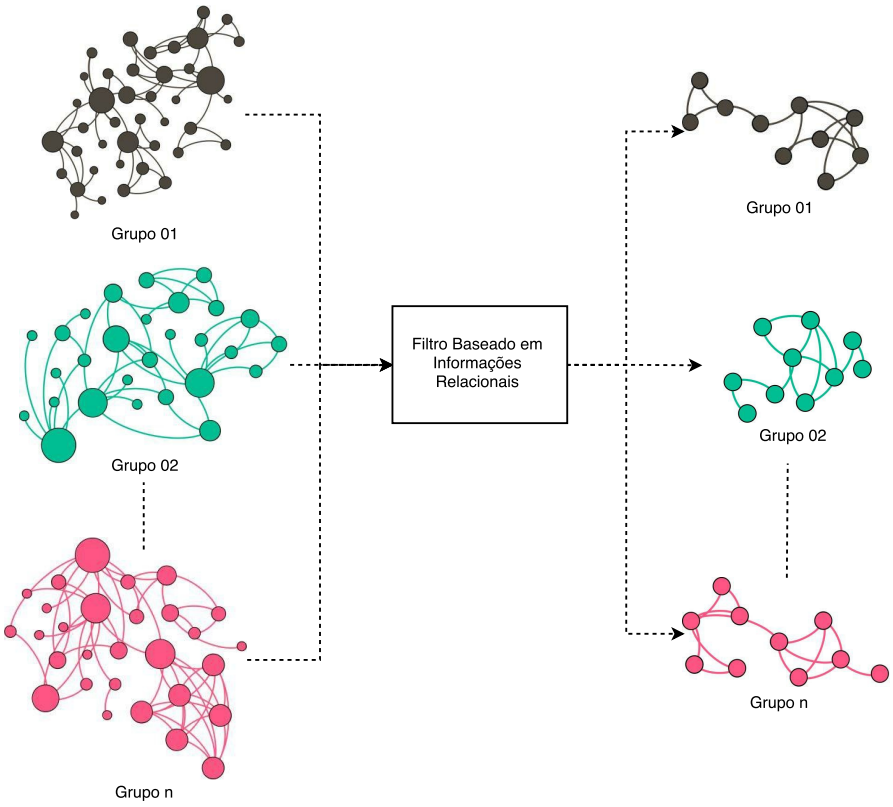

Figura 7 – Visão geral do processo de seleção de nós realizado pelo Filtro de Baseado em Informações Relacionais. Imagem gerada usando o framework Gephi para a Rede arXiv, mesma utilizada no experimentos.

Observe que apesar de estarmos adicionando mais uma etapa ao processo de caracte-

rização de grupos, verifica-se que o principal propósito da abordagem é buscar avanços no tratamento da escalabilidade. De fato o principal gargalo, atividade mais onerosa, na abordagem global (Seção 2.3) tende a ser a grande quantidade de dados a serem processados para a Representação dos Usuários. O tempo gasto no pré-processamento dos atributos é reduzido quando está limitado ao conteúdo gerado pelos nós centrais, reduzindo o processo de Representação dos Usuários (detalhes na Seção 3.2.2). Logo, o tempo economizado no pré-processamento tende a compensar o custo adicional dos cálculos da centralidade dos nós no processo de filtragem.

Todavia, a escolha entre as abordagens, baseada em centralidade ou não, dependerá do tamanho do gráfico $G = (V, E)$ e da dimensão do vetor d de recursos associados a cada vértice $v_i \in V$ considerado. Portanto, trata-se de um problema de compensação (do inglês, *tradeoff problem*), dependendo das entradas em cada módulo/etapa. Sabendo-se que a inclusão da etapa de filtragem diminui a quantidade de dados considerados, e que isso reflete consequentemente nas duas etapas subsequentes (Representação de Usuários e Caracterização das Comunidades), evidencia-se que a abordagem baseada em informações relacionais tende a apresentar vantagens em relação a complexidade computacional global. Na próxima Seção, é apresentado o módulo de Representação dos Usuários, nela ficará claro o quão oneroso é o processamento dos atributos, reafirmando a necessidade de tratamento/filtragem de usuários a priori.

3.2.2 Representação dos Usuários Baseada em Comunidades

A principal dificuldade desse módulo encontra-se em considerar apenas as informações relevantes, desconsiderando tudo que não for importante à caracterização. Normalmente se erra de duas formas na representação dos usuários: (1) seleção de atributos não relevantes ao domínio dos grupos, resultando na possibilidade de inclusão de características sem valores semânticos, que podem ser selecionadas e descreverem equivocadamente as comunidades; (2) desconsideração de atributos essenciais a alguma comunidade, esse erro apresenta uma consequência pior, pois o anterior pode ser corrigindo pelo método de caracterização (não selecionando os atributos), todavia a ausência da característica desconsidera todas as possibilidades de seleção do atributo na caracterização das comunidades, elevando consideravelmente a probabilidade de descrições incorretas das comunidades.

Tradicionalmente se considera todos os atributos em um único conjunto de dados (para atributos textuais, saco de palavras) e sobre esse é executado o módulo de representação de usuários, eliminando as características consideradas irrelevantes. Na prática um bom módulo (com boas técnicas de pré-processamento) obterá um resultado favorável, todavia esse processo desconsidera uma avaliação particular sobre cada comunidade, podendo levar a uma representação limitada de alguma(s) comunidade(s). Sabendo-se que a representação dos usuários é à base de toda a caracterização, se faz necessário esse olhar singular a cada comunidade durante esse processo.

Buscando resolver os problemas supracitados, propomos a Representação dos Usuários Baseada em Comunidades. Nesta abordagem, aplicamos as técnicas de seleção de atributos sobre os conjuntos de características de cada comunidade individualmente. Após termos a execução para todas as comunidades, realizamos a integração de todas as representações em um único subconjunto de características. Essa abordagem tende a ser mais equilibrada, garantindo que todas as comunidades possuam um universo de características relevantes que às representem. Evitando, assim, que os usuários de uma comunidade sejam bem representados, enquanto outra comunidade não possua uma boa representação para seus membros. A seguir, na Figura 8, é apresentada uma visão geral do funcionamento do módulo de Representação dos usuários Baseado em Comunidades.

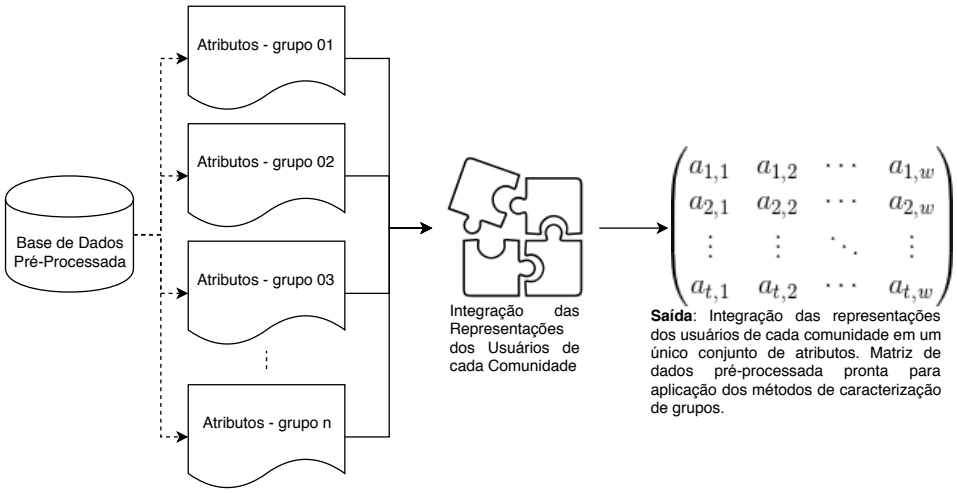

Figura 8 – Visão geral do processo de representação de usuários baseada em comunidades.

O módulo de representação de usuários proposto é realizado em dois passos. Inicialmente são coletados da base de dados os atributos individuais dos usuários de todas as comunidades (usuários considerados para caracterização), todos os atributos de cada comunidade; posteriormente, sobre esses são aplicadas técnicas de pré-processamento, realizando o processo de filtragem/seleção; De posse dos atributos selecionados para representar os usuários de cada comunidade, realiza-se a integração de todos em uma única representação. Objetiva-se que a representação de usuários seja simultaneamente, eficaz na identificação dos atributos relevantes e equilibrada para todas as comunidades. Por isso, módulo tende a favorecendo os métodos de caracterização na geração de perfis descritivos.

3.2.3 Caracterização das Comunidades

Finalmente neste módulo da metodologia os atributos caracterizadores de cada comunidade são selecionados. De posse da matriz de dados resultante da Representação dos Usuários, o módulo de Caracterização das Comunidades identifica os principais atributos de cada comunidade, gerando os perfis. Na Figura 9 o processo é ilustrado. Explanaremos o

que já foi supracitado inicialmente, a metodologia proposta independente do método de caracterização adotado. Assim, iremos dispor das possíveis possibilidades de abordagem, de acordo com a literatura.

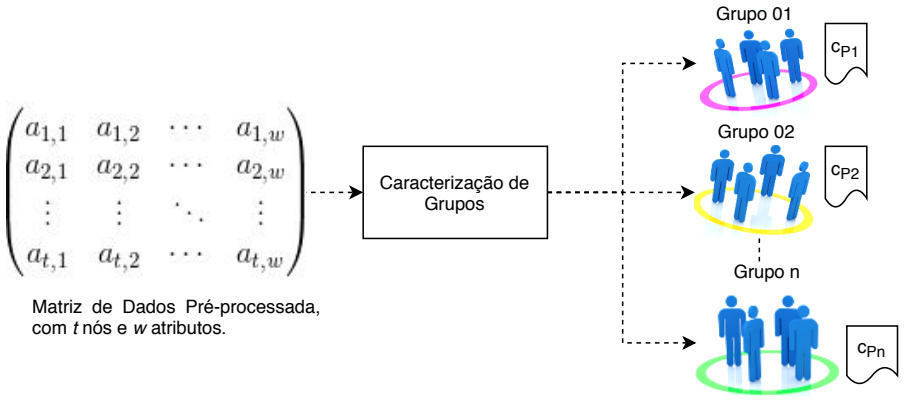

Figura 9 – Visão geral do processo de geração dos perfis. Dada a matriz de dados pré-processada, o método de caracterização identifica e seleciona os atributos caracterizadores das comunidades (perfis).

3.3 CONSIDERAÇÕES FINAIS

Este capítulo descreveu a abordagem proposta, reverenciando sua relevância para área de caracterização de grupos sociais, através de argumentos baseados na literatura. Para um melhor entendimento da arquitetura, foram delineados cada um dos seus módulos.

Dessa forma, vimos que o Filtro Baseado em Informações Relacionais visa à seleção dos principais nós autores a partir das informações relacionais, e que para a abordagem, somente os nós selecionados serão considerados no processo de caracterização dos grupos. Ou seja, a ideia é reduzir a matriz de dados, selecionando apenas nós que representem/generalizem as comunidades. Independente o método de caracterização adotado, baseado em diferenciação ou aglomeração, o módulo de filtragem baseado em centralidade pode ser adotado/adicionado. O que evidencia os benefícios proporcionados com as informações relacionais na caracterização de comunidades. Considera-las no processo tende a proporcionar grandes benefícios ao estudo e compreensão de comunidades em redes sociais, podendo se apresentar como uma grande solução para correção dos gargalos da área.

Outro módulo apresentado foi o de Representação dos Usuários. Sabendo-se que a representação dos usuários é a base de toda a caracterização, se faz necessário um olhar singular a cada comunidade durante esse processo. Assim, propomos a Representação dos Usuários Baseada em Comunidades. Nesta abordagem de representação, aplicamos as técnicas de seleção de atributos sobre os conjuntos de características de cada comunidade individualmente. Após termos a execução para todas as comunidades, realizamos a integração de todas as representações em um único subconjunto de características. Essa abordagem tende a ser mais equilibrada, garantindo que todas as comunidades possuam

um universo de características relevantes que às representem. Evitando, assim, que os usuários de uma comunidade sejam bem representados, enquanto outra comunidade não possua uma boa representação para seus membros. Por fim, no módulo de Caracterização de Grupos, evidenciamos o que já havíamos defendido anteriormente, todos os métodos de caracterização podem ser adotados pela abordagem proposta, sejam eles baseados em agregação, diferenciação (global ou egocêntrica), ou até mesmo hierárquicos.

Buscamos com essa metodologia melhorias sobre os três princípios básicos de uma abordagem de caracterização de grupos sociais, apontados por (TANG; WANG; LIU, 2011). Dessa forma, almejamos uma melhor *Descrição*, através da Representação dos Usuários Baseada em Comunidades (entendimento individual); *Robustez*, através da aplicação de técnicas sofisticadas de pré-processamento (Processamento de Linguagem Natural (PLN), filtragem de atributos, etc), efetivas no tratamento sobre os dados ruidosos; *Escalabilidade*, através do Filtro Baseado em Informações Relacionais, almeja-se reduzir consideravelmente os dados de entrada sem perdas na qualidade, tendo como consequência uma redução do tempo de execução global do processo de caracterização. A seguir, no Capítulo 4, os experimentos realizados são delineados, assim como, apresentamos a viabilidade da abordagem de Caracterização de Grupos Baseada em Informações Relacionais através de exemplos práticos.

4 EXPERIMENTOS

Como visto nos capítulos anteriores, um dos objetivos da ARS é compreender o processo de formação e evolução das redes. A caracterização de grupos, portanto, tem papel fundamental nesse processo investigativo, pois delineia as estruturas da rede, em termos de grupos, descrevendo-os em perfis e buscando justificar as suas formações.

Neste trabalho, foi desenvolvido um estudo de caso com o objetivo de caracterizar comunidades de pesquisadores. As comunidades desempenham um papel crucial nas redes de coautoria, uma vez que refletem a base dos relacionamentos de colaboração entre os autores. Há tempos se percebe que a análise de redes de coautoria pode favorecer na compreensão da estrutura e evolução das sociedades acadêmicas correspondentes. Embora amplamente estudada em seus aspectos estáticos e evolutivos, pouca atenção tem sido dedicada à exploração da natureza das suas comunidades Martin-Borregon et al. (2014). Tais dados normalmente são usados para analisar previsões de novas relações entre colaboradores, ou seja, predição de relacionamentos Lü e Zhou (2011), Albert e Barabási (2002), Lopes et al. (2010).

Existem vários motivos que podem levar à formação de comunidades em redes de coautoria Savić et al. (2015). Assim, uma questão importante seria: podemos descrever/justificar as comunidades científicas em uma rede de coautoria? A resposta para essa questão nós proporcionaria consideráveis avanços na compreensão das estruturas acadêmicas. A investigação das características e peculiaridades das comunidades pode levar a novas abordagens que podem fortalecer e expandir as redes de colaboração científica.

Dessa forma, o presente estudo de caso tem por objetivo validar empiricamente a abordagem proposta, e para tanto, realizamos três experimentos: Seção 4.2 - estudo comparativo com a adaptação e aplicação de diferentes técnicas de caracterização de grupos sociais em um rede de coautoria; Seção 4.3 - verificação do uso das informações relacionais no processo de caracterização de comunidade, caracterização de grupos baseada em informações relacionais; por fim, Seção 4.4 - análise do módulo de representação de usuários baseado em comunidades. A seguir a configuração inicial de toda a base de dados é apresentada, bem como o processo de avaliação qualitativa adotado. Por fim, são apresentadas as considerações finais sobre o experimento.

4.1 CONFIGURAÇÃO DOS EXPERIMENTOS

Nesta seção são apresentadas as configurações adotadas sobre a base de dados considerada no estudo, bem como todo o processo de avaliação qualitativa. Logo, as demais subseções descrevem o conjunto de dados da rede de coautoria, o procedimento de detecção de comunidades, os métodos utilizados e as estratégias de avaliação para analisar os perfis

atribuídos pelos métodos de caracterização de grupos considerados.

4.1.1 Base de Dados

Para avaliação do estudo, uma rede de coautoria foi utilizada nos experimentos realizados. As redes de coautoria representam a colaboração de pesquisadores em trabalhos científicos. Os nós são autores e as arestas ocorrem quando dois autores escrevem um artigo juntos Albert e Barabási (2002). Redes de coautoria estão entre as redes sociais mais estudadas, dada a importância de compreender a estrutura e a evolução das sociedades acadêmicas. Uma quantidade significativa de dados abertamente acessíveis sobre publicações científicas tem incentivado o uso de métodos de ARS para analisar a estrutura de coautoria. O objetivo desse estudo de caso consiste em caracterizar as comunidades, descrevendo-as e buscando justificar as suas formações.

Optou-se por esse tipo de rede pelo fato da colaboração científica, ser uma estrutura inerentemente evolutiva, o que faz dela uma fonte de dados ideal para a investigação das formações de grupos. Além disso, as redes de coautorias são propícias à atividade de caracterização de grupos, pois se trata de uma rede onde os usuários geram consideráveis quantidades de informações nas suas interações, em sua grande maioria, motivados por interesse acadêmicos comuns. Descrever comunidades de coautoria pode auxiliar pesquisadores na formação de novas colaborações científicas, inclusão de novos membros aos grupos, surgimento e integração de comunidades, entre outras possibilidades. Estudos demonstram que grupos de pesquisa bem conectados tendem a ser mais produtivos Lopes et al. (2010). Apesar de não ser uma tarefa trivial, tendo em vista a grande quantidade de informações e propriedades consideradas, a caracterização dos grupos pode facilitar o entendimento das estruturas acadêmicas, visualização e navegação da rede, acompanhamento a mudanças de temas de um grupo, casos alarmantes e sugestões/divulgações direcionadas a grupos específicos.

4.1.2 Coleta de Dados e Geração da Rede

Para realização dos experimentos, selecionou-se como base de dados a biblioteca digital arXiv[1]. Essa se trata de uma biblioteca digital operada pela *Cornell University* e apresenta 1.416.072 documentos cadastrados (informação coletada em 15 de julho de 2018), envolvendo as seguintes áreas: física, matemática, ciência da computação, biologia, financias e estatísticas. Optou-se pela arXiv, uma vez que essa disponibiliza uma API (*Application programming interface*) que fornece todo o suporte para coleta das informações necessárias para geração da rede e realização do estudo aqui proposto.

As informações consideradas para geração da rede foram: título, resumo, data de publicação, área e autor(es). Na Figura 10 apresentamos como exemplo um artigo coletada

[1] Disponível em: <http://arxiv.org/>

da área de computação simbólica (código “cs.SC”). Essas informações foram serializadas em arquivos XML e armazenadas em um repositório local, com arquivos individuais para cada artigo.

```
<article id="1404.0002v1">
<title>Factoring Differential Operators in n Variables</title>
<abstract>In this paper, we present a new algorithm and an experimental implementation
for factoring elements in the polynomial n'th Weyl algebra, the polynomial n'th
shift algebra, and ZZ^n-graded polynomials in the n'th q-Weyl algebra.
  The most unexpected result is that this noncommutative problem of factoring
partial differential operators can be approached effectively by reducing it to
the problem of solving systems of polynomial equations over a commutative ring.
In the case where a given polynomial is ZZ^n-graded, we can reduce the problem
completely to factoring an element in a commutative multivariate polynomial
ring.
  The implementation in Singular is effective on a broad range of polynomials
and increases the ability of computer algebra systems to address this important
problem. We compare the performance and output of our algorithm with other
implementations in commodity computer algebra systems on nontrivial examples.</abstract>
<published>01/03/2014</published>
<category>cs.SC</category>
<author>Mark Giesbrecht</author>
<author>Albert Heinle</author>
<author>Viktor Levandovskyy</author>
</article>
```

Figura 10 – Template XML para aquisição do corpos

O processo de aquisição foi implementado através de um *script* (cliente *WebService*) que faz requisição aos repositórios de artigos. Esse *script* pode ser substituído por qualquer cliente *WebService* com mesmas funcionalidades e padrões. O povoamento da base de dados recebe como entrada os arquivos XML, e inicia o processo de desserialização desses arquivos. Utilizou-se um *parser* para processar cada arquivo que guarda as informações sobre um artigo, e os dados assim obtidos são armazenados na base de dados.

A Figura 11 apresenta o modelo lógico da base de dados que representa a rede utilizada, assim como as tabelas auxiliares utilizadas para a geração das redes. As entidades básicas do modelo (‘paper’ e ‘author’) são utilizadas para armazenamento dos corpora de artigos, no qual um autor pode escrever um ou vários artigos e um artigo pode ser escrito por um ou vários autores. Essa relação é caracterizada como um relacionamento de M para N. Nesse tipo de relacionamento é gerada uma nova tabela associativa (‘author-paper’) que armazena as chaves estrangeiras das entidades básicas. A coluna ‘date-paper’ (data de publicação do paper) é utilizada para definir os intervalos de tempo para coleta de dados da rede de coautoria. Os nós representam os autores, e as ligações, os artigos publicados em coautoria. O número de coautorias de um artigo é calculado da seguinte forma: $(n \times (n-1))/2$, levando em consideração um artigo com n autores ($n > 1$). O conteúdo formado pelo titulo e resumo de cada artigo, após a fase de processamento de texto (detalhes das técnicas aplicadas na Seção 3.2.2) é armazenado na tabela ‘author-document’, aqui temos a unificação de todas as informações de cada autor. Finalmente, para representar os relacionamentos (coautoria) entre os nós da rede, foi criada a tabela auxiliar ‘co-authored-a’, essa indica os relacionamentos entre os nós apresentando em um

mesmo registro identificadores dos autores que já publicaram ao menos um artigo juntos. Ressaltando que para esse estudo os links não foram ponderados.

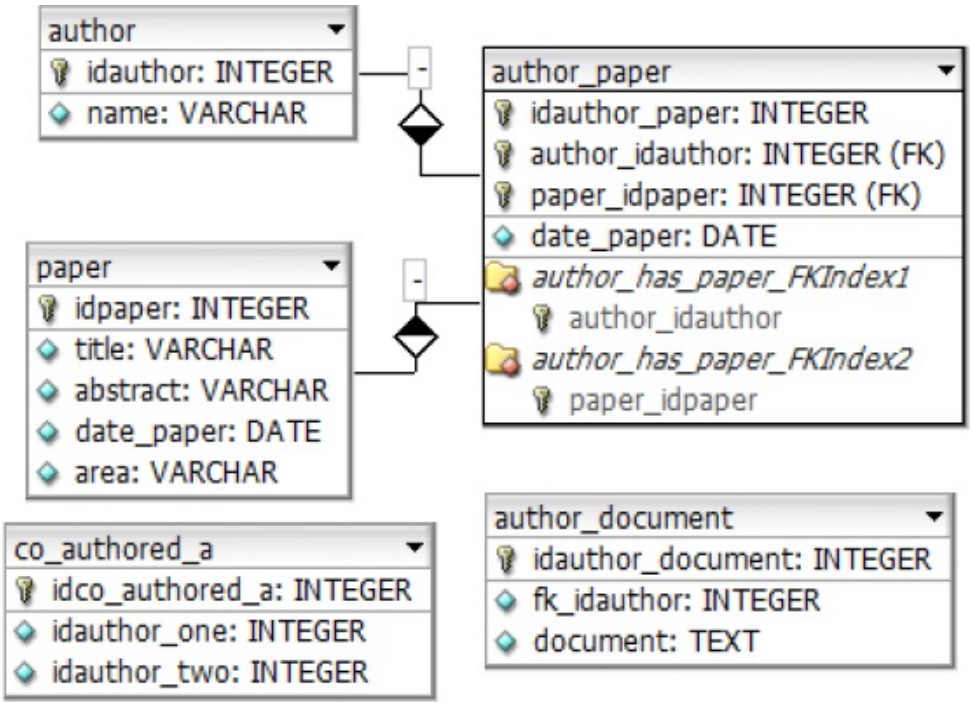

Figura 11 – Modelo lógico da base de dados da rede de coautoria

Como a base de dados não apresenta perfis anotados das comunidades, para as avaliações dos experimentos foram necessárias análises qualitativas feitas por humanos especialista. Visando facilitar o encontro de avaliadores qualificados para os experimentos, apenas artigos no campo de inteligência artificial ("cs.IA"), publicados entre 2012 e 2014, foram considerados neste estudo. Na rede gerada, cada nó representa um autor, e dois nós estão conectados se eles publicaram juntos de pelo menos um artigo. A rede resultante, pós-coleta, contem 1850 autores com 2560 relacionamentos. As principais estatísticas da rede são apresentados na Tabela 3.

Tabela 3 – Estatísticas da rede de coautoria arXiv de inteligencia artificia, entre os anos 2012 e 2014

#Autores	1850
#Links	2560
Densidade	0.001
Grau Médio	2.768
Diâmetro	19

Utilizou-se como ferramenta para visualização e geração da rede a API Gephi[2]. O Gephi é o software livre utilizado para visualização e exploração de grafos, apresentando

[2] www.gephi.org/

considerável suporte às estatísticas tradicionais Bastian, Heymann e Jacomy (2009). Os nós possuem todas as informações selecionadas na tabela 'author-document', e as arestas representam as coautorias coletadas da tabela 'co-authored-a'. De posse da estrutura da rede, avança-se para a fase de identificação das comunidades, detalhada a seguir.

4.1.3 Detecção de Comunidades

Estudos realizados sobre ARS com a biblioteca arXiv não apresentaram nenhuma aplicação no contexto de grupos, não havendo comunidades explícitas. Dessa forma, aplicou-se o algoritmo de detecção de comunidades Multi-level Aggregation Method (MAM) para identificação dos grupos. Esse é um método de passos múltiplos baseado na modularidade local otimizada Girvan e Newman (2002). Segundo (FORTUNATO; LANCICHINETTI, 2009), o MAM está entre os métodos que apresentaram melhor desempenho sobre redes não direcionadas e não ponderada, tais como a utilizada no presente estudo; além de ser muito rápido, essencialmente em redes de tamanho linear. Para maiores detalhes sobre a implementação do algoritmo MAM verificar Seção 2.2.2.

Antes da aplicação do algoritmo MAM para a identificação das comunidades na rede, realizou-se o pré-processamento sobre os dados coletados. Como o objetivo do estudo é caracterizar comunidades científicas a partir dos atributos individuais dos autores, os nós isolados (*singletons*) foram removidos, uma vez que esses autores não poderiam estar presentes em nenhuma comunidade. Em seguida, o algoritmo de detecção de comunidades foi executado sobre a rede, obtendo 439 comunidades. Este número considerável de comunidades detectadas deve-se principalmente a identificação de um elevado número de comunidades pequenas e isoladas, formadas por nós com poucos relacionamentos e não conectadas a nenhuma outra comunidade maior.

Com o objetivo de selecionar somente as comunidades mais conectadas para os experimentos de caracterização de grupos, realizou-se alguns processamentos de seleção sobre as mesmas de forma empírica. Grupos com menos de 10 autores foram removidos, pois foram considerados muito pequenos e irrelevantes para o estudo. Para os grupos restantes, calculou-se a densidade de cada um, essa é uma medida comum que verifica quão bem conectada a rede é, ou seja, no presente estudo, quão unida são as comunidades Tang e Liu (2010a).

Logo executou-se diversos filtros com base na densidade na busca de um quantitativo de comunidades relevante e por uma rede resultante densamente conectada. Obteve-se como rede resultante pós filtro um grafo com dez comunidades, realizando um *ranking* dos 10 grupos mais densos (TOP-10). As estatísticas do conjunto de dados resultante para a rede selecionada são sumarizadas na Tabela 4, onde são apresentados: o número de usuários e *links*, a densidade, grau médio, diâmetro e o número de grupos selecionados. Na Tabela também podemos verificar o estado original da rede, comparando-o com o pós-filtragem sobre as comunidades. Já a plotagem da rede resultante é apresentada na

Figura 12.

Tabela 4 – Medidas relacionando a rede de coautoria arXiv inicial e pós a aplicação do filtro sobre as comunidades.

Medidas	Original	Filtrada
#Autores	1850	372
#Links	2560	654
Densidade	0.001	0.009
Grau	2.768	3.516
Diâmetro	19	19
Número de Comunidades	439	10

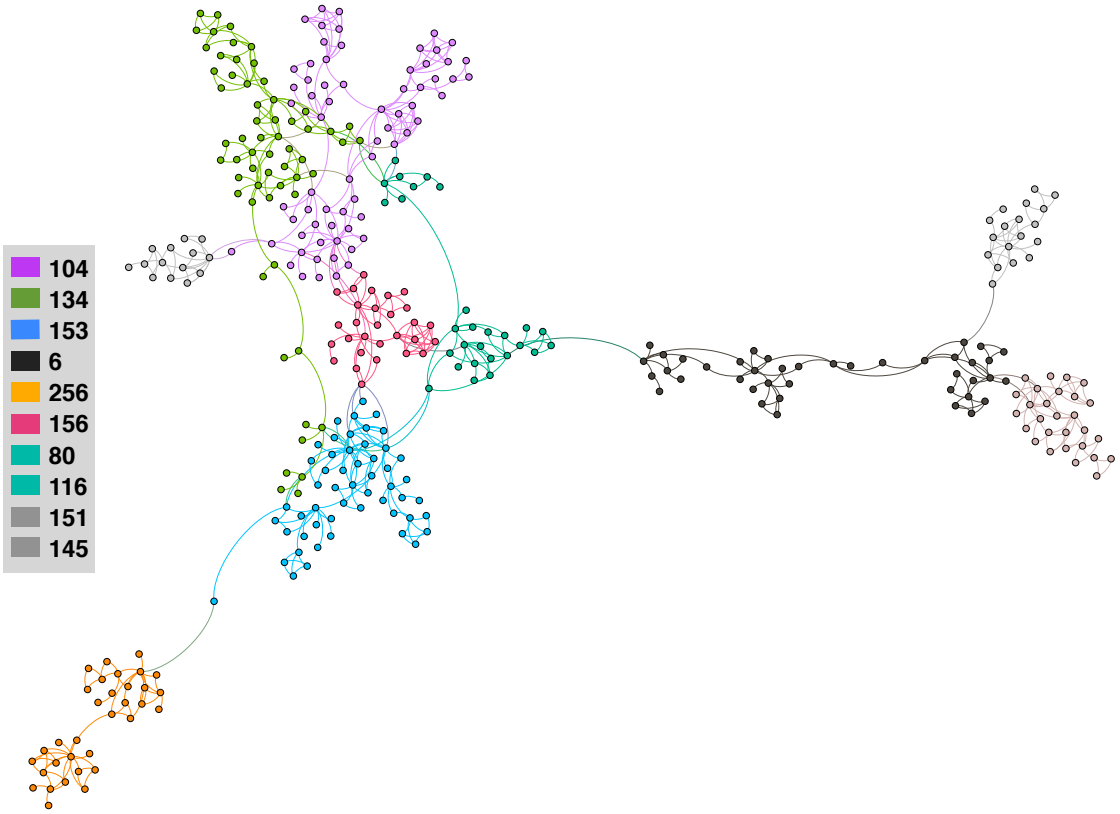

Figura 12 – Rede resultante e comunidades detectadas.

Analisando a Tabela 4, pode-se verificar que o propósito da filtragem foi alcançado, obtendo um considerável ganho na densidade da rede resultante (nove vezes maior). As estatísticas de cada grupo são apresentados na Tabela 5, incluindo o tamanho, densidade, grau médio e coesão. A coesão do conteúdo das comunidades foi calculada com base na métrica dos cossenos Cha (2007). Para tanto, se verificou a similaridade entre o conteúdo de cada artigo do grupo confrontado com o conteúdo de todos os demais artigos nele

presentes, a média geral dos valores dos cossenos obtida para cada artigo representa a grau de coesão das informações do grupo. A ideia desse cálculo é verificar se as comunidades apresentam pesquisadores que atuam em tema latente, ou seja, artigos predominantes vinculados a uma área de pesquisa; ou contrário em casos de comunidades formadas por pesquisadores atuantes em diferentes vertentes de pesquisa.

Analisando-se a Figura 12, inicialmente é notado que os grupos 151 e 256 demonstram ser os mais distantes das demais comunidades, verificando que a grande maioria dos seus relacionamentos são entre seus próprios autores, além de serem grupos visivelmente bem conectados. Já as comunidades 104, 134 e 153 são facilmente visualizadas como as comunidades com maior número de autores. Outro grupo que chama a atenção é o grupo 116, esse demostra ter bom número de links entre seus usuários, se apresentando como uma comunidade bem conectada. Na figura, ainda notamos uma boa conectividade entre as comunidades, não apresentando nenhuma comunidade isolada na rede resultante, reafirmando a efetividade do processo de seleção da rede resultante.

Tabela 5 – Estatísticas das comunidades

Comunidades	**Tamanho**	**Grau Médio**	**Densidade**	**Coesão**
6	41	2.829	7.1%	0.248
80	28	2.929	10.8%	0.384
104	68	3.735	5.6%	0.307
116	28	3.929	14.6%	0.363
134	60	3.167	5.4%	0.299
145	14	3.429	26.4%	0.377
151	18	3.333	19.6%	0.346
153	53	3.321	6.4%	0.337
156	29	3.724	13.3%	0.352
256	33	3.273	10.2%	0.316

A partir dos dados apresentados na Tabela 5, é possível verificar as análises realizadas inicialmente sobre a Figura 12. O grupo 104, 134 e 153 são de fato os grupos maiores, apresentando uma quantidade considerável de autores em relação os demais grupos. Também é possível identificar na tabela, comunidades que são mais densas que outras, por exemplo, as comunidades de 145 e 134. Indiferente, o grupo 116 apresenta o maior grau médio. Entre as comunidades mais coesas, destacam-se os grupos 80 e 145.

4.1.4 Representação dos Usuários

Neste módulo, os atributos relevantes para a caracterização da rede e posteriormente dos grupos, são selecionados a partir do conjunto de dados. Basicamente é realizado um filtro

sobre os dados coletados da base, objetivando a desconsideração dos atributos irrelevantes, sem influência semântica na caracterização das comunidades analisadas.

O módulo de representação de usuários é realizado em dois passos. Inicialmente são (a) coletados da base de dados os atributos individuais dos usuários, e posteriormente sobre esses é realizado o (b) processo de filtragem (seleção), resultando no conjunto de atributos "mais representativos" para caracterização das comunidades da rede social em estudo. Salientando que a realização da filtragem dos atributos representativos da rede não é uma tarefa trivial, se faz necessário o acompanhamento e validação por especialistas do domínio da rede em estudo. Do contrário, poderá resultar em resultados incompletos ou até mesmo incorretos. Como por exemplo: a exclusão indevida de um atributo representativo.

Os atributos utilizados para caracterização das comunidades na rede arXiv, são dados textuais coletados dos títulos e resumo dos artigos de cada autor. Cada nó da rede (isto é, autor) está associado às informações de suas publicações, que descrevem as suas características como pesquisador. E cada atributo corresponde a uma palavra ou composição de palavras (por exemplo, "inteligência artificial").

Dessa forma, para aquisição do conjunto de atributos foram realizados procedimentos de pré-processamento textual. Em primeiro lugar, todos os trabalhos publicados de cada autor foram coletados e combinados em um único documento (conforme detalhado na Seção 4.1.2). Em seguida, técnicas de processamento de linguagem natural foram aplicadas para cada conjunto de artigos de um autor (documento). Abaixo são listadas as etapas/técnicas aplicadas no pré-processamento dos documentos.

- Tokenização: é utilizada para decompor o documento em termos que o compõe. Os delimitadores utilizados para tokenização são: o espaço em branco entre os termos, quebras de linhas, tabulações, e alguns caracteres especiais.

- *Stopwords*: depois da tokenização, cada termo obtido passa pela etapa de limpeza. São removidas as *stopwords*, com base em uma lista pré-existente de termos irrelevantes. Essa lista é composta por preposições, artigos, advérbios, números e pronomes.

- Filtragem por Classe Gramatical: remoção de termos que não pertenciam a classe dos substantivos ou adjetivos; de acordo com (BARRERA; VERMA, 2012) é uma boa pratica quando o propósito é sumarizar/caracterizar.

- *Stemming*: é o método para redução de uma palavra ao seu radical, removendo as desinências, afixos, e vogais temáticas. Com sua utilização, as palavras derivadas de um mesmo radical serão contabilizados como um único termo.

- *n-Gramas*: uma sequência de palavras pode começar de uma maneira conhecida, mas terminar por uma palavra desconhecida. Um modo de agrupar todas as sequências de tamanho n que começam pelas mesmas $n-1$ palavras em uma classe de equivalência é supor que o contexto local prévio afeta a palavra seguinte, e construir o modelo

de Markov de ordem $(n-1)$ ou modelo de n-Gramas (sendo a última palavra do n-grama a que está sendo prevista). Quanto maior o valor de n, isto é, maior o número de classes que dividem os dados, maior a confiabilidade da inferência. No entanto, o número de parâmetros a serem estimados cresce exponencialmente em relação a n Manning e Schütze (1999), Figueroa e Atkinson (2012). Por isso, geralmente são utilizados bigramas ou trigramas em aplicações de natureza semelhante. Aqui consideramos $n = \{1, 2, 3\}$, ou seja, combinações de até três termos formando um único atributo.

Como saída da fase de pré-processamento, os textos são armazenados na tabela 'author-document' junto com o identificador dos seus respectivos autores. Não foi empregado peso de relevância para uma palavra ou sentença, por exemplo, título. Ao final da etapa de representação dos usuários, temos para cada nó da rede seus atributos expressivos. De posse dos dados de entrada, atributos dos usuários e suas comunidades estão aptos a aplicar as abordagens de caracterização de grupos, geradoras dos perfis descritores. Como não há nenhuma orientação (perfis das comunidades rotulados) ou avaliação padrão para apontar a qualidade dos perfis gerados, recorreu-se a uma avaliação qualitativa realizada por especialistas, descrita a seguir.

4.1.5 Avaliação Qualitativa

O presente trabalho tem como metodologia a pesquisa qualitativa, já que esta responde questões muito particulares (descrição dos perfis apresentados). Segundo Minayo e Sanches (1993), a pesquisa qualitativa se preocupa com um nível de realidade que não pode ser quantificado. Ou seja, ela trabalha com o universo de significados, motivos, aspirações, crenças, valores e atitudes, o que corresponde a um espaço mais profundo das relações, dos processos e dos fenômenos que não podem ser reduzidos à operacionalização de variáveis. Idem o problema posto, não temos um parâmetro de comparação para os perfis gerados, temos por consequência a busca por análises qualitativas.

Dessa forma, para definir uma avaliação imparcial sobre a qualidade dos perfis gerados pelos métodos de caracterização de comunidades, todos os métodos concorrentes foram avaliados nas mesmas condições, ou seja, a mesma rede, comunidades, máquina e representação dos usuários (pré-processamentos). Para cada avaliador foi apresentado um formulário eletrônico contendo:

1. Os títulos dos dez artigos mais coesos[3] do grupo (com um link para o resumo do artigo na página web da arXiv). Essa seleção foi necessária devido a quantidade de artigos presentes em cada grupo, tornando inviável para os avaliadores a consideração de todos no processo de avaliação. De acordo com o estudo piloto de (TANG; WANG;

[3] Um artigo é considerado coeso se apresenta grande semelhança com o conteúdo presente no grupo.

LIU, 2011), os avaliadores tendem a atribuir classificações aleatórias se a tarefa é muito demorada.

2. Uma tabela com os perfis gerados para as comunidades, isto é, os dez termos mais representativos, selecionados por cada um dos métodos (uma coluna para cada método). Ressaltamos que para cada grupo apresentado ao avaliador, as colunas são geradas aleatoriamente, logo as posições dos métodos na tabela não são fixas, evitando assim um viés associado as posições.

3. Pergunta de avaliação: "Com base nesses artigos, qual método produziu o melhor perfil para a comunidade"?

4. Finalmente, um espaço para a seleção do melhor método, melhor perfil dentre os quatro apresentados.

Em cada página de avaliação, os quatro métodos foram indicados como "Método I","Método II", "Método III" e "Método IV", evitando assim a identificação dos métodos.Conforme ressaltado, para evitar o viés associado com a posição da coluna, a ordem de apresentação dos perfis da comunidade foi gerada de forma aleatória para cada página de avaliação. Suponhamos que, para uma comunidade as quatro colunas são preenchidas por TF-IDF, χ^2, WRS e BNS, respectivamente; A próxima vez que esse grupo ou outro grupo for escolhido, as três colunas podem corresponder a métodos em uma ordem totalmente diferente. Esse estudo foi dividido em três experimentos que serão detalhados nas próximas seções.

4.2 EXPERIMENTO I - CARACTERIZAÇÃO DE GRUPOS EM REDES DE COAUTORIA: UM ESTUDO COMPARATIVO

Como já apresentado, uma análise descritiva das comunidades pode ser feita basicamente de duas formas distintas Tang et al. (2008): CGBA e CGBD (detalhes no Capítulo 2). Na primeira abordagem, o objetivo é encontrar atributos que são mais prováveis de ocorrer dentro do grupo, ignorando o resto da rede; na CGBD selecionam-se características que diferenciam um grupo dos outros na rede, considerando-se todos os usuários.

Estudos anteriores demonstraram que a CGBA é mais aplicável em ambientes relativamente livres de ruído Tang, Wang e Liu (2011). Para atributos ruidosos, como postagens em mídias sociais ou interesses auto-relatados, as técnicas CGBD superam consistentemente a abordagem CGBA, porém com um custo computacional maior. Isso se deve a quantidade de usuários e informações consideradas para caracterizar as comunidades.

Embora tenha apresentado bons resultados, (TANG; WANG; LIU, 2011) consideraram apenas um método de cada abordagem: BNS Forman (2003) (CGBD e CGBDE) e TF Treeratpituk e Callan (2006) (CGA). Em (GOMES et al., 2013), um novo método de CGBD

foi proposto, o teste WRS foi adaptado para caracterização de grupos fazendo uso de atributos numéricos (idade, quantidade de acessos, número de postagens, entre outros). Apesar da abordagem CGBD tenha sido efetiva nas descrições, poucos métodos baseados em diferenciação foram explorados; por exemplo, trabalhos anteriores Tang, Wang e Liu (2011), Gomes et al. (2013), consideraram apenas um único método de CGBD, analisando apenas BNS e WRS, respectivamente. O presente experimento tem por objetivo, explorar de forma concreta outras possibilidades, dando um destaque à possíveis adaptações facilmente implementadas ao problema.

Com relação aos objetivos, a caracterização de grupos sociais é bastante semelhante a uma classe de algoritmos mais tradicional, conhecida como rotulagem de agrupamentos (*Cluster Labeling*). As principais diferenças entre elas são as fontes de informações e o processo de identificação de grupos, aprendizagem tradicional não supervisionada, por exemplo, clustering; e detecção de comunidade em redes. Inerentemente as correlações entre as áreas, temos a facilidade de adaptação de métodos de clustering para caracterização de comunidades sociais. Adicionalmente teríamos a inclusão das informações relacionais entre os nós, mas a sua aplicação também dar-se-ia pela identificações de termos/atributos descritores para as classes/comunidades.

Sabendo-se da facilidade de adaptação, é relevante analisar o desempenho de métodos consolidados de rotulagem de agrupamentos no contexto de comunidades sociais. Ressaltando, ainda, que apesar de alguns estudos iniciais sobre a análise de comunidades em redes de coautoria Savić et al. (2015), não há nenhum estudo prévio direcionado para caracterização de comunidades acadêmicas.

Para abordar as limitações de estudos anteriores Tang, Wang e Liu (2011), Gomes et al. (2013), este experimento apresenta uma estudo comparativo de métodos baseados em diferenciação para a caracterização de comunidades em uma rede de coautoria Gomes, Prudêncio e Nascimento (2016). Nesse cenário, os nós da rede representam os autores, dotados de informações colaterais, codificadas como atributos textuais extraídos dos títulos e resumos dos artigos publicados. Sobre as informações utilizadas para diferenciar os grupos, duas classes distintas de métodos foram consideradas no presente estudo: métodos já aplicados ao contexto de redes e em tradicionais métodos de rotulagem em agrupamentos. A primeira classe de métodos incorpora a estrutura de rede no processo de criação de perfil de grupo. Neste trabalho, foram considerados dois métodos distintos baseados em rede: o BNS Tang, Wang e Liu (2011) e foi proposto uma variação para atributos textuais do WRS, definido em (GOMES et al., 2013). Os métodos convencionais de rotulagem em agrupamentos incluem métodos que tradicionalmente não consideram nenhuma informação da rede na tarefa de rotulação. Para essa classe, também se considerou dois métodos distintos: o TF-IDF Treeratpituk e Callan (2006), e a selecção de qui-quadrado (Chi-Squared Selection (χ^2)) Witten e Frank (2005), ambos foram adaptados para a tarefa de rotulagem em comunidades sociais (todos os métodos foram detalhados no Capítulo 2).

4.2.1 Estratégia de Caracterização de Comunidades

Dada uma rede dividida em comunidades, onde cada nó é representado por um conjunto de atributos, a caracterização de comunidades concentra-se na seleção automática das características mais descritivas para cada comunidade. Neste experimento, definimos uma estratégia dividida em três fases. Inicialmente, precisamos identificar as comunidades. Algumas redes apresentam comunidades explícitas (através de aplicações para formação de grupos), mas caso as comunidades estejam implícitas na rede em estudo, se faz necessário a identificação via algoritmos de detecção. Dessa forma, essa etapa trata justamente da identificação das comunidades, ou via assinaturas dos usuários ou através da aplicação de algoritmos de detecção de comunidades, (1) fase de **Detecção de Comunidades** (detalhes na Subseção 4.1.3). Na sequência, é realizado o pré-processamento sobre a base de dados, identificando os atributos individuais dos usuários. Após o pré-processamento da base, os principais atributos dos usuários são selecionados: (2) fase de **Representação dos Usuários** (detalhes na Subseção 4.1.4). Finalmente, é iniciada a (3) fase de **Caracterização dos Grupos**, com a aplicação de um método de caracterização sobre a rede pré-processada na fase anterior. Na Figura 13, a arquitetura proposta é apresentada Gomes, Prudêncio e Nascimento (2016). Nas subseções a seguir, serão apresentados os métodos utilizados, os resultados obtidos e as limitações do Experimento I.

4.2.2 Métodos CPBD considerados no Experimento I

Como já mencionado, os objetivos dos métodos de caracterização de comunidades são obter descrições concisas e significativas para as comunidades. Neste experimento, visamos confrontar os métodos CGBD já propostos, bem como verificar a eficácia da adaptação de métodos de rotulagem de agrupamentos em documentos ao problema de caracterização de comunidades.

Dessa forma, foram considerados quatro métodos nesses experimentos: dois métodos baseados em redes, o BNS Forman (2003) (detalhes na Subseção 2.3.2.1) e propomos uma variação do WRS definido em (GOMES et al., 2013) (detalhada subsequentemente); e dois métodos tradicionais de rotulagem em agrupamentos, TF-IDF Treeratpituk e Callan (2006) (detalhes na Subseção 2.3.2.3), e o *Chi-Squared Selection* (χ^2) Witten e Frank (2005) (detalhes na Subseção 2.3.2.2).

Conforme apontado em Gomes et al. (2013), a adaptação do teste estatístico não-paramétrico *Wilcoxon Rank Sum* para a caracterização de grupos sociais, proporcionou bons resultados com atributos numéricos. Todavia, para atributos textuais se faz necessário repensar esse método; termos poucos presentes em um grupo podem apresentar grandes diferenças em relação ao restante da rede e serem selecionados como caracterizadores da comunidade analisada. Sabendo que o propósito é caracterizar o grupo e não a rede, aqui propomos a adição de uma restrição inicial à seleção de atributos a serem analisados

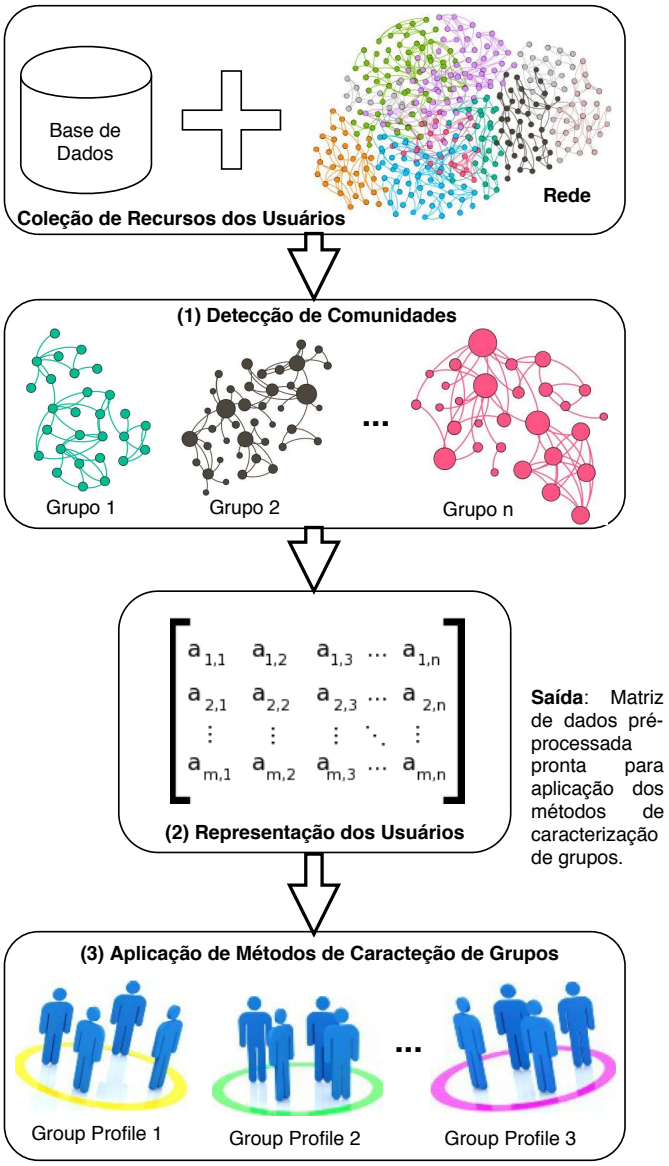

Figura 13 – Uma visão geral da estratégia de caracterização de grupos - Experimento I.

pelo teste WRS. Assim, selecionam-se apenas atributos que apresentarem média no grupo superior a amostra de comparação (restante da rede). Dessa forma, temos a seguinte restrição:

$$m_{P_A} > m_{G'_A} \tag{4.1}$$

onde m_{P_A} e $m_{G'_A}$ são as médias do atributo A no grupo (P) e no restante da rede (G'), respectivamente. Sem essa restrição poderíamos selecionar um atributo (termo) pouco presente na comunidade, o que não seria relevante para sua caracterização. Demais detalhes desse método na Seção 2.3.2.1.

4.2.3 Resultados Experimento I

Um total de 34 pessoas vinculadas à computação (graduação, pós-graduandos, professores universitários) participaram da pesquisa, o que resultou em 340 avaliações. A porcentagem de respostas em que cada método foi marcado como "melhor perfil" é apresentado na Tabela 6.

Tabela 6 – Porcentagem de respostas que cada método foi marcado como "melhores rótulos" (média calculada sobre todas as comunidades).

WRS	χ^2	**BNS**	**TF-IDF**
23.82%	27.65%	17.66%	**30.9**%

Embora o método de TF-IDF seja tradicionalmente um método simples, esse alcançou o maior desempenho geral no experimento. Em contraste, o método BNS obteve o menor desempenho global. Os métodos WRS e Chi-Square (χ^2) foram bastante similares em média. O WRS foi apontado como o melhor perfil para os grupos 134 e 151, enquanto o χ^2 foi o melhor nos grupos de 104, 153 e 156 (Detalhes Figura 21). Para o grupo 145 os avaliadores não apresentaram uma unanimidade sobre o melhor perfil, apresentando porcentagem de seleção equivalente para os métodos WRS, BNS e TF-IDF. Analisando as estatísticas na Tabela 5, verificamos que esse é precisamente o grupo de maior densidade, o que demonstra a importância dessa medida no desempenho de todos os métodos de caracterização. Fatos esperado, pois essa medida representa o grau de relacionamentos do grupo de autores.

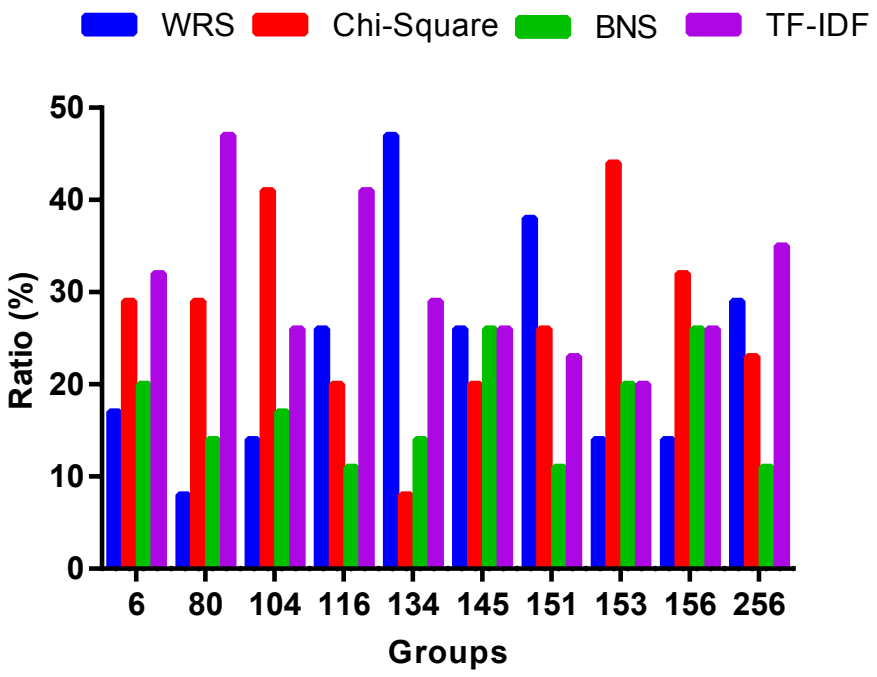

Figura 14 – Gráfico de colunas com os resultados para cada grupo.

Para avaliar se houve diferenças estatísticas entre os métodos, foram aplicados testes de hipóteses[4]. Para nosso experimento foi aplicado o teste não-paramétrico de Friedman Friedman et al. (1999), embora seja um teste relativamente conservador, permite a comparação de múltiplos métodos, verificando se existe diferença estatisticamente significante entre eles. Dessa forma, considerou-se com um intervalo de confiança de 95% ($\alpha = 0.05$), que resultou na não rejeição de H_0, obtendo um $F = 1.2$ e um $p - value = 0.753$. Uma vez que não apresentaram diferenças significativas entre os métodos, o pós teste de Nemenyi não foi aplicado. Na Figura 15, os valores obtidos pelo teste são apresentados na forma de diagrama de diferenças críticas (*critical difference diagram*) (DEMŠAR, 2006).

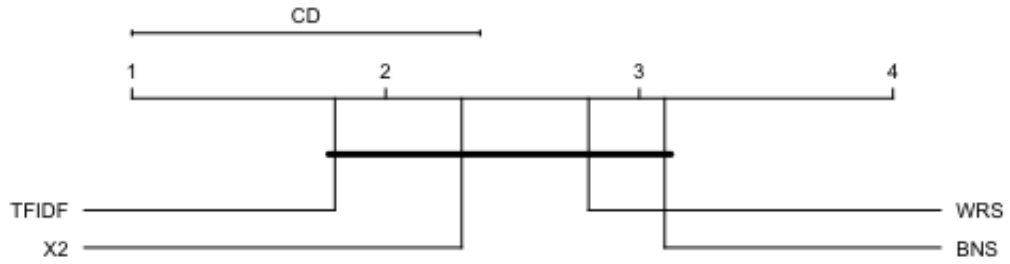

Figura 15 – Apresentação do diagrama de diferenças críticas, comparando o desempenho dos métodos de CGBD sobre as comunidades analisadas. Os métodos conectados não apresentaram diferenças significativas.

Como podemos ver no diagrama de diferenças críticas, os métodos analisados não apresentaram diferenças estatisticamente significativas de desempenho entre eles. Isso mostra que para esse experimento nenhum método foi soberano sobre outro. Isso atesta que a variação aqui proposta se comportou de forma equivalente ao estado da arte, o que aprova a viabilidade do teste WRS para caracterização de grupos em redes complexas, fazendo uso de atributos textuais. Bem como, a adaptação de métodos tradicionais de rotulagem de agrupamentos em documentos ao problema de caracterização. Esse apontamento reforça a ideia que o principal gargalo da caracterização das comunidades socais está na escalabilidade, uma vez que temos consideráveis possibilidades de métodos tradicionais para rotulagem. Essa questão é analisada no Experimento II (Seção 4.3).

A fim de verificar se os perfis indicados por cada método são similares ou não, os perfis gerados de cada método também foram comparados. Para tanto, o coeficiente de similaridade de *Jaccard* Cha (2007) foi aplicado, esse retorna uma medida de similaridade entre conjuntos de amostras finitas. Sendo definido como o tamanho da intersecção dividida pelo tamanho da união dos conjuntos de amostras. A seguir, na Tabela 7, a análise de similaridade é delineada.

Analisando a Tabela 7, verificamos que todos os métodos retornaram perfis pouco semelhantes em média, com exceção do TF-IDF e *Chi-Square*, similaridade média de 0.23.

[4] Os testes de hipóteses são guias em relação à confiança com que se assume que existe realmente uma diferença de desempenho. Na hipótese nula (H_0) não há diferença de desempenho, enquanto que na hipótese alternativa (H_1) esta diferença existe

Tabela 7 – Coeficiente de Jaccard sobre geração de perfis.

	WRS	CHI-SQUARE	BNS
CHI-SQUARE	0.0052	-	-
BNS	0.0000	0.0491	-
TF-IDF	0.0458	0.2300	0.0105

Dentre os grupos que os dois métodos geraram perfis mais similares, destacam-se: grupos 80 e 153 ambos com similaridade de 0.428. Esse resultado mostra que cada método gerar perfis diferentes, e se faz necessário estudos mais aprofundados para verificar quais as peculiaridades de cada um, e as possíveis relações com as medidas da rede (grau, densidade, entre outros) e a coesão.

Para realizar uma análise do desempenho dos diferentes métodos, vamos mostrar dois exemplos concretos: os grupos 80 e 134. Esses foram os grupos que apresentaram o maior número de avaliadores com opiniões semelhantes sobre os perfis (ambos 47%). As Tabelas 8 e 9, apresenta os perfis extraídos para descrever esses grupos com base nos títulos e resumos dos artigos publicados pelos autores. Os rótulos são classificados por ordem alfabética, visando facilitar a comparação entre os perfis gerados por cada método.

Tabela 8 – Perfis gerados por cada método para a comunidade 80.

Comunidade 80			
WRS	Chi-Square	BNS	TF-IDF
convention	bayesian	continuo	algorithms
effort	bayesian networks	control	approximate
engineering	belief	empirical	bayesian
evidence	inference	expert	bayesian network
fault	method	method	inference
inference bayesian	networks	powerful	method
knowledge engineering	papers	simple	models
positive	probabilistic	system	networks
sensitivity analysis	probability	time	probabilistic
SPI	utility	utility	probability

O grupo 80 possui 28 autores, sendo o grupo com maior coesão dentre todos os analisados (detalhes em Tabela 5). Nesse grupo podemos observar que as características selecionadas pelos mótodos WRS e BNS (Table 8) são, em grande maioria, termos que não agregam ao entendimento do conteúdo do grupo, são exemplos: ‘evidence’, ‘fault’, ‘positive”, ‘convention’, ‘continuo’, ‘control’, ‘time’, ‘utility’, etc. Todavia, observando os termos selecionados pelo método *Chi-Square*, temos: ‘Bayesian’, ‘Bayesian networks’, ‘probabilistic’; esses já dão uma visão para estudos em aprendizagem em redes bayesianas. Que acaba por ser mais detalhado no perfil gerado pelo método TF-IDF, com adição de

outros termos, como: 'models' e 'probability'. Analisando, 6 dos 10 termos selecionados se repetem nos métodos TF-IDF e *Chi-Square*, o que reafirma o direcionamento apontado pelos perfis. Dessa forma, podemos concluir que a comunidade 80 é um grupo de aprendizagem de máquina baseado em modelos probabilísticos, focado no estudo das redes bayesianas e em modelos estocásticos e dinâmicos de aprendizado.

Tabela 9 – Perfis gerados por cada método para a comunidade 134.

Comunidade 134			
WRS	Chi-Square	BNS	TF-IDF
DEC-POMDPs	algorithms	algorithms	basis
heuristic	approach	approach	circuit
MRE	decision	decision	concept
multiagent	functions	heuristics	hierarchy
observable markov decision	optimality	method	kernel
online	papers	model	linear programming
optimal policy	representations	networking	practice
policy	sampling	optimal	reward
search	search	planning	suitable
speedups	value	policy	value function

Já o grupo 134 é o segundo maior grupo do experimento, com 60 usuários. Trata-se do grupo com menor densidade (5.4%), como já visto uma medida de grande relevância nos estudos das comunidades em redes sociais. O método WRS conseguiu apontar bons termos caracterizadores, tais como: *'DEC-POMDPs*[5](decentralized partially observable Markov decision process), 'MER'[6] (Most Relevant Explanation), 'observable Markov decision', 'Optimal policy'[7], 'multiagent' e 'heuristic'. Igualmente ocorrido na comunidade 80, para a comunidade 134 os métodos Chi-Square e BNS, geraram perfis bastante superficiais, abstraindo o conteúdo central da comunidade. Não muito diferente, o método TF-IDF apresentou alguns termos poucos descritores. Com base no perfil gerado pelo método WRS, podemos interpretar o grupo 134 como um grupo focando em estudos sobre agentes baseados em conhecimento e sistemas multiagentes, considerando a aplicações de técnicas como: MER, Modelos de Markov e uso de funções heurísticas. Essa pequena análise incentiva novos estudos sobre a influência da densidade no desempenho dos métodos caracterização de grupos.

Este experimento teve por objetivo avaliar a eficácia de métodos de caracterização de grupos aplicados à uma rede de coautoria. Até então não se tinha indicativos de estudos de descrição de comunidades nesse contexto. Esse resultado se fez relevantes, tanto

[5] É um modelo muito geral para a coordenação entre os vários agentes.
[6] Um método para identificar automaticamente as variáveis-alvo mais relevantes.
[7] Um algoritmo para sistemas multiagente.

para seleção dos melhores métodos com para um maior embasamento para aplicação da abordagem baseada em informações relacionais, validada no Experimento II. A seguir a Seção 4.3 estende esse experimento considerando a rede de coautoria arXiv aqui adotada, bem como os quatro métodos comparados. No Experimento II, analisamos a aplicação de uma abordagem baseada em informações relacionais para caracterização de comunidades.

4.3 EXPERIMENTO II - CARACTERIZAÇÃO DE GRUPOS BASEADO EM INFORMAÇÕES RELACIONAIS

Neste experimento, a consideração das informações relacionais no processo de caracterização de grupos é avaliada. Para tal, os métodos considerados na Seção 4.2, ou seja, BNS, TF-IDF, Qui-Quadrado e WRS, foram aplicados na abordagem baseada em informações relacionais (apresentada no Capítulo 3). É importante ressaltar que a abordagem proposta é genérica, também aplicável a métodos baseados em agregação. As medidas de centralidade consideradas foram: Grau (*Degree*), Proximidade (*Closeness*), Intermediação (*Betweenness*), Autovetor (*Eigenvector*), *PageRank* e Excentricidade (*Eccentricity*).

Com base no tamanho das comunidades da rede (consulte Tabela 5), aplicou-se o filtro baseado em informações relacionais (detalhes na Seção 3.2.1), selecionando para cada comunidade da rede seus dez nós com maiores pontuações de acordo com a centralidade. Considerou-se no filtro dez autores após análises quantitativas relacionadas a generalização do conteúdo das comunidades da rede em estudo, ou seja, a cobertura da quantidade de atributos considerados. Dessa forma, a partir da rede de coautoria original, diferentes subredes são produzidas para cada medida de centralidade adotada, como mostra a Figura 16. As estatísticas gerais de cada subrede gerada são apresentadas na Tabela 10.

Tabela 10 – Medidas das subredes de centralidade.

	Grau	Proximidade	Intermediação	Autovetor	*PageRank*	Excentricidade
#Autores	100	100	100	100	100	100
#Links	139	75	141	209	157	83
Densidade	2.8%	1.5%	2.8%	4.2%	3.2%	0.017
Grau Médio	2.78	1.5	2.82	4.18	3.14	1.66
Diâmetro	8	2	19	14	16	4

De acordo com a Tabela 10, para todas as subredes resultantes obtevesse aumento da densidade em relação a rede original (ver Tabela 4), sendo a subrede autovetor a mais proeminente, refletindo o destaque no valor do Grau Médio. Esses ganhos em densidade são relevantes para o propósito da abordagem proposta, que visa selecionar nós bem conectados que generalizem o conteúdo das comunidades. Outras medidas de centralidade que geraram subredes com consideráveis ganhos em densidade foram: grau, intermediação e *pagerank*.

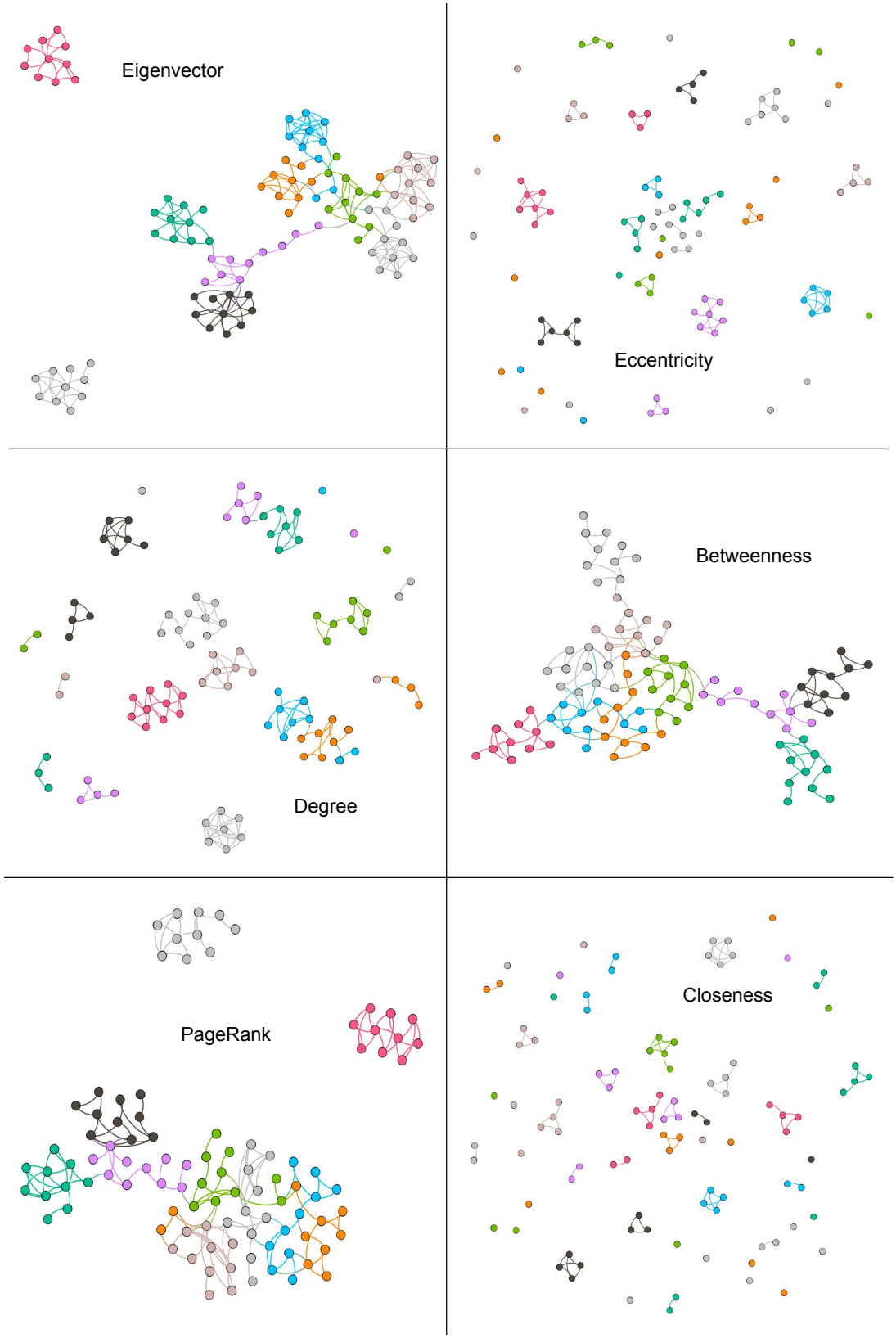

Figura 16 – Subgrafos resultantes de cada medida de centralidade após a aplicação do filtro de centralidade.

Analisando a Figura 16, nas subredes obtidas pelas medidas proximidade e excentricidade, a maioria das comunidades se dissolveram, perdendo suas conexões internas,

resultando em um grafo com nós e algumas comunidades isoladas. A centralidade de grau apresentou um grafo com predominância de comunidades isoladas densamente conectadas, destacando a ocorrência de fusões de algumas comunidades, 6+156 e 104+134. Já o grafo resultante da intermediação, destaca-se pela considerável similaridade visual com a rede original, ressaltando a boa conectividade, não apresentando nenhum nó isolado. Semelhante ao grafo de intermediação, o *pagerank* obteve uma rede densamente conectada, destacando a presença duas comunidades isoladas, 256 e 145. Na Tabela 11, para cada subrede de centralidade, são apresentadas as densidades de suas comunidades.

Tabela 11 – Densidades dos grupos em cada subrede de centralidade resultante.

Grupo	Grau	Proximidade	Intermediação	Autovetor	*PageRank*	Excentricidade
6	20.0%	15.6%	22.2%	28.9%	22.2%	28.9%
80	22.2%	15.6%	26.7%	33.3%	24.4%	11.1%
104	28.9%	20.0%	26.7%	53.3%	24.4%	28.9%
116	33.3%	22.2%	31.1%	44.4%	31.1%	24.4%
134	20.0%	8.9%	22.2%	31.1%	22.2%	6.7%
145	40.0%	20.0%	31.1%	37.8%	40.0%	28.9%
151	26.7%	13.3%	28.9%	44.4%	42.2%	17.8%
153	31.1%	15.6%	26.7%	57.8%	33.3%	15.6%
156	48.9%	31.1%	26.7%	55.6%	35.6%	17.8%
256	28.9%	4.4%	22.2%	40.0%	28.9%	4.4%

Analisando a Tabela 11 é possível verificar ganhos de densidade para as comunidades em comparação com a rede primária (consulte Tabela 5). Isso demonstra a efetividade da abordagem baseada em informações relacionais na obtenção de subredes densamente conectadas. Ressaltando que a abordagem tem como propósito o uso das informações relacionais para a seleção dos usuários proeminentes de cada comunidades, possibilitando-se com isso a generalização dos conteúdos e reflexões das peculiaridades de suas comunidades. A seguir na Tabela 12, para uma melhor visualização das generalizações de conteúdos alcançadas pela abordagem baseada em informações relacionais, são apresentados para cada subrede e rede global, as quantidades de artigos presentes em cada comunidade.

Combinando-se as informações da Tabela 12 com os dados apontados pela Figura 16 e Tabela 10, confirma-se a hipótese inicial: a filtragem/redução de uma rede original à uma subrede densamente conectada, pode generalizar o conteúdo das comunidades analisadas. Reafirmando, ainda, a proeminência das medidas *pagerank* e intermediação, na seleção de nós que generalizam o conteúdo das comunidades. Para uma ampla exploração da capacidade de generalização de conteúdos da abordagem baseada em informações relacionais, na Figura 17 são confrontadas as proporções (em percentagem (%)) relacionando a quantidade de usuários considerados e os conteúdos coletados para cada subrede de centralidade.

Tabela 12 – Quantidade de artigos considerados em cada comunidade na rede orginal e subredes resultantes de cada centralidade considerada.

Grupo	**Global**	Grau	Proximidade	Intermediação	Autovetor	*PageRank*	Excentricidade
6	**88**	37	6	51	47	51	6
80	**128**	61	17	90	82	90	15
104	**226**	94	23	154	119	155	22
116	**260**	107	30	184	138	186	38
134	**327**	140	40	229	175	235	55
145	**344**	158	49	245	190	253	70
151	**373**	181	68	273	208	271	92
153	**467**	208	84	351	252	349	109
156	**522**	228	97	403	287	392	140
256	**562**	259	112	434	317	423	155
Total Artigos	**3297**	1473	526	2414	1815	2405	702

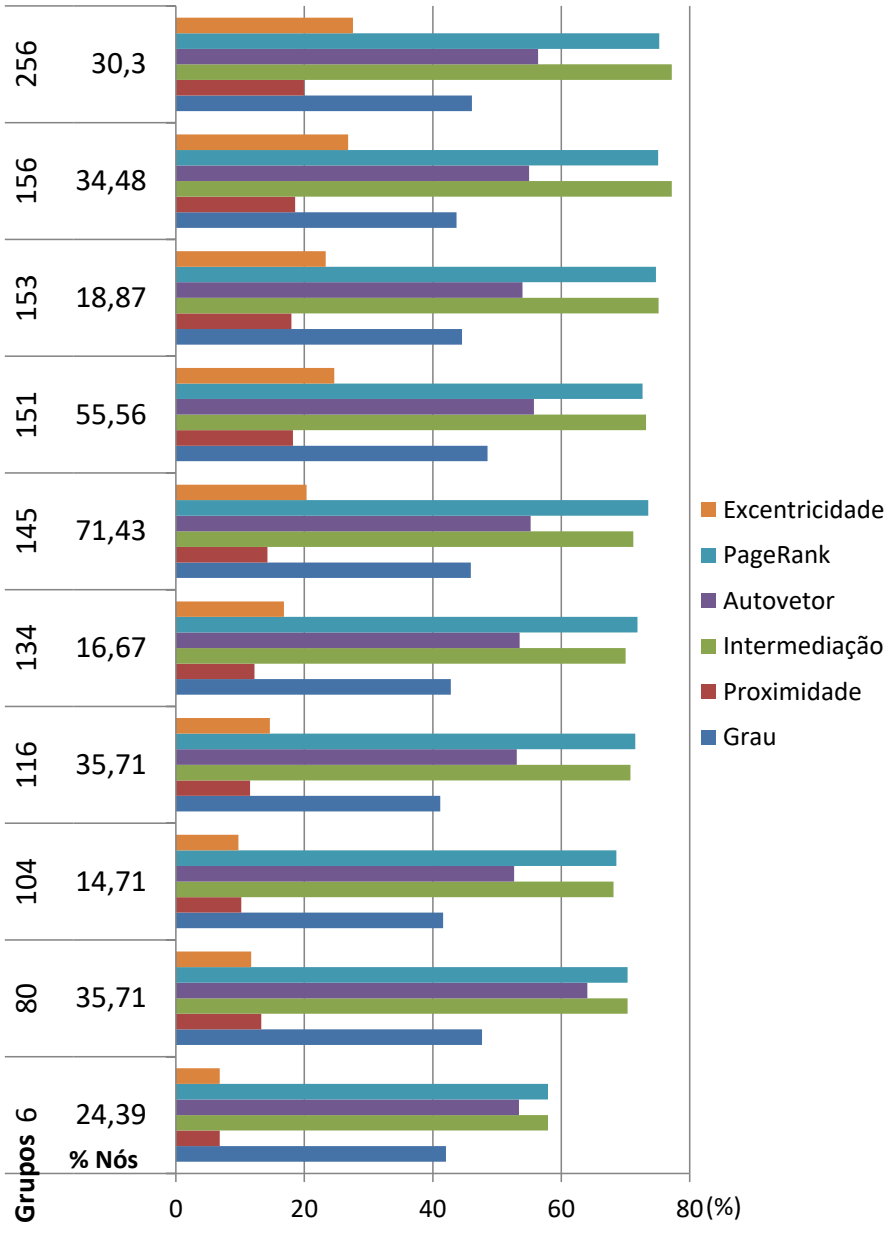

Figura 17 – Apresentação das proporções em gráfico de barras, contendo a percentagem dos usuários considerados em cada comunidade pela abordagem baseada em

Observando os dados do gráfico de barras, Figura 17, verifica-se a considerável redução dos nós considerados na caracterização das comunidades aplicando o filtro de centralidade, apenas 33,78% dos usuários da rede original são apreciados, com perdas suaves em conteúdos para as subredes densamente conectadas. Por exemplo, as subredes intermediação e *pagerank*, foram efetivas, generalizando o conteúdo da rede em 71,11% e 71,15%, respectivamente. As medidas autovetor e grau apresentaram dados medianos em relação à conteúdos, 55,3% e 44,4%, respectivamente. Todavia, as subredes proximidade e excentricidade não se mostraram efetivas, com 14,3% e 18,24% do conteúdo geral, respectivamente; concomitantemente, essas subredes apresentaram dissolução e considerável perda de conectividade entre os nós das comunidades. Isso novamente reafirma a hipótese inicial, podemos utilizar as informações relacionais para geração de subredes densamente conectadas, capazes de generalizar e selecionar os conteúdos pertinentes das comunidades; resultando em uma considerável redução dos dados considerados na caracterização.

Conforme supracitado, além da capacidade de generalização dos conteúdos das comunidades, objetiva-se que os nós selecionados pelo filtro de informações relacionais possuam as informações relevantes das comunidades, possibilitando uma caracterização efetiva. Com o objetivo de verificar a efetividade da abordagem baseada em informações relacionais na caracterização das comunidades realizou-se três experimentos: (1) similaridade entre os perfis gerados por diferentes métodos - Subseção 4.3.1; (2) análise de similaridade com os melhores perfis gerados no Experimento I - Seção 4.3.2; e (3) uma avaliação qualitativa. Finalmente na Subseção 4.3.4, é apresentado um estudo de viabilidade da abordagem de caracterização de grupos baseada em informações relacionais, nesse analisamos sua escalabilidade, comparando o tempo de execução da abordagem baseada em centralidade em relação a abordagem global.

4.3.1 Análise da Similaridade entre Perfis de Grupos Considerando Diferentes Métodos CGBD

Para este experimento não realizamos uma avaliação qualitativa com avaliadores humanos. Para verificar a eficácia da abordagem baseada em informações relacionais, analisou-se a similaridade entre os perfis gerados por cada método, variando a abordagem de caracterização de grupos (global versus baseada em informações relacionais). Nesta averiguação também usou-se o índice de Jaccard (CHA, 2007) como medida de verificação de similaridade.

Neste experimento, analisou-se a similaridade dos perfis gerados por cada método, variando as abordagens e medidas de centralidade. Como este é o primeiro estudo nesse contexto, o objetivo é verificar o grau de semelhança entre os perfis gerados considerando-se todos os nós ou apenas os nós mais centrais, bem como o desempenho de cada medida de centralidade nos métodos CGBD adotados. A Tabela 13 apresenta a semelhança média entre os perfis gerados para os dez grupos, comparando cada método individualmente na abordagem global e baseada em informações relacionais. A média de cada medida de

centralidade nos quatro métodos também é apresentada.

Tabela 13 – Similaridade de Jaccard (%) entre os perfis gerados por cada método CGBD considerando todos os nós (global) e apenas os nós mais centrais (baseado em informações relacionais). As maiores similaridades de cada método são apresentadas em negrito.

Médida	Qui-quadrado	BNS	TF-IDF	WRS	Average
Grau	**41.00**	**4.97**	31.32	**38.15**	28.86
Proximidade	21.76	2.10	17.75	7.79	12.35
Intermediação	39.60	1.63	**45.12**	24.85	27.80
Autovetor	23.25	2.16	39.94	17.06	20.60
PageRank	35.05	1.64	44.07	31.87	28.16
Excentricidade	25.22	3.22	26.70	10.87	16.50

De acordo com a Tabela 13, a centralidade de grau apresentou a maior similaridade média entre os métodos analisados. Apresentando também os maiores valores de similaridade para os métodos Qui-quadrado (41%) e WRS (38,15%), bem como uma similaridade média razoável para o método TF-IDF (31,32%). As medidas de intermediação e *pageRank* também obtiveram boa similaridade média, especialmente para o método TF-IDF (45,12% e 44,07%, respectivamente). As centralidades de proximidade e excentricidade apresentaram perfis semelhantes em média, e a medida autovetor apresentou média geral razoável, mostrando boa similaridade no método TF-IDF (39,94%). Considerando-se o método BNS na geração dos perfis, todas as medidas de centralidade apresentaram baixa similaridade média quando comparada com a abordagem global. Para ilustrar os resultados obtidos, as Tabelas 14 e 15 apresentam os perfis gerados paca os grupos 145 e 151, respectivamente, fazendo uso da centralidade de grau para os métodos WRS e BNS. Idem experimento da Seção 4.2, os rótulos são classificados em ordem alfabética.

Para o grupo 145, conforme apontado na Figura 17, por trata-se de um comunidade pequena (apenas 14 usuários), a abordagem baseada em informações relacionais considerou 71,43% dos nós, alcançando apenas 45,6% do conteúdo dos usuários com a centralidade de grau. Apesar da fragilidade na generalização total do conteúdo, verificamos que a subrede de grau foi eficaz na seleção das características relevantes, obtendo como reflexos perfis muitos semelhantes, praticamente idênticos, conforme apontado na Tabela 14. Ou seja, os nós escolhidos foram trouxeram consigo as informações pertinentes e peculiares da comunidade 145. A seguir, na Tabela 15, apresenta uma situação oposta, os perfil gerado para o grupo 151 é totalmente dissimilar considerando apenas os nós centrais apontados pela centralidade de Grau.

Analisando os resultados obtidos pelo método BNS no Experimento I, menos descritivo, reflete aqui em uma seleção muito dissimilar. Todavia, um perfil diferente pode fornecer uma visão excêntrica da comunidade, o que propulsionaria conclusões mais refinadas. Sabendo-se

Tabela 14 – Perfis Gerados pelo Método WRS para o Grupo 145, Global e Baseado em Grau.

Global WRS	**Centrality WRS**
bayesian belief network	bayesian belief network
features	features
hierarchy	hierarchy
independent	independent
key	key
mdp	limit
net	mdp
network models	network models
path	path
random variables	random variables

Tabela 15 – Profiles Generated by the BNS Method for Group 151, Global and Degree-Based.

Global BNS	**Centrality BNS**
application	conditional
approach	context
goal	distribution
graph	heuristic
joint	learning
linear programming	powerful
main	previous
natural	reinforcement
resource	selection
random value	size

que na visão/abordagem global o método BNS apresentou perfis pouco descritivos, pode-se sugerir uma visão baseada em informações relacionais, filtrando as informações irrelevantes e proporcionando uma melhor descrição. Esta é uma questão que será discutida mais adiante na Seção 4.3.3. A seguir, Seção 4.3.2, analisa-se as similaridades entre os perfis gerados pela abordagem baseada em informações relacionais e os melhores perfis apontados pelos avaliadores no Experimento I (Seção 4.2).

4.3.2 Similaridade Entre Perfis Baseados em Centralidade e Melhores Globais

O objetivo deste experimento é avaliar quão similares aos "melhores perfis"[8] são as descrições obtidas aplicando-se a abordagem baseada em informações relacionais. Ressaltando

[8] Como "melhores perfis", refere-se as melhores descrições fornecidas pelos métodos, ou seja, primeiro lugar no ranking de avaliações do estudo qualitativo descrito na seção 4.2.

que para cada caracterização de comunidade foi considerado o mesmo método em ambas as abordagens; método que obteve o melhor perfil para a comunidade analisada no Experimento I. A Tabela 16 apresenta a semelhança média de Jaccard sobre as dez comunidades, ao comparar os perfis obtidos por cada combinação de método/medida com os melhores perfis.

Tabela 16 – Semelhança média entre os perfis gerados na análise de centralidade e os melhores perfis identificados pelos avaliadores. Os melhores resultados para cada método são apresentados em negrito.

	Methods				
Medida	**Chi-Square**	**BNS**	**TF-IDF**	**WRS**	**Média**
Grau	17.7	1.11	19.08	**19.19**	14.27
Proximidade	16.47	**3.27**	11.85	5.32	9.23
Intermediação	**26.8**	1.11	26.2	9.37	15.87
PageRank	18.38	1.05	**30.46**	12.2	15.52
Autovetor	11.42	0.53	27.23	7.32	11.62
Excentricidade	16.78	1.58	13.83	3.93	9.03
Global	39.6	1.64	51.58	32.3	31.28

Um aspecto interessante que pode ser observado são as medidas de centralidade de grau, intermediação e *PageRank* continuarem a alcançar maiores similaridades com os melhores perfis. Além disso, o método TF-IDF alcançou a maior similaridade média quando combinado com o *PageRank* (30.46%), seguido de perto pelo Qui-Quadrado/intermediação (26.8%) e WRS/grau (19.19%). Os resultados obtidos pelo algoritmo BNS alcançaram pontuações em gerais mais baixas, embora a combinação BNS/Closeness tenha aumentado à similaridade quando comparada com a abordagem global. Na Tabela 17, são apresentados para cada grupo o método que gerou o melhor perfil no experimento da Seção 4.2, bem com as medidas de centralidade que combinadas com o método, geraram o perfil mais semelhante e o mais dissimilar comparada ao melhor perfil.

A Tabela 17 reafirma o equilíbrio entre as medidas grau, intermediação e *pagerank*, sendo que a última a que apresentou o maior número de perfis similares aos melhores (quatro perfis de grupo). Observa-se também que a centralidade de grau apresentou as maiores semelhanças com os melhores perfis quando combinada com o método WRS (grupos 134, 145 e 151). O mesmo acontece com o *pagerank*, quando combinado com o TF-IDF, a medida apresentou bons resultados (grupos 6, 80, 256). Essas associações apresentam indícios de efetividade, se fazendo necessária uma avaliação mais profunda em outros contextos. Outra medida que apresentou perfis semelhantes ao melhor perfil foi a centralidade de intermediação, apresentando maiores semelhanças em três grupos quando combinada com os métodos Qui-quadrado e TF-IDF. Curiosamente, as três medidas supracitadas também geraram subredes densamente conectados (detalhes na Tabela 10 e

Tabela 17 – Para cada comunidade a combinação Método/Medida que geraram, respectivamente, o perfil mais semelhante e dissimilar comparado ao melhor perfil.

Grupo	Melhores Métodos	Medida Dissimilar	Medida Similar
6	TF-IDF	Proximidade	PageRank
80	TF-IDF	Grau	PageRank
104	Chi-Square	Proximidade	PageRank
116	TF-IDF	Proximidade	Intermediação
134	WRS	Excentricidade	Grau
145	WRS	Excentricidade	Grau
151	WRS	Proximidade	Grau
153	Chi-Square	Eigenvector	Intermediação
156	Chi-Square	PageRank	Intermediação
256	TF-IDF	Proximidade	PageRank

Figura 16), levando novamente indícios de que, para subredes bem conectados, pode-se obter perfis similares a abordagem global com redução dos dados.

Como mostrado na Tabela 15, alguns perfis gerados pelas duas abordagens foram bastante dissimilares. Analisando a Tabela 17, podemos verificar as medidas de proximidade e excentricidade foram as que apresentaram os perfis mais dissimilares ao melhores do Experimento I. Curiosamente, essas subredes apresentaram características opostas ao debatido anteriormente, estas medidas geraram subredes dispersas e de baixa densidade (ver Tabela 10 e Figura 16). Em contra partida, a aplicação da abordagem baseada em centralidade possibilita, através da variação das medidas de centralidade, a obtenção de diferentes visões/entendimentos para comunidades que tiveram perfis poucos descritivos por algum método na abordagem global. Especialmente quando as subredes geradas pela abordagem baseada em informações relacionais são não densas. Podendo, assim, possibilitar novos entendimentos e conclusões sobre as comunidades.

4.3.3 Caracterização de Grupos Baseada em Informações Relacionais: Uma Avaliação Qualitativa

De modo a avaliar qualitativamente os perfis gerados pela abordagem baseada em informações relacionais, neste experimento realizou-se uma avaliação qualitativa semelhante à descrita na Subsecção 4.1.5. A única modificação realizada foi na quantidade de opções apresentadas aos avaliadores, apresentando três métodos a serem escolhidos. Dessa forma, os perfis gerados pelas configurações apresentadas na Tabela 17 foram selecionados. Logo, para o método que gerou as melhores descrições para as comunidades no experimento da Subseção 4.2, foram selecionadas as medidas de centralidade que geraram perfis mais similares e dissimilares, ambas utilizando a abordagem baseada em informações relacionais (detalhes na Tabela 17).

Para obter um resultado comparável ao experimento anterior (Seção 4.2), solicitou-se a avaliação do presente experimento aos mesmos 34 avaliadores do anterior. Dos 34 avaliadores, 16 realizaram essa avaliação. Logo, 160 avaliações foram submetidas e, portanto, cada grupo foi avaliado 16 vezes. A porcentagem de respostas em que cada método foi marcado como "melhor perfil" é apresentado na Tabela 18.

Tabela 18 – Porcentagem de respostas média para todos os grupos em que cada método/-medida foi marcado como "melhor perfil".

Melhores Perfis	**Medida Dissimilar**	**Medida Similar**
38.125%	30%	31,875%

Analisando o Table 18, os métodos baseados em diferenciação aplicados à abordagem global alcançaram o maior desempenho geral no survey. No entanto, as duas configurações de métodos/medidas analisadas na abordagem baseada em informações relacionais, apesar de termos perfis semelhantes ou dissimilares, apresentaram resultados competitivamente equivalentes. A abordagem global foi identificada como a melhor para os grupos 134 e 256. Para os grupos 6, 80 e 104, os melhores perfis foram obtidos com a combinação da centralidade *pagerank* combinada aos métodos TF-IDF e Chi-quadarado (mais similares com o perfil global). Para os grupos 145 e 156, por sua vez, os melhores perfis deste experimento foram os gerados pela combinação com uma medida de centralidade mais dissimilar aos melhores perfis globais. Esse é um resultado interessante, verificando que a abordagem baseada em informações relacionais pode apresentar para as comunidades entendimentos diferente da visão global. Em outras palavras, uma visão diferente e mais relevante da comunidade foi produzida, descrevendo de forma mais efetiva as comunidades. Para as comunidades 151, 153 e 116 não se obteve uma escolha evidente sobre o melhor perfil, detalhes na Figura 18.

Analisando as combinações método/medida adotadas pela abordagem baseada em informações relacionais, com os resultados obtidos no estudo qualitativo, há uma ênfase considerável na centralidade *PageRank*. Quando considerada essa medida gerou os melhores perfis do experimento para os grupos 6, 80, 104 e 156, com maior destaque para o último, com 68,75% dos votos. Ressaltando que para o grupo 156 o pagerank combinado ao Qui-quadrado foi a medida de perfil mais dissimilar a melhor descrição global; nota-se, ainda, que a comunidade 156 apresentou perfis avaliados de forma equilibrada no Experimento I (verificar Figura 21), não atestando uma uninimidade sobre o perfil mais adequado na visão global. Novamente o método WRS adotado na abordagem global gerou um perfil proeminente nas escolhas dos avaliadores (Figura 21), obtendo 68.75% dos votos no presente experimento.

Idem experimento anterior, para realizar uma análise do resultado para diferentes métodos/medidas, aqui se apresenta dois exemplos concretos: grupos 6 e 156. Nesse experimento, os dois grupos foram identificados como melhores descrições os perfis gerados

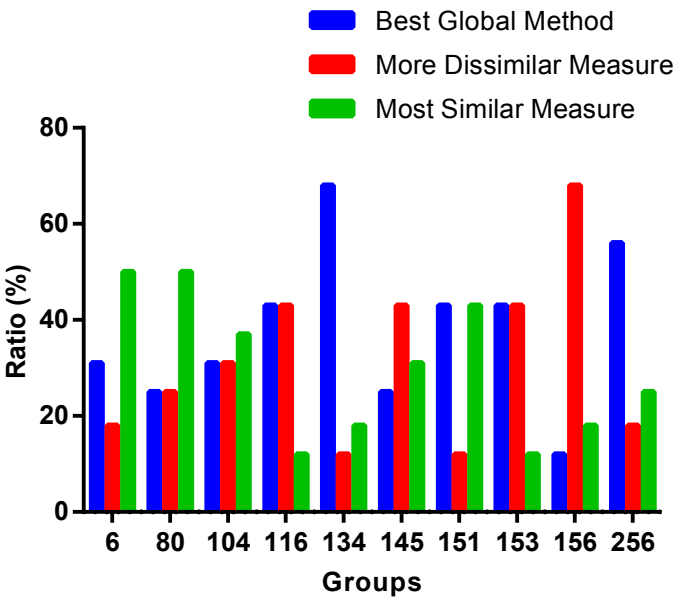

Figura 18 – Porcentagens que indicam quantas vezes cada método avaliado forneceu o melhor perfil no *survey* qualitativo.

pela abordagem baseada em informações relacionais, ambos fazendo uso da medida de centralidade PageRank. As Tabelas 19 e 20 apresentam os perfis extraídos para descrever essas comunidades.

Tabela 19 – Perfis gerados para o grupo 6

TF-IDF/PageRank	TF-IDF/Closeness	TF-IDF
algorithm	arc	algorithms
approach	arc consistency	approaches
bayesian	argumentation	bayesian
complexity	bounds	complexity
computational	causal	computations
diagrams	consistency	consistent
model	cost	model
network	links	paper
paper	localization	probability
probabilities	wcsp	processes

Analisando Tabela 19, existe uma grande similaridade entre os perfis gerados pela abordagem global (TF-IDF) e a baseada em informações relacionais fazendo uso do PageRank (TF-IDF/PageRank). Os termos gerados como perfis são bastante gerais, 'Bayesian', 'probabilities', 'model' e 'network', esses geram indicações de que se trata de uma comunidade estocástica de aprendizado de máquina com foco em redes bayesianas. No entanto, o perfil gerado pelo TF-IDF/Closeness, dissimilar aos outros dois, apresentou

alguns termos aparentemente mais específicos da atuação dos autores do grupo, tais como: nome de métodos e problemas investigados, tais como: '*arc consistency*' - um método de inferência fortemente investigado; e '*WCPS*'- *Weighted Constraint Satisfaction Problem*, generalização de um problema de satisfação de restrições (CSP)). As duas visões de entendimento, geral e específica, são interessantes e uma análise conjunta tende a melhorar a qualidade das conclusões sobre os perfis obtidos. Dessa forma, conclui-se que para alguns caso específicos, a aplicação de diferentes abordagens para a definição do perfil de grupo é relevante, uma vez que a análise das diversas visões pode levar a conclusões mais ricas.

Tabela 20 – Perfis gerados para o grupo 156.

Chi-Square/Betweenneess	Chi-Square/PageRank	Chi-Square
algorithms	approaches	algorithms
approximations	approximations	approach
bayesian	bayesian networks	bayesian
belief	conditional	belief
inference	data	inference
model	distribution	model
networks	function	networks
paper	probabilistic	papers
probability	reasoning	probability
variables	structure	variables

O perfil mais votado para o grupo 6, Tabela 20, foi o gerado pelo Qui-Quadrado combinado com a centralidade PageRank. Para este grupo, os três perfis gerados mostraram-se gerais, mas todos com termos voltados para a aprendizagem de máquina, com foco em estudos estocásticos (como '*belief*', '*inference*', '*probabilistic*' e '*Bayesian networks*'). A mesma conclusão obtida no grupo 6, indica que as comunidades, em geral, têm o mesmo foco de estudo, e com os perfis disponíveis é possível desenvolver pesquisas ou mesmo a unificação das comunidades. Analisando-se a Figura 19, as comunidades 6 e 156 têm relacionamentos com alguns autores da comunidade 80, o que reafirma a deescrição fundamenta nos perfis gerados, realmente tratam-se de comunidades de autores com atuações não dissimilares. Podendo, ainda, tratar-se da necessidade de um acompanhamento temporal das comunidades; esse seria facilitado pela abordagem baseada em informações relacionais, restringindo a acompanhar apenas uma amostra dos autores (mais relevantes).

Continuando a análise, de maneira geral os três perfis gerados para o grupo 156 levam a conclusões semelhantes. Isso ratifica o propósito da abordagem baseada em informações relacionais, se podem reduzir os dados de entrada (consequentemente, tempo necessário de coleta os dados), considerando apenas os usuários centrais, sem percas na qualidade das descrições. Para os grupos 6 e 156, por exemplo, o número de nós de entrada foi reduzido para 24,39% e 34,48%, respectivamente. Para obter mais detalhes sobre os ganhos

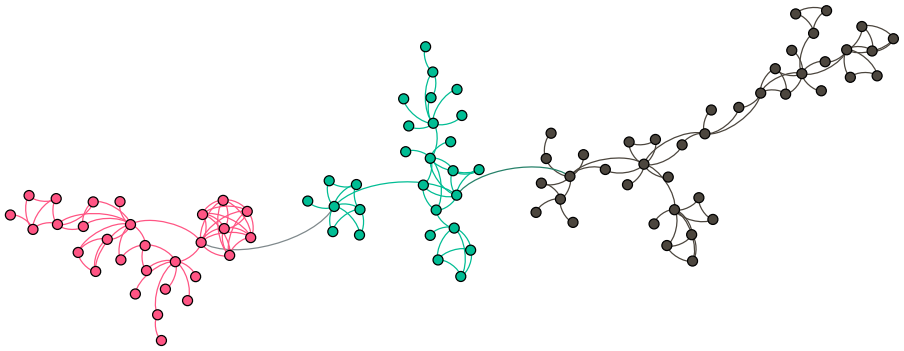

Figura 19 – Relacionamentos comuns das comunidade 6 (grafite) e 156 (rosa), com o grupo 80 (verde).

obtidos com a redução de dados de entrada, na próxima subseção 4.3.4, é apresentada uma avaliação comparando os tempos de execução das abordagens global e baseada em informações relacionais.

4.3.4 Avaliação Baseada na Viabilidade do Tempo de Execução

Com o propósito de comprovar a redução da complexidade de tempo obtida fazendo uso da abordagem baseada em informações, realizou-se uma avaliação confrontando-a com a abordagem global. Dessa forma, analisaram-se os tempos de execução para todas as combinações consideradas, abordagens (global e baseada em centralidade) e métodos/medidas. Sabendo-se que o tempo de execução pode variar, para cada método foi executado 30 vezes, computando-se a média e o desvio padrão do tempo de execução, conforme detalhado na Tabela 21. Todos os experimentos foram executados nas mesmas condições, com a seguinte configuração de hardware: processador Intel (R) Core (TM) i7-4510U, 16 GB de RAM e sistema operacional Windows 8.1 Professional de 64 bits. Todos os quatro algoritmos (BNS, Qui-Quadrado, WRS e TF-IDF) foram implementados e executados no Matlab. A API Java Gephi (BASTIAN; HEYMANN; JACOMY, 2009) foi usada para calcular as medidas de centralidade, bem como para efetuar a filtragem baseada em informações relacionais.

Como visto na Tabela 21, uma redução significativa no tempo de execução foi observada através da implementação da abordagem baseada em informações relacionais, independentemente do método de caracterização de grupos adotado. Isso de fato altera o estado da arte, na redução da complexidade computacional para problemas de caracterização de comunidades sociais. A Tabela 21 também apresenta a quantidade de tempo gasto em cada etapa da geração do perfil: o filtro de centralidade, o processamento de texto e a execução do próprio método de criação do perfil. Naturalmente, como a abordagem global não executa a filtragem dos usuários, nenhum custo é considerado nesta etapa. Analisando o tempo médio de cada passo, verifica-se que, para esse experimento, o processo menos custoso é a filtragem baseada em informações relacionais. Esse é um fato relevante, pois

Tabela 21 – Tempos médios de execução em segundos e desvio padrão das 30 execuções para cada Medida/Método (maiores períodos de duração indicados em negrito).

Médida/Método	**Aplicação do Método**	**Geração da Matriz de Dados**	**Aplicação do Filtro**	**Tempo de Execução Total Em Média**
Global				
BNS	2.9426 ± 0.7148	1.0868 ± 0.3822	-	**4.029**
Chi-Square	7.9215 ± 0.5446	1.0868 ± 0.3822	-	**9.008**
WRS	6.3192 ± 0.4140	0.9521 ± 0.4583	-	**7.271**
TF-IDF	31.0837 ± 5.4206	0.9521 ± 0.4583	-	**32.0358**
Degree				
BNS	2.1824 ± 0.3073	0.3181 ± 0.1933	0.0013 ± 0.00201	2.502
Chi-Square	5.7279 ± 0.9536	0.3181 ± 0.1933	0.0013 ± 0.00201	6.047
WRS	3.6382 ± 0.3398	0.3818 ± 0.2036	0.0013 ± 0.00201	4.021
TF-IDF	29.5975 ± 1.0786	0.3818 ± 0.2036	0.0013 ± 0.00201	29.9806
Eigenvector				
BNS	1.8307 ± 0.6582	0.2687 ± 0.2031	0.00173 ± 0.0015	2.101
Chi-Square	6.3216 ± 0.3567	0.2687 ± 0.2031	0.00173 ± 0.0015	6.592
WRS	3.4398 ± 0.6411	0.3648 ± 0.2246	0.00173 ± 0.0015	3.806
TF-IDF	30.0022 ± 1.8959	0.3648 ± 0.2246	0.00173 ± 0.0015	30.3687
Eccentricity				
BNS	1.6122 ± 0.5744	0.2249 ± 0.1369	0.0015 ± 0.00154	1.839
Chi-Square	5.2867 ± 0.7322	0.2249 ± 0.1369	0.0015 ± 0.00154	5.513
WRS	3.1766 ± 0.4826	0.1467 ± 0.1343	0.0015 ± 0.00154	3.325
TF-IDF	29.8897 ± 2.3267	0.1467 ± 0.1343	0.0015 ± 0.00154	30.0379
Betweenneess				
BNS	1.7759 ± 0.7144	0.6471 ± 0.2727	0.00266 ± 0.00754	2.426
Chi-Square	7.0254 ± 0.4406	0.6471 ± 0.2727	0.00266 ± 0.00754	7.675
WRS	3.6850 ± 0.5378	0.6012 ± 0.2450	0.00266 ± 0.00754	4.289
TF-IDF	29.8975 ± 1.6743	0.6012 ± 0.2450	0.00266 ± 0.00754	30.5014
Pagerank				
BNS	2.036 ± 0.5094	0.6707 ± 0.2934	0.00176 ± 0.00167	2.708
Chi-Square	5.6169 ± 1.8497	0.6707 ± 0.2934	0.00176 ± 0.00167	6.289
WRS	2.7994 ± 1.0794	0.5032 ± 0.2908	0.00176 ± 0.00167	3.304
TF-IDF	29.1369 ± 1.8495	0.5032 ± 0.2908	0.00176 ± 0.00167	29,6418
Closeness				
BNS	2.9950 ± 0.5532	0.2248 ± 0.1490	0.0018 ± 0.0019	3.222
Chi-Square	4.8156 ± 0.6756	0.2248 ± 0.1490	0.0018 ± 0.0019	5.042
WRS	1.4427 ± 0.6266	0.1259 ± 0.1767	0.0018 ± 0.0019	1.57
TF-IDF	28.1850 ± 2.7850	0.1259 ± 0.1767	0.0018 ± 0.0019	28.3127

se trata do único procedimento em que a abordagem baseada em informações relacionais considera todos os nós da rede. Para as outras duas etapas (processamento textual e execução do método de caracterização de grupos), apenas 26.88% dos nós da rede são considerados. Sabendo-se que o tempo de execução é proporcional ao número de nós de entrada, temos com isso uma forte indicação da eficácia da abordagem.

Observe que o principal gargalo na abordagem global é a quantidade de dados textuais a serem processados. O tempo gasto no pré-processamento de texto é reduzido quando está limitado ao conteúdo gerado pelos nós centrais. O tempo economizado no pré-processamento compensou o custo adicional de calcular a centralidade do nó em cada grupo. Na verdade, a escolha entre as abordagens dependerá do tamanho do gráfico $G = (V, E)$ e do vetor d-dimensional de recursos associados ao vértice $v_i \in V$ ($\mathbf{a} \in \mathbb{D}^d, \mathbf{a} = (a_1, a_2, ..., a_d)$, em que $\mathbb{D}$ compreende o domínio do atributo); portanto, é um problema de compensação (*tradeoff problem*), dependendo das entradas em cada módulo/etapa. Considerando que a introdução da etapa de filtragem diminui a quantidade de dados, bem como o tempo de execução geral nas duas etapas subsequentes, fica claro que a abordagem baseada em centralidade é uma vantagem no presente cenário.

Para o estudo de caso dessa tese, uma pequena rede composta por vértices com muitos atributos foi analisada. A abordagem baseada em informações relacionais foi eficaz, apresentando redução considerável no estágio de geração da matriz de dados e, consequentemente, menor tempo médio total. Como visto nos experimentos, o filtro de centralidade representa uma pequena parte do tempo total de execução para geração dos perfis. Esses resultados indicam a capacidade da abordagem proposta em produzir perfis de boa qualidade, ao mesmo tempo em que reduz significativamente os tempos computacionais. Desta forma, concluímos que a abordagem baseada em informações relacionais é efetiva para pequenas redes ricas em atributos (similar a adotada no estudo de caso). Esse resultado preliminar apresenta consideráveis indícios da efetividade da abordagem em redes maiores, uma vez que a filtragem dos nós repercute na redução dos dados para as etapas subsequentes. Todavia, se fazem necessários experimentos mais aprofundados, especialmente na análise de grandes redes, verificando, assim, o comportamento do filtro de centralidade em outros contextos. A seguir, Seção 4.4, é apresentado o experimento que valida o módulo de representação de usuários baseado em comunidades, detalhado na Seção 3.2.2.

4.4 EXPERIMENTO III - CARACTERIZAÇÃO DE GRUPOS COM REPRESENTAÇÃO DE USUÁRIOS BASEADA EM COMUNIDADES

Esses Experimentos foram conduzidos para analisar a eficácia do módulo de representação de usuários baseada em comunidades, denominado Community-based Users' Representation (CUR), na geração de perfis de grupos (Figura 20). Para alcançar este objetivo, as representações de usuários baseada em comunidades e convencional foram aplicadas nas

abordagens de caracterização de grupos global e baseada em informações relacionais. Conforme detalhado na Seção 3.2.2, o método de representação de usuários baseado em comunidades, visa atribuir conjuntos de caraterísticas relevantes para cada comunidade em particular, evitando o viés causado pelo desbalanceamento entre as comunidades na representação de usuários convencional. Dessa forma, os perfis das comunidades nesse experimento são obtidos em três etapas:

1. Seleção de Característica: coleta, pré-processamento e filtragem de atributos dos usuários, representando individualmente cada comunidade. Ou seja, uma abordagem baseada em agregação é aplicada, selecionando características internas de cada comunidade;

2. Integração de recursos: integração das representações de todas as comunidades em uma única representação de usuários;

3. Criação de perfil: aplicação de métodos de caracterização de grupos baseados em diferenciação para identificação dos rótulos que descrevem as comunidades.

No processo acima o módulo CUR corresponde aos passos 1) e 2). Inicialmente, todos os recursos individuais dos usuários são coletados, seguido por aplicações de técnicas de pré-processamento textuais (por exemplo, PNL), selecionando as informações com conteúdo semântico que serão possivelmente consideradas na descrição das comunidades. Em seguida, as características resultantes do processo são incorporadas para a filtragem/seleção (Figura 8 - A). Nesta etapa, algoritmos de seleção de recursos, como *stopwords*, *stemming*, remoção de não-substantivos e não-adjetivos (BARRERA; VERMA, 2012), são aplicados sobre cada conjunto de atributos das comunidade de forma isolada, selecionando as características semanticamente relevantes para os grupos (abordagem AGP). O objetivo desta etapa é gerar uma seleção de recursos efetiva e justa para cada grupo, desconsiderando seu tamanho relativo ao tamanho geral da rede. Depois disso, os recursos selecionados são então integrados em uma única representação (Figura 8 - B), simplesmente concatenando os recursos selecionados. Finalmente, o método de caracterização de grupos ser aplicado sobre a representação final dos usuários (Figura 8 - C)).

Nesse experimento, adotamos duas abordagens distintas para caracterização de grupos, Global (considerando toda a rede) e baseada em informações relacionais (apenas os nós centrais). Para ambas aplicou-se o método IDF Maqbool e Babri (2005) (detalhado no Capitulo 2), por se tratar de um algoritmo pouco robusto, requisitando, assim, uma boa represetação de usuários. Para a abordagem baseada em informações relacionais foi considerada a medida de centralidade de maior destaque no experimento II, o *pagerank*. Fazendo uso dos mesmos parâmetros de (GOMES; PRUDÊNCIO; NASCIMENTO, 2018a), ou seja, $n = 10$, selecionando os dez nós com as maiores pontuações de centralidade do *pageRank* em cada comunidade, a serem considerados no processo de caracterização. As

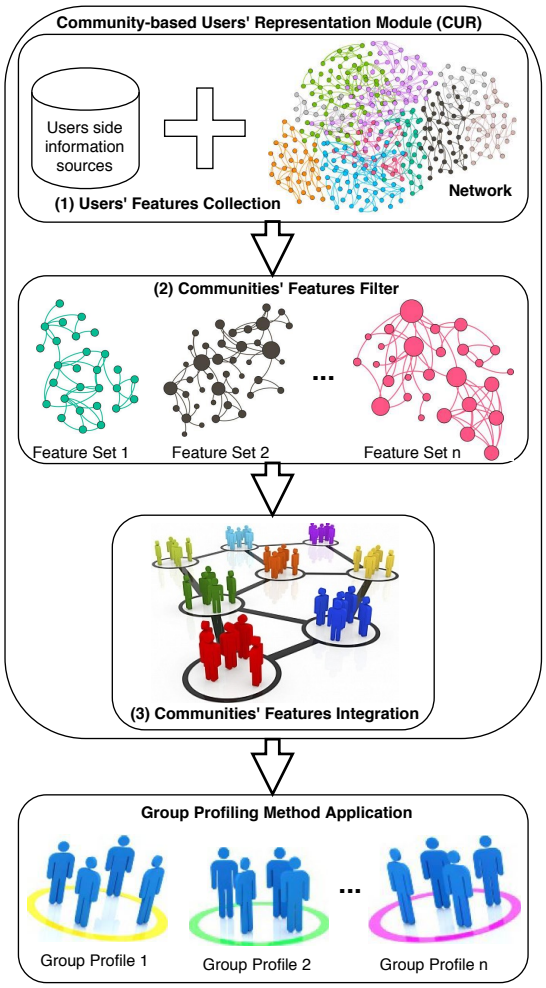

Figura 20 – Uma visão geral caracterização de grupos com representação de usuários baseada em comunidades. O processo de criação de perfil de grupo é híbrido, no qual o módulo CUR conta com uma abordagem baseada em agregação (passos (1) e (2)), enquanto a geração final dos perfis é executada por um método baseado em diferenciação (passo (3)).

estatísticas gerais da rede original e subrede pagerank são fornecidas na Tabela 22. A seguir as Seções 4.4.1 e 4.4.2, apresentam os processos de seleção dos atributos das comunidades e a integração dos dados, respectivamente. Por fim, na Seção 4.4.3, os resultados obtidos no experimento são apresentados e discutidos.

4.4.1 Seleção dos Atributos das Comunidades

Nesse experimento, considerou-se o mesmo *dataset* dos experimentos anteriores e comparamos o módulo CUR com a representação convencional de usuários. Ambos começaram com a mesma coleção de características dos autores. Idem os experimentos anteriores, os artigos publicados de cada autor foram coletados e combinados em um único documento. Em seguida, uma série de etapas de pré-processamento foram aplicadas para cada documento: tokenização, remoção de stopwords, stemming, remoção de não-substantivos, não-adjetivos e, por fim, extração de n-grams (com $n = \{1, 2, 3\}$) Barrera e Verma (2012).

Para cada comunidade isoladamente o módulo CUR selecionou as k características mais frequentes, ou seja, se aplicou o método baseado em agregação TF. Nesse experimento, consideramos $k = 1.000$ para cada um das dez comunidades, resultando em um total de ≤ 10.000 atributos, dada as possíveis repetições de termos entre as comunidades. Para a representação convencional do usuário obteve-se 10.000 termos, considerando-se toda a rede (10 comunidades).

4.4.2 Integração dos Atributos

De posse dos atributos selecionados para representar os usuários em cada comunidade, realiza-se a integração de todos em uma única representação (saco de palavras). As características selecionadas são integradas em uma representação única, pela união de todas as representações da etapa de seleção. Objetiva-se que a representação de usuários seja simultaneamente, eficaz na identificação dos atributos relevantes e equilibrada para todas as comunidades, favorecendo os métodos de caracterização na geração de perfis descritivos.

Embora essa seja uma abordagem bastante simples, o módulo CUR evita selecionar recursos que sejam relevantes apenas para comunidades maiores. Assim, o módulo retorna uma seleção de recursos balanceada para cada grupo, desconsiderando seus tamanhos/dimensões. Ressaltamos que essa etapa não é considerada na abordagem de representação dos usuários convencionais. O número final de recursos para cada abordagem é apresentado na Tabela 22.

Tabela 22 – Medidas de rede e quantidade de característica resultantes das representações de usuários para a rede global e subrede *pagerank*.

Medida	**Global**	**PageRank**
#Autores	372	100
#Links	654	157
#Grupos	10	10
Densidade	0.009	0.032
Grau Médio	3.516	3.14
Diâmetro	19	16
#Atributos Convencional	10000	10000
#Atributos CUR	7.794	7.959

Analisando a Tabela 22 verifica-se que para ambas as redes, global e *pagerank*, o módulo CUR selecionou menos atributos que a convencional, apresentando um pouco menos de 80% da quantidade de termos. Isso já era esperado, sabendo-se que alguns dos termos seriam selecionados para diferentes comunidades. A seguir na Seção 4.4.3, os resultados do experimento são apresentados e debatidos.

4.4.3 Caracterização de Grupos com Representação de Usuários Baseada em Comunidades: Um Estudo Comparativo

O presente experimento analisou a eficácia do módulo de representação dos usuários CUR. Sabendo-se que o propósito do experimento é avaliar a importância da etapa de representação dos usuários na qualidade dos perfis gerados, comparou-se a representação dos usuários convencional com o módulo CUR, os aplicando sobre duas abordagens de caracterização de grupos, global Tang, Wang e Liu (2011) e baseada em informações relacionais Gomes, Prudêncio e Nascimento (2018a). Como já supracitado, para se exigir mais da etapa de representação dos usuários durante o processo, considerou-se como método de caracterização de grupos, o simples algoritmo IDF.

Para a avaliação dos perfis gerados, realizou-se um survey qualitativo. Esse reuniu um total de 23 pessoas vinculadas a computação (graduação, pós-graduandos, professores universitários), resultando em 230 avaliações, como 115 verificações para cada abordagem de representação dos usuários (CUR e convencioal). Para a abordagem baseada em informações relacionais, se definiu os mesmos paramentos considerados em (GOMES; PRUDÊNCIO; NASCIMENTO, 2018a). Dessa forma, considerou-se n igual a 10, selecionando para cada comunidade os 10 nós com maiores scores de centralidade *PageRank*. As estatísticas gerais das redes resultantes, global e PageRank, estão disponíveis na Tabela 22. De posse das redes, estamos aptos a aplicar o algoritmo IDF para caracterização das comunidades.

Conforme defendido inicialmente, um melhor balanceamento entre as características selecionadas para as comunidade melhorou a qualidade de suas descrições. Analisando a Figura 21, observamos que, independente da abordagem de caracterização de grupos adotada (global ou baseada em informações relacionais), para todas as comunidades, os perfis gerados aplicando o módulo CUR foram melhores que o convencinal. Analisando de uma forma global, verifica-se no presente experimentos que os ganhos qualitativos obtidos com a aplicação do módulo CUR foram relevantes, alcançando a média global de 76,54%. Esses dados ficam mais proeminentes para as comunidades 256, 153 e 6, todas superiores a 80%.

Para realizar uma análise detalhada do desempenho das duas representações dos usuários, vamos mostrar dois exemplos concretos: os grupos 80 e 256. O primeiro, grupo 80, apresentou destaque entre os perfis debatidos em (GOMES; PRUDÊNCIO; NASCIMENTO, 2016) (Seção 4.2); já o grupo 256, foi o que apresentou maior unanimidade entre os avaliadores desse experimento, obtendo 87% das indicação para o módulo CUR. Nas Tabelas 23 e 24, apresentam os perfis extraídos para descrever esses grupos com base nos títulos e resumos dos artigos publicados pelos autores. Os rótulos novamente são classificados por ordem alfabética, visando facilitar a comparação entre os perfis gerados por cada método.

De acordo com o (GOMES; PRUDÊNCIO; NASCIMENTO, 2016), os termos selecionados pelos métodos *Chi-Square* e TF-IDF (mais robustos que o método IDF) para o grupo 80,

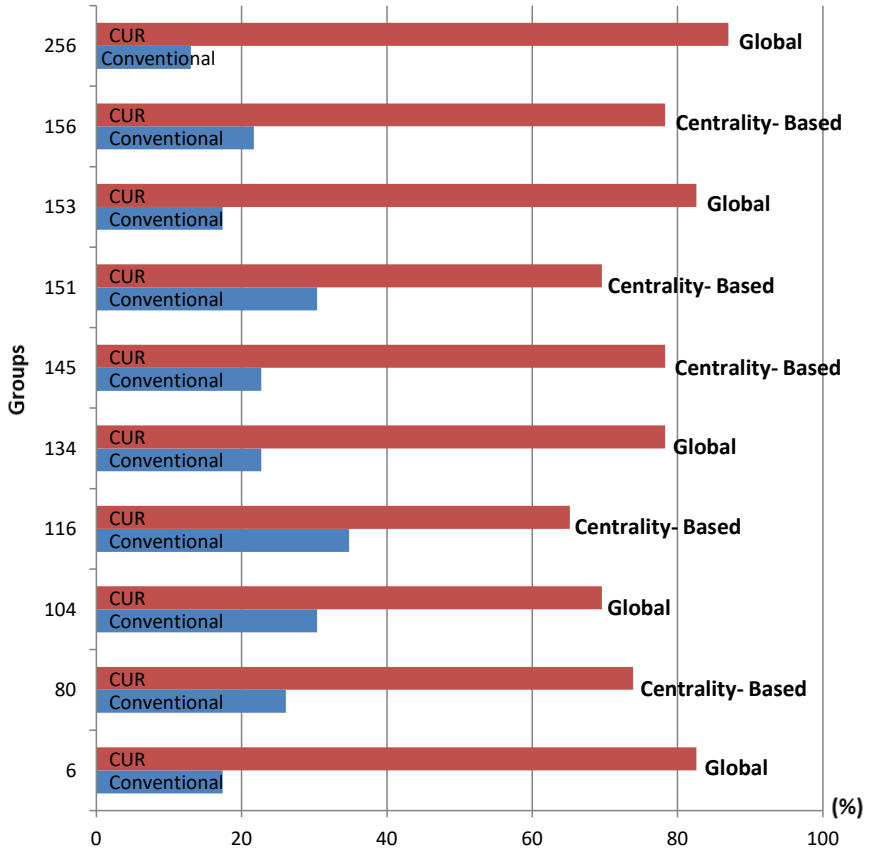

Figura 21 – Porcentagens indicando quantas vezes cada método avaliado forneceu o melhor perfil na pesquisa. As configurações possíveis são: Global + Convencional, Global + módulo CUR, Baseada em Centralidade + Convencional, e Baseada em Centralidade + módulo CUR.

apresentaram uma visão de um grupo de colaboradores focados em estudos de aprendizagem em redes bayesianas. Destacam-se entre os temos selecionados: 'Bayesian', 'Bayesian networks', 'probabilistic', 'models' e 'probability'. Analisando a Tabela 23, essa conclusão acaba por ser melhor detalhada no presente experimento, com a adição outros termos igualmente relevantes, como: 'binary bayesian networks', 'SPIC'[9], 'qualitative belief propagation' . Dessa forma, podemos concluir que a comunidade 80, trata-se de um grupo de aprendizagem de máquina baseado em modelos probabilísticos, focado no estudo das redes bayesianas e em modelos estocásticos e dinâmicos de aprendizado.

Já para o grupo 256, apesar da grande quantidade de escolhas para o perfil gerado adotando o módulo CUR, verificamos que a maioria dos termos selecionados como caracterizadores pelo algoritmo IDF foram poucos descritivos, são exemplos: 'real world', 'method single', 'observation construction', 'single outward propagation', 'diagrams approach' e 'tool adaptive systems'. Isso pode ter ocorrido pela característica do algoritmo IDF, esse atribui maior relevância aos termos que diferenciam a comunidade analisada das demais. Ou seja,

[9] Symbolic Probabilistic Inference Continuous - extensão do algoritmo SPI para lidar com redes bayesianas compostas de variáveis contínuas.

Tabela 23 – Perfis gerados para o Grupo 80 rede de coautoria arXiv.

Grupo 80	
CUR - PageRank	Convencional - PageRank
binary bayesian networks	ancestor
car comparison	bad
discovery dbcl presence	bar
edge-by-edge correction	crisis
half	discrete-valued
knowledge engineering version	half
netview subnetworks leak	illustrative
numerical probabilities	netview
qualitative belief propagation	requirmenets
SPIC	SPIC

Tabela 24 – Perfis gerados para o Grupo 256 rede de coautoria arXiv.

Grupo 256	
CUR - Global	Convencional - Global
analogous potentials cliques	appendix
diagrams approach	bags
method single	conp/poly
observation construction	distortions
posterior marginal	extent
QPNs	papamichail
qualitative reasoning	plastic
real world	pool
single outward propagation	QPNs
tool adaptive systems	spaghetti

como todos os autores da rede são da área de inteligência artificial, há a possibilidade da comunidade 256 ser "eclética", formada por membros atuantes em diferentes vertentes. Essa característica da comunidade pode ter levando o algoritmo IDF a identificar termos pouco descritivos como o perfil da comunidade.

Mas como justificar tamanha adesão dos avaliadores para os perfis gerados adotando o módulo CUR? Ambos os perfis apresentaram termos com baixo poder de descrição, porém podemos observar a inclusão de alguns termos interessantes no módulo CUR, o que justificaria uma escolha pelo perfil “menos ruim”. Dentre os atributos selecionados, destacando-se: ‘posterior marginal’ - rótulo que define o uso de marginais posteriores na estimação da verossimilhança marginal por meio de amostragem de importância; ‘qualitative reasoning’ - Sobre raciocínio em redes com incerteza qualitativa; and, ‘QPNs’

(Qualitative probabilistic networks) - rótulo referente às redes bayesianas, esse trata do fornecimento de uma semântica probabilística para afirmações qualitativas sobre a probabilidade. Pode-se concluir que, apesar da grande importância da representação dos usuários na garantia da qualidade dos perfis, o método de caracterização de grupos deve ser eficaz na identificação dos termos relevantes à descrição das comunidades.

Analisando os perfis apresentados nas Tabelas 23 e 24, com o módulo CUR obteve-se termos relevantes não considerados na representação convencional. Isso valida a hipótese inicial, confirmando que um olha singular para cada comunidade, evita a desconsideração de atributos relevantes durante o processo de representação dos usuários. Assim como, o experimento evidenciou a eficácia do módulo CUR, melhorando consideravelmente a qualidade dos perfis gerados. Todavia, ressalta-se a necessidade de experimentos mais profundos, principalmente com o foco na aplicação do módulo CUR conjuntamente a métodos de caracterização de grupos mais robustos (como por exemplos, Palavras Frequentes e Preditivas (POPESCUL; UNGAR, 2000) e Indexação Semântica Latente (KUHN; DUCASSE; GÍRBA, 2007)).

4.5 CONSIDERAÇÕES FINAIS

O presente estudo de caso, apresentou três experimentos analisados sobre a rede de coautoria arXiv, considerando publicações em Inteligência Artificial entre 2012 e 2014. Os estudos foram publicados em: (GOMES; PRUDÊNCIO; NASCIMENTO, 2016), (GOMES; PRUDÊNCIO; NASCIMENTO, 2018a) e (GOMES; PRUDÊNCIO; NASCIMENTO, 2018b). O Experimento I, Seção 4.2, teve como foco a adaptação de métodos de rotulagem de agrupamentos em documentos ao problema de caracterização de grupos. Dessa forma, foi realizando um experimento confrontando os métodos CGBD já propostos (BNS e WRS), bem como se verificou a eficácia da adaptação de métodos tradicionais de rotulagem de agrupamentos em documentos (TF-IDF e Chi-Square). Os resultados desse experimento apresentaram uma grande similaridade entre as performances dos quatro métodos, o que consolida a expectativa inicial e levanta indícios positivos para a adaptação de métodos mais robustos aplicados na caracterização de comunidades. Até onde se sabe, foi a primeira vez que as técnicas de rotulagem de cluster foram adaptadas e comparadas com algoritmos de caracterização de grupos. Bem como, representou o primeiro estudo de caso com introdução de técnicas de caracterização de grupos para a análise de colaboração científica. Esses resultados ajudam a explicar os motivos pelos quais autores se conectam e interagem em uma rede de coautoria, podendo eventualmente levar a novas abordagens para fortalecer e expandir as redes de colaboração científica.

No segundo experimento (Seção 4.3), investigamos o uso das informações relacionais na caracterização das comunidades. Uma nova estratégia baseada no uso da centralidade para filtragem dos indivíduos foi proposta. A abordagem introduziu o uso de informações relacionais para a caracterização das comunidades, Caracterização de Grupos Baseada

em Informações Relacionais (GOMES; PRUDÊNCIO; NASCIMENTO, 2018a). A principal motivação foi à busca por melhorias na escalabilidade, reduzindo significativamente os dados de entrada a serem processados pelos métodos de caracterização de grupos, sem perdas expressivas de informações. As medidas de centralidade consideradas no presente estudo foram: Grau, Proximidade, Intermediação, Autovetor, *PageRank* e Excentricidade. Os Experimentos demonstraram que a abordagem baseada em informações relacionais obteve desempenho não diferente da abordagem global, enquanto ainda reduziu consideravelmente o tempo de execução necessário para caracterização das comunidades. Para avaliar os perfis gerados pela abordagem baseada em informações relacionais, inicialmente foi realizada uma análise baseada na similaridade dos perfis. A investigação da similaridade apontou para a geração de perfis semelhantes e dissimilares entre as abordagens global e baseada em informações relacionais. A qualidade desses perfis foi investigada, realizando uma pesquisa qualitativa com seres humanos. Como conclusão, verificou-se que, para algumas medidas de centralidade, a abordagem baseada em informações relacionais poderia reduzir os dados de entrada e obter resultados equivalentes; Já para outras medidas, apresentar perfis dissimilares, todavia apontados como melhores pelos avaliadores, proporcionando assim uma nova perspectiva de entendimento para as comunidades, através da variação das medidas de centralidade. Analisaram-se subsequentemente os tempos de execução para todas as combinações das abordagens consideradas, global e baseada em informações relacionais, verificando uma redução significativa no tempo de execução com a abordagem nesta tese.

Por fim, na Seção 4.4, foi apresentada uma nova estratégia de representação de usuários, O módulo de representação do usuário baseado na comunidade (CUR) (GOMES; PRUDÊNCIO; NASCIMENTO, 2018b). Esse tratou o desbalanceamento entre os dados das comunidades, selecionando os recursos dos usuários para cada comunidade de forma singular e uniforme. Com o módulo CUR, obtiveram-se conjuntos de recursos relevantes para cada comunidade em particular, evitando o viés causado pelas comunidades maiores na representação geral dos usuários. Os Experimentos demonstraram que a aplicação do módulo CUR levou ao equilíbrio entre as concepções das comunidades e, ao mesmo tempo, elevou a qualidade dos perfis gerados. Até onde se sabe essa também foi à primeira vez que as comunidades foram consideradas no processo de representação dos usuários. No contexto da rede de coautoria, o equilíbrio entre as comunidades colaborou na extração de interesses latentes, ou seja, as razões pelas quais os autores se conectam uns com os outros em redes científicas. Essas percepções podem levar a novas abordagens para fortalecer e expandir as redes de colaboração científica. Analisando os resultados, com o módulo CUR se obtive rótulos adequados não considerados na representação convencional. Pode-se notar que para todas as comunidades, independentemente da abordagem de caracterização de grupos, CGBD ou baseada em informações relacionais, os perfis obtidos após a aplicação do módulo CUR foram melhores, atingindo uma média geral de 76,54% das indicações dos avaliadores. Isso

valida à hipótese inicial, confirmando que uma representação única para cada comunidade evita à desconsideração de atributos relevantes durante o processo de representação dos usuários. A seguir, no Capítulo 5, são apresentadas as conclusões e trabalhos futuros desta tese.

5 CONCLUSÃO

Este capítulo apresenta as considerações finais deste trabalho. Inicialmente são expostas as conclusões e contribuições obtidas pela pesquisa, traçando um paralelo entre os objetivos do trabalho e os resultados alcançados por ele. Em seguida, são discutidas algumas limitações do estudo e, por fim, direcionamentos para trabalhos futuros.

Comunidades em redes sociais virtuais têm recebido bastante atenção nos últimos anos, dada a sua importância no entendimento e análise das redes. Grande parte desse esforço está direcionada para a detecção dos agrupamentos implícitos nas redes, detecção de comunidades Xie, Kelley e Szymanski (2011); entretanto igualmente relevante é a atividade de rotulagem dos grupos, denominada caracterização de comunidades. Essa visa à descrição das comunidades a partir dos atributos individuais dos usuários Tang, Wang e Liu (2011).

Dentre as redes, as de coautoria, vêm sendo objeto de estudo contínuo, principalmente na avaliação de técnicas e abordagem para predição de links (surgimento de novas colaborações científicas) Lü e Zhou (2011). Porém o estudo das comunidades acadêmicas também se apresenta como uma atividade estratégica, uma vez que essas apresentam conceitos chaves sobre as colaborações acadêmicas e a formação dos grupos de pesquisa Savić et al. (2015). Descrever comunidades acadêmicas é algo fundamental, proporcionando entendimento e acompanhamento das pesquisas, bem como o acompanhamento das mudanças de temas nas comunidades. Isso posto, definimos como estudo de caso para esta tese as redes de coautoria, mais precisamente utilizou-se a biblioteca arXiv.

Baseado em problemas de rotulagem de agrupamentos em documentos e caracterização de comunidades sociais (intrinsecamente mineração de links), este trabalho propôs uma abordagem para caracterização de grupos sociais baseada em informações relacionais. Buscando embasamento para tanto, investigou-se não só, os aspectos gerais da caracterização de comunidades (análise de redes sociais), como também o contexto de rotulagem de agrupamentos em documentos, problema bastante similar. O entendimento das dificuldades e deficiências foi primordial para a definição dos objetivos almejados e contribuições deste trabalho para a comunidade acadêmica. Em suma, os fundamentos aprendidos ajudaram na definição do escopo do trabalho proposto, sob o qual foi possível a aplicação de conceitos, abordagens, técnicas e metodologias das duas áreas de pesquisa.

Como primeiro experimento, confrontamos os métodos de CGBD já propostos (BNS e WRS), propondo uma variação do WRS para atributos textuais, bem como verificamos a eficácia da adaptação de métodos tradicionais de rotulagem de agrupamentos em documentos (TF-IDF e Qui-quadrado) ao problema de comunidades sociais. Os resultados desse experimento não apresentaram diferenças estatísticas entre as performances dos quatro métodos, o que consolida a expectativa inicial e levanta indícios positivos para a

adaptação de abordagens de rotulagem em documento ao problema de caracterização de comunidades.

No segundo experimento investigou-se o uso das informações relacionais na caracterização das comunidades. Para tanto, o filtro baseado em informações relacionais foi avaliado, esse seleciona com base nas medidas de centralidade os usuários a serem considerados no processo de caracterização; ou seja, nós que generalizam o conteúdo das comunidades. Como conclusão, verificou-se que, para alguns casos, a abordagem baseada em informações relacionais poderia reduzir os dados de entrada, obtendo resultados equivalentes (perfis bem similares ao global). Ainda pode-se observar a geração de diferentes perfis variando as medidas de centralidade, e apesar de dissimilares ao global, alguns perfis foram selecionados como melhores, fornecendo assim uma nova perspectiva de entendimento para as comunidades. A análise subsequente dos tempos de execução para todas as combinações das abordagens consideradas (global e baseado em centralidade) mostrou uma redução significativa no tempo de execução com a abordagem baseada na centralidade. Dessa forma, a principal motivação foi alcançada, reduzir consideravelmente os dados de entrada a serem processados por um método de criação de perfil de grupo sem perdas significativas de informações relevantes. As medidas de centralidade consideradas foram: Grau, Proximidade, Intermediação, Autovetor, *PageRank* e Excentricidade. Experimentos com uma rede de coautoria do mundo real demonstraram que a aplicação da estratégia de seleção proposta leva a um desempenho muito semelhante à abordagem global, enquanto ainda reduz o tamanho do conjunto de dados de entrada e consequentemente o tempo necessário para rotulagem.

Por fim, no Experimento III, verificou-se o tratamento inicial dos dados, buscando um balanceamento adequado entre os atributos selecionados como representativos para cada comunidade. Dessa forma, foi avaliada a representação de usuários baseada em comunidades (módulo CUR). Essa diferentemente da abordagem convencional, que pré-processa os atributos considerando toda a rede, representa de forma única cada comunidade, integrando posteriormente as representações. Como experimento, comparou-se a representação proposta com a convencional. Como o propósito era verificar melhorias nos perfis a partir do pré-processamento inicial, se considerou um dos métodos mais simples dentre os baseados em diferenciação, o IDF. Analisando os resultados, com a representação de usuários baseada em comunidades se obtive rótulos adequados não considerados na representação convencional. Pode-se notar que em todas as comunidades, independentemente da abordagem caracterização de grupos, CGBD ou baseada em Informações Relacionais, os perfis obtidos após a aplicação da representação baseada em comunidades foram melhores, atingindo uma média global de 76,54% das avaliações. Isso valida à hipótese inicial, confirmando que o tratamento do desbalanceamento através de uma representação única para cada comunidade, evita à desconsideração de atributos relevantes a algumas comunidades durante o processo de representação dos usuários.

5.1 PRINCIPAIS CONTRIBUIÇÕES

As contribuições desta tese se dão em cinco linhas principais:

- **Método WRS para dados textuais:** criação de uma restrição ao métodos proposto em (GOMES et al., 2013), tornando-o viável à caracterização de grupos fazendo uso de atributos textuais;

- **Adaptação dos métodos de rotulagem de agrupamentos em documentos para o problema de caracterização de comunidades sociais:** os métodos Chi-quadrado, TF-IDF e IDF foram aplicados e validados na caracterização de comunidades sociais. Os resultados foram similares e para algumas comunidades até superiores, esses dados além de validarem os seus usos abrem um leque considerável de possibilidades de métodos e abordagens a serem aplicadas para geração de perfis descritivos;

- **Abordagem de caracterização de grupos baseada em informações relacionais:** como visto dentre os principais desafio, certamente a escalabilidade é um deles. Abordagens escaláveis não são apenas de interesse da comunidade acadêmica, são necessárias para um funcionamento viável da caracterização. Dessa forma, certamente, a mais relevante das contribuições desta tese é a caracterização de grupos baseada em informações relacionais. Essa conseguir reduzir drasticamente a complexidade temporal, onde apenas 26,88% dos nós são considerados, mantendo a essência dos conteúdos das comunidades; em alguns casos gerando melhores perfis que a abordagem CGBD (Global). Na Tabela 21, são apresentados todos os tempos de execução para as diferentes variações, e pode-se verificar a eficácia da abordagem no tratamento da complexidade. Esse resultado de fato altera o estado da arte, na redução da complexidade computacional para problemas de caracterização de comunidades sociais.;

- **Módulo CUR - representação de usuários baseada em comunidades:** visivelmente uma boa descrição para as comunidades dependerá da representação dos usuários, fase anterior, responsável por todo o pré-processamento dos atributos. Consideráveis desbalanceamentos entre as comunidades são normais, e muito se deve as diferentes dimensões entre as comunidades (por exemplos, tamanho e quantidade de conteúdos gerados). Dessa forma, a metodologia de representação de usuários baseada em comunidades foi eficaz no tratamento do desbalanceamento entre as comunidades, realizando pré-processamentos singulares para cada comunidade, evitando assim, representações superficiais. Foi à primeira metodologia a considerar informações sobre as comunidades durante o processo de representação dos usuários. No contexto da rede de coautoria, o equilíbrio entre as comunidades ajudou a extrair

interesses latentes, ou seja, as razões pelas quais os autores se conectam uns com os outros em redes científicas;

- **Estudo de caso sobre a caracterização de comunidades acadêmicas:** Até onde sabe-se esses foram os primeiros experimentos voltados para a introdução de métodos de caracterização de grupos na análise de dados de colaboração científica de coautoria. Esses esclarecimentos ajudam a explicar por que os autores se conectam e interagem com eles na rede de autoria, podendo eventualmente levar a novas abordagens para fortalecer e expandir as redes de colaboração científica.

Além destes aspectos, acredita-se que os resultados obtidos nesta tese abrem espaço para uma série de trabalhos na caracterização de grupos sociais, não apenas em redes de coautoria, mas em redes complexas em geral.

5.2 PRODUÇÃO BIBLIOGRÁFICA

5.2.1 Artigos em Periódicos

1. **GOMES, J. E. A.**; PRUDÊNCIO, R. B. C.; NASCIMENTO, A. C. A. *Centrality-Based Group Profiling: a comparative study in co-authorship networks.* New Generation Computing, [S.l.], v.36, n.1, p.59–89, Jan 2018.

5.2.2 Artigos em Conferências

1. **GOMES, J. E. A.**, PRUDÊNCIO, R. B. C.; *Educational Social Network Group Profiling: An Analysis of Differentiation-Based Methods.* In: IV Brazilian Workshop on Social Network Analysis and Mining (BraSNAM), 2015.

2. **GOMES, J. E. A.**; PRUDÊNCIO, R. B. C.; NASCIMENTO, A. C. A. *A Comparative Study of Group Profiling Techniques in Co-authorship Networks.* In: BRAZILIAN CONFERENCE ON INTELLIGENT SYSTEMS (BRACIS), 2016. Anais. . . [S.l.: s.n.], 2016. p.373–378.

3. **GOMES, J. E. A.**; PRUDÊNCIO, R. B. C.; NASCIMENTO, A. C. A. *CUR: group profiling with community-based users' representation.* In: ENIAC 2018 (), São Paulo. Anais. . . [S.l.: s.n.], 2018.

5.3 LIMITAÇÕES

Analisando o estudo de caso, realizado em três experimentos, pode-se apontar algumas limitações. No Experimento I, observou-se o desequilíbrio entre as comunidades (classes), em relação a quantidade de informações. Esse problema foi inicialmente tratado com a abordagem baseada em informações relacionais, e posteriormente, refinado com o Módulo

CUR (Experimentos II e III). Com a abordagem baseada em informações relacionais, equilibrou-se os dados das comunidades considerando a mesma quantidade de nós para cada grupo, ou seja, os 10 nós mais centrais. Todavia, sabe-se que a quantidade de informações em cada comunidade depende da atuação dos seus autores, logo comunidades com pesquisadores mais produtivos continuariam a refletir um maior quantitativo de informações no processo de caracterização. Ciente disso, no Experimento III, validou-se a representação de usuários baseada em comunidade, módulo CUR, esse seleciona as informações das comunidades de forma singular e uniforme, garantindo de forma definitiva o balanceamento entre as classes/comunidades.

Outra limitação foi a consideração de um único conjunto de dados. Isso ocorreu devido à ausência de bases de dados com perfis anotados, o que também cominou na necessidade de avaliações qualitativas feitas com seres humanos. No entanto, este foi um estudo de caso original, com principal foco no tratamento da complexidade computacional para geração de perfis descritores de comunidades. Porém, requer análises mais aprofundadas sobre o ganho em qualidade com o uso da centralidade na caracterização. Em outras palavras, ficaram evidentes os ganhos em complexidade computacional com o uso da abordagem baseada em informações relacionais, gerando perfis igualmente bons, enquanto, ao mesmo tempo, reduziu significativamente os dados de entrada, sem comprometer a qualidade dos perfis produzidos. Todavia, infelizmente para os experimentos atuais, não podemos afirmar de forma absoluta quais perfis são mais descritivos. Para alcançar esses níveis de avaliação, mais estudos são necessários.

Consequentemente, a avaliação qualitativa adotada na pesquisa também apresenta limitações, podendo ser melhorada. A simples seleção do melhor perfil, infelizmente limita as análises realizadas sobre os resultados obtidos. Inviabilizando, assim, outras ponderações relacionadas ao nível de descrição dos perfis obtidos. Reafirmando a necessidade de estudos mais aprofundados sobre as estruturas internas das comunidades, bem como os relacionamentos com membros fora do grupo, essas informações tendem a estar diretamente relacionadas à geração de perfis.

Apesar de ser visualmente observada a importância da consideração das informações relacionais no processo de caracterização de grupos, perfis descritivos com considerável redução da complexidade de tempo. A abordagem proposta realiza uma seleção de instâncias, seleção de nós em uma representação baseada em informações relacionais. Com base nisso seria interessante um estudo comparativo com outras técnicas tradicionais de seleção de instâncias. Tendo assim, outras demonstrações práticas da importância das informações relacionais na caracterização das comunidades, bem como, a eficácia das medidas de centralidade na seleção dos dados mais preponderantes.

Por fim, o presente estudo de caso considerou apenas algoritmos simples/tradicionais, BNS, TF-IDF, Qui-Quadrado, WRS e IDF. O propósito foi forçar ao máximo a abordagem proposta baseada em informações relacionais e o módulo CUR, bem como analisa-los

com os métodos tradicionalmente adotados. No entanto, se faz necessária a verificação em experimentos mais aprofundada, especialmente com o foco nas suas aplicações com métodos mais robustos.

5.4 TRABALHOS FUTUROS

Com base nas limitações identificadas e nas lições aprendidas, como trabalhos futuros são sugeridos as seguintes linhas de investigação para a extensão e melhoria das abordagens propostas:

- Realização de experimentos com diferentes redes e contextos, especialmente com o foco na aplicação do módulo CUR e da abordagem baseada em centralidade com métodos mais robustos e *baselines* (abordagem baseada em agregação);
- Estudar melhores possibilidades de formatação para a apresentação dos perfis, objetivando facilitar o entendimento dos usuários finais (melhor descrição). São exemplos de possíveis saídas: taxonomias, sumarização e regras.
- Caracterização de grupos ponderada por informações relacionais: avaliar o uso de informações relacionais como pesos durante o processo de caracterização das comunidades;
- Estudo comparativo entre o filtro baseado em informações relacionais e métodos convencionais de seleção de instâncias não supervisionadas;
- Propor uma metodologia de avaliação qualitativa efetiva, capaz de realizar uma análise comparativa mais robusta, proporcionando a obtenção de métricas que ponderem de forma efetiva os níveis de descrições dos perfis obtidos.

REFERÊNCIAS

AIELLO, L. M. Group types in social media. In: PALIOURAS, G.; PAPADOPOULOS, S.; VOGIATZIS, D.; KOMPATSIARIS, Y. (Ed.). *User Community Discovery*. Springer International Publishing, 2015, (Human–Computer Interaction Series). p. 97–134. ISBN 978-3-319-23834-0. Disponível em: <http://dx.doi.org/10.1007/978-3-319-23835-7_5>.

AIELLO, L. M. The nature of social structures. In: _____. *Encyclopedia of Social Network Analysis and Mining*. New York, NY: Springer New York, 2017. p. 1–16. ISBN 978-1-4614-7163-9. Disponível em: <https://doi.org/10.1007/978-1-4614-7163-9_110180-1>.

AIELLO, L. M.; SCHIFANELLA, R.; STATE, B. Reading the source code of social ties. In: *Proceedings of the 2014 ACM Conference on Web Science*. New York, NY, USA: ACM, 2014. (WebSci'14), p. 139–148. ISBN 978-1-4503-2622-3. Disponível em: <http://doi.acm.org/10.1145/2615569.2615672>.

ALBERT, R. Z.; BARABáSI, A. L. Statistical mechanics of complex networks. *Reviews of Modern Physics*, v. 74, n. 1, p. 47–97, 2002.

ALMACK, J. The school childs choice of companions. *School and Society 16*, p. 529–530, 1922.

ALMEIDA, L. J. *Detecção de comunidades em rede complexas utilizando estratégia multinível*. Tese (MSc em Ciência da Computação e Matemática Computacional) — Universidade de São Paulo, 2009.

APPEL, A. P.; CUNHA, R. L. F.; AGGARWAL, C. C.; TERAKADO, M. M. Temporally evolving community detection and prediction in content-centric networks. In: BERLINGERIO, M.; BONCHI, F.; GÄRTNER, T.; HURLEY, N.; IFRIM, G. (Ed.). *Machine Learning and Knowledge Discovery in Databases*. Cham: Springer International Publishing, 2019. p. 3–18. ISBN 978-3-030-10928-8.

BACK, K. W. Exchange and power in social life. by peter m. blau. new york: John wiley amp; sons, 1964. 352 pp. *Social Forces*, v. 44, n. 1, p. 128, 1965. Disponível em: <http://dx.doi.org/10.2307/2574842>.

BACKSTROM, L.; HUTTENLOCHER, D.; KLEINBERG, J.; LAN, X. Group formation in large social networks: membership, growth, and evolution. In: *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. [S.l.: s.n.], 2006. (KDD '06), p. 44–54.

BARABáSI, A. L. *Linked: a nova ciência dos networks*. São Paulo, Brasil: Leopardo, 2009.

BARBIER, G.; TANG, L.; LIU, H. Understanding online groups through social media. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, John Wiley & Sons, Inc., v. 1, n. 4, p. 330–338, 2011. ISSN 1942-4795. Disponível em: <http://dx.doi.org/10.1002/widm.37>.

BARRERA, A.; VERMA, R. Computational linguistics and intelligent text processing: 13th international conference, cicling 2012, new delhi, india, march 11-17, 2012, proceedings, part ii. In: _____. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012.

cap. Combining Syntax and Semantics for Automatic Extractive Single-Document Summarization, p. 366–377.

BASTIAN, M.; HEYMANN, S.; JACOMY, M. *Gephi: An Open Source Software for Exploring and Manipulating Networks*. 2009. Disponível em: <http://www.aaai.org/ocs/index.php/ICWSM/09/paper/view/154>.

BAUMES, J.; GOLDBERG, M.; MAGDON-ISMAIL, M.; WALLACE, A. Discovering hidden groups in communication networks. In: *IN PROCEEDINGS OF THE 2ND NSF/NIJ SYMPOSIUM ON INTELLIGENCE AND SECURITY INFORMATICS*. [S.l.: s.n.], 2004.

BIGGS, N. *Graph theory, 1736-1936*. New York, USA: Oxford University Press, 1976.

BLONDEL, V.; GUILLAUME, J.; LAMBIOTTE, R.; LEFEBVRE, E. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, v. 10, p. 8, 2008.

BOLLEN, J.; GONçALVES, B.; RUAN, G.; MAO, H. Happiness is assortative in online social networks. *Artif. Life*, MIT Press, Cambridge, MA, USA, v. 17, n. 3, p. 237–251, ago. 2011.

BOLLOBAS, B. *Modern Graph Theory*. [S.l.]: Springer, 1998.

BONACICH, P. Factoring and weighting approaches to status scores and clique identification. *Journal of Mathematical Sociology*, v. 2, n. 1, p. 113–120, 1972.

BORGATTI, S. P.; EVERETT, M. G. A graph-theoretic perspective on centrality. *Social Networks*, v. 28, n. 4, p. 466 – 484, 2006. ISSN 0378-8733. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0378873305000833>.

BOTT, H. Observation of play activities in a nursery school. *Genetic Psychology Monographs 4*, p. 44–88, 1928.

BOUDIN, F. *A Comparison of Centrality Measures for Graph-Based Keyphrase Extraction*. 2013.

BRANDES, U. A faster algorithm for betweenness centrality. *Journal of Mathematical Sociology*, v. 25, p. 163–177, 2001.

BRANDãO, W.; PARREIRAS, F.; SILVA, A. *Redes em Ciência da Informação: evidências comportamentais dos pesquisadores e tendências evolutivas das redes de coautoria.* [S.l.], 2007. Disponível em <http://www.uel.br/revistas/uel/index.php/informacao/article/view/1778/1516>. Último acesso em 05/10/2018.

CARMEL, D.; ROITMAN, H.; ZWERDLING, N. Enhancing cluster labeling using wikipedia. In: *Proceedings of the 32Nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York, NY, USA: ACM, 2009. (SIGIR '09), p. 139–146. ISBN 978-1-60558-483-6. Disponível em: <http://doi.acm.org/10.1145/1571941.1571967>.

CHA, S.-H. *Comprehensive Survey on Distance/Similarity Measures between Probability Density Functions*. 2007.

CHAHINE, C. A.; CHAIGNAUD, N.; KOTOWICZ, J.; PÉCUCHET, J. Context and keyword extraction in plain text using a graph representation. *CoRR*, abs/0912.1421, 2009. Disponível em: <http://arxiv.org/abs/0912.1421>.

CHEN, W.-Y.; CHU, J.-C.; LUAN, J.; BAI, H.; WANG, Y.; CHANG, E. Y. Collaborative filtering for orkut communities: discovery of user latent behavior. In: *Proceedings of the 18th international conference on World wide web*. [S.l.: s.n.], 2009. (WWW '09), p. 681–690.

CHINTALAPUDI, S. R.; PRASAD, M. H. M. K. A survey on community detection algorithms in large scale real world networks. In: *Computing for Sustainable Global Development (INDIACom), 2015 2nd International Conference on*. [S.l.: s.n.], 2015. p. 1323–1327.

CLAUSET, A.; MOORE, C.; NEWMAN, M. E. J. Hierarchical structure and the prediction of missing links in networks. *Nature*, v. 453, n. 7191, p. 98–101, 2008.

COHEN, R.; EREZ, K.; AVRAHAM, D. ben; HAVLIN, S. *Breakdown of the internet under intentional attack*. [S.l.], 2001. ArXiv. Disponível em <https://arxiv.org/pdf/cond-mat/0010251.pdf/>. Último acesso em 05/10/2018.

COSTA, L. F.; RODRIGUES, F. A.; TRAVIESO, G.; BOAS, P. R. V. Characterization of complex networks: A survey of measurements. In: *Advances in Physics*. [S.l.: s.n.], 2007.

DAKICHE, N.; TAYEB, F. B.-S.; SLIMANI, Y.; BENATCHBA, K. Tracking community evolution in social networks: A survey. *Information Processing & Management*, 2018. ISSN 0306-4573. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0306457317305551>.

DEMŠAR, J. Statistical comparisons of classifiers over multiple data sets. *J. Mach. Learn. Res.*, JMLR.org, v. 7, p. 1–30, dez. 2006. ISSN 1532-4435. Disponível em: <http://dl.acm.org/citation.cfm?id=1248547.1248548>.

DEO, N. *Graph theory with applications to engineering and computer science*. Englewood Cliffs, New Jersey, USA: Prentice Hall, 1974.

EVERT, S. The statistics of word cooccurrences: word pairs and collocations. *Unpublished doctoral dissertation, Institut f*
"ur maschinelle Sprachverarbeitung, Universit
"at Stuttgart, 2004. Disponível em: <http://scholar.google.de/scholar.bib?q=info:6tx-Vnyw1ooJ:scholar.google.com/&output=citation&hl=de&as_sdt=2000&ct=citation&cd=0>.

FIGUEROA, A.; ATKINSON, J. Contextual language models for ranking answers to natural language definition questions. *Comput. Intell.*, Blackwell Publishers, Inc., Cambridge, MA, USA, v. 28, n. 4, p. 528–548, nov. 2012. ISSN 0824-7935. Disponível em: <http://dx.doi.org/10.1111/j.1467-8640.2012.00426.x>.

FORMAN, G. An extensive empirical study of feature selection metrics for text classification. *J. Mach. Learn. Res.*, JMLR.org, v. 3, p. 1289–1305, mar. 2003. ISSN 1532-4435. Disponível em: <http://dl.acm.org/citation.cfm?id=944919.944974>.

FORTUNATO, S.; LANCICHINETTI, A. Community detection algorithms: a comparative analysis: invited presentation, extended abstract. In: *Proceedings of the Fourth International ICST Conference on Performance Evaluation Methodologies and Tools*. [S.l.: s.n.], 2009. (VALUETOOLS '09), p. 27:1–27:2.

FREEMAN, L. C. Centrality in social networks conceptual clarification. *Social Networks*, v. 1, n. 3, p. 215–239, 1979.

FRIEDMAN, N.; GETOOR, L.; KOLLER, D.; PFEFFER, A. Learning probabilistic relational models. In: *Proceedings of the 16th International Joint Conference on Artificial Intelligence*. [S.l.: s.n.], 1999. (IJCAI '99), p. 1300–1309.

GETOOR, L.; DIEHL, C. P. Link mining: a survey. *SIGKDD Exploration Newsletter*, v. 7, n. 2, p. 3–12, 2005.

GIRVAN, M.; NEWMAN, M. E. J. Community structure in social and biological networks. *Proceedings of the National Academy of Sciences*, v. 99, n. 12, p. 7821–7826, 2002.

GLOVER, E. J.; TSIOUTSIOULIKLIS, K.; LAWRENCE, S.; PENNOCK, D. M.; FLAKE, G. W. Using web structure for classifying and describing web pages. In: LASSNER, D.; ROURE, D. D.; IYENGAR, A. (Ed.). *WWW*. ACM, 2002. p. 562–569. ISBN 1-58113-449-5. Disponível em: <http://dblp.uni-trier.de/db/conf/www/www2002.html#GloverTLPF02>.

GOMES, J.; PRUDÊNCIO, R. B. C.; MEIRA, L.; FILHO, A. A.; NASCIMENTO, A. C. A.; OLIVEIRA, H. Group profiling for understanding educational social networking. In: *The 25th International Conference on Software Engineering and Knowledge Engineering, Boston, MA, USA, June 27-29, 2013.* [S.l.: s.n.], 2013. p. 101–106.

GOMES, J. E. A.; PRUDÊNCIO, R. B. C.; NASCIMENTO, A. C. A. A comparative study of group profiling techniques in co-authorship networks. In: *2016 5th Brazilian Conference on Intelligent Systems (BRACIS)*. [S.l.: s.n.], 2016. p. 373–378.

GOMES, J. E. A.; PRUDÊNCIO, R. B. C.; NASCIMENTO, A. C. A. Centrality-based group profiling: A comparative study in co-authorship networks. *New Generation Computing*, v. 36, n. 1, p. 59–89, Jan 2018. ISSN 1882-7055. Disponível em: <https://doi.org/10.1007/s00354-017-0028-9>.

GOMES, J. E. A.; PRUDÊNCIO, R. B. C.; NASCIMENTO, A. C. A. Cur: Group profiling with community-based users' representation. In: *XV Encontro Nacional de Inteligência Artificial e Computacional (ENIAC)*. São Paulo: [s.n.], 2018.

HAGE, P.; HARARY, F. Eccentricity and centrality in networks. *Social Networks*, v. 17, n. 1, p. 57 – 63, 1995. ISSN 0378-8733. Disponível em: <http://www.sciencedirect.com/science/article/pii/0378873394002489>.

HAN, E.-H.; KARYPIS, G. Centroid-based document classification: Analysis and experimental results. In: *Proceedings of the 4th European Conference on Principles of Data Mining and Knowledge Discovery*. London, UK, UK: Springer-Verlag, 2000. (PKDD '00), p. 424–431. ISBN 3-540-41066-X. Disponível em: <http://dl.acm.org/citation.cfm?id=645804.669671>.

HANNA, M. Data Mining in the e-Learning Domain. *Campus-Wide Information Systems*, v. 21, n. 1, p. 29–34, 2004.

HE, J.; CHEN, D. A fast algorithm for community detection in temporal network. *Physica A: Statistical Mechanics and its Applications*, v. 429, p. 87 – 94, 2015. ISSN 0378-4371. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0378437115001922>.

HE, J.; CHEN, D.; SUN, C.; FU, Y.; LI, W. Efficient stepwise detection of communities in temporal networks. *Physica A: Statistical Mechanics and its Applications*, v. 469, p. 438 – 446, 2017. ISSN 0378-4371. Disponível em: <http://www.sciencedirect.com/science/article/pii/S037843711630824X>.

HENNIG, P.; BERGER, P.; STEUER, C.; WUERZ, C.; MEINEL, C. Cluster labeling for the blogosphere. In: *Big Data and Cloud Computing (BdCloud), 2014 IEEE Fourth International Conference on*. [S.l.: s.n.], 2014. p. 416–423.

HULPUS, I.; HAYES, C.; KARNSTEDT, M.; GREENE, D. An eigenvalue-based measure for word-sense disambiguation. In: YOUNGBLOOD, G. M.; MCCARTHY, P. M. (Ed.). *FLAIRS Conference*. AAAI Press, 2012. ISBN 978-1-57735-558-8. Disponível em: <http://dblp.uni-trier.de/db/conf/flairs/flairs2012.html#HulpusHKG12>.

HULPUS, I.; HAYES, C.; KARNSTEDT, M.; GREENE, D. Unsupervised graph-based topic labelling using dbpedia. In: *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*. New York, NY, USA: ACM, 2013. (WSDM '13), p. 465–474. ISBN 978-1-4503-1869-3. Disponível em: <http://doi.acm.org/10.1145/2433396.2433454>.

IENCO, D.; MEO, R. Towards the automatic construction of conceptual taxonomies. In: SONG, I.-Y.; EDER, J.; NGUYEN, T. (Ed.). *Data Warehousing and Knowledge Discovery*. Springer Berlin Heidelberg, 2008, (Lecture Notes in Computer Science, v. 5182). p. 327–336. ISBN 978-3-540-85835-5. Disponível em: <http://dx.doi.org/10.1007/978-3-540-85836-2_31>.

JEONG, H.; MASON, S. P.; BARABASI, A. L.; OLTVAI, Z. N. Lethality and centrality in protein networks. *Nature*, Nature Publishing Group, v. 411, n. 6833, p. 41–42, maio 2001. ISSN 0028-0836. Disponível em: <http://dx.doi.org/10.1038/35075138>.

KHRIBI, M.; JEMNI, M.; NASRAOUI, O. Automatic recommendations for e-learning personalization based on web usage mining techniques and information retrieval. In: *Advanced Learning Technologies, 2008. ICALT '08. Eighth IEEE International Conference on*. [S.l.: s.n.], 2008. p. 241–245.

KIMBALL, A. W. Methods of statistical analysis. cyril h. goulden. *The Quarterly Review of Biology*, v. 29, n. 2, p. 201–202, 1954. Disponível em: <https://doi.org/10.1086/400231>.

KOUZNETSOV, A.; ZOUAQ, A. Knowledge management and acquisition for smart systems and services: 13th pacific rim knowledge acquisition workshop, pkaw 2014, gold cost, qld, australia, december 1-2, 2014. proceedings. In: _____. Cham: Springer International Publishing, 2014. cap. A Comparison of Graph-Based and Statistical Metrics for Learning Domain Keywords, p. 260–268. ISBN 978-3-319-13332-4. Disponível em: <http://dx.doi.org/10.1007/978-3-319-13332-4_21>.

KUHN, A.; DUCASSE, S.; GÍRBA, T. Semantic clustering: Identifying topics in source code. *Inf. Softw. Technol.*, Butterworth-Heinemann, Newton, MA, USA, v. 49, n. 3, p. 230–243, mar. 2007. ISSN 0950-5849. Disponível em: <http://dx.doi.org/10.1016/j.infsof.2006.10.017>.

KUMAR, R.; NOVAK, J.; TOMKINS, A. Structure and evolution of online social networks. In: *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*. [S.l.: s.n.], 2006. p. 611–617.

KURKA, D. B.; GODOY, A.; ZUBEN, F. J. V. Online social network analysis: A survey of research applications in computer science. *CoRR*, abs/1504.05655, 2015.

Lü, L.; ZHOU, T. Link prediction in complex networks: A survey. *Physica A*, v. 390, n. 6, p. 11501170, 2011.

LEHMANN, J.; ISELE, R.; JAKOB, M.; JENTZSCH, A.; KONTOKOSTAS, D.; MENDES, P. N.; HELLMANN, S.; MORSEY, M.; KLEEF, P. van; AUER, S.; BIZER, C. DBpedia - a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web Journal*, v. 6, n. 2, p. 167–195, 2015. Disponível em: <http://jens-lehmann.org/files/2014/swj_dbpedia.pdf>.

LESKOVEC, J.; LANG, K. J.; DASGUPTA, A.; MAHONEY, M. W. Statistical properties of community structure in large social and information networks. In: *Proceedings of the 17th international conference on World Wide Web*. [S.l.: s.n.], 2008. (WWW '08), p. 695–704.

LIU, X.; BOLLEN, J.; NELSON, M. L.; SOMPEL, H. V. de. Co-authorship networks in the digital library research community. *CoRR*, abs/cs/0502056, 2005. Disponível em: <http://arxiv.org/abs/cs/0502056>.

LOPES, G. R.; MORO, M. M.; WIVES, L. K.; OLIVEIRA, J. P. M. de. Collaboration recommendation on academic social networks. In: TRUJILLO, J.; DOBBIE, G.; KANGASSALO, H.; HARTMANN, S.; KIRCHBERG, M.; ROSSI, M.; REINHARTZ-BERGER, I.; ZIMÁNYI, E.; FRASINCAR, F. (Ed.). *Advances in Conceptual Modeling – Applications and Challenges*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010. p. 190–199. ISBN 978-3-642-16385-2.

LUHN, H. P. A statistical approach to mechanized encoding and searching of literary information. *IBM J. Res. Dev.*, IBM Corp., Riverton, NJ, USA, v. 1, n. 4, p. 309–317, out. 1957. ISSN 0018-8646. Disponível em: <http://dx.doi.org/10.1147/rd.14.0309>.

MALLIAROS, F. D.; VAZIRGIANNIS, M. Clustering and community detection in directed networks: A survey. *CoRR*, abs/1308.0971, 2013. Disponível em: <http://arxiv.org/abs/1308.0971>.

MANNING, C. D.; RAGHAVAN, P.; SCHüTZE, H. *Introduction to Information Retrieval*. New York, NY, USA: Cambridge University Press, 2008. ISBN 0521865719, 9780521865715.

MANNING, C. D.; SCHüTZE, H. *Foundations of Statistical Natural Language Processing*. Cambridge, MA, USA: MIT Press, 1999. ISBN 0-262-13360-1.

MAQBOOL, O.; BABRI, H. Interpreting clustering results through cluster labeling. In: *Emerging Technologies, 2005. Proceedings of the IEEE Symposium on*. [S.l.: s.n.], 2005. p. 429–434.

MARTIN-BORREGON, D.; AIELLO, L.; GRABOWICZ, P.; JAIMES, A.; BAEZA-YATES, R. Characterization of online groups along space, time, and social dimensions. *EPJ Data Science*, Springer Berlin Heidelberg, v. 3, n. 1, 2014.

MCPHERSON, M.; SMITH-LOVIN, L.; COOK, J. M. Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, v. 27, p. 415–444, 2001.

MCPHERSON, M.; SMITH-LOVIN, L.; COOK, J. M. Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, v. 27, n. 1, p. 415–444, 2001.

MEI, Q.; CAI, D.; ZHANG, D.; ZHAI, C. Topic modeling with network regularization. In: *Proceedings of the 17th international conference on World Wide Web*. [S.l.: s.n.], 2008. (WWW '08), p. 101–110.

MIHALCEA, R.; TARAU, P. Textrank: Bringing order into texts. In: LIN, D.; WU, D. (Ed.). *Proceedings of EMNLP 2004*. Barcelona, Spain: Association for Computational Linguistics, 2004. p. 404–411. Disponível em: <http://www.aclweb.org/anthology/W04-3252>.

MINAYO, M.; SANCHES, O. Quantitativo-qualitativo: Oposição ou complementaridade? *Cadernos De Saude Publica - CAD SAUDE PUBLICA*, v. 9, 09 1993.

MITCHELL, T. M. *Machine Learning*. 1. ed. New York, NY, USA: McGraw-Hill, Inc., 1997. ISBN 0070428077, 9780070428072.

NEWMAN, M. E. J. Clustering and preferential attachment in growing networks. *Physical Review E*, v. 64, 2001.

NEWMAN, M. E. J. The structure and function of complex networks. *Society for Industrial and Applied Mathematics Review*, v. 45, n. 2, p. 167–256, 2003.

Newman, M. E. J. From the cover: Modularity and community structure in networks. *Proceedings of the National Academy of Science*, v. 103, p. 8577–8582, 2006.

OLSHANNIKOVA, E.; OLSSON, T.; HUHTAMÄKI, J.; KÄRKKÄINEN, H. Conceptualizing big social data. *Journal of Big Data*, v. 4, n. 1, p. 3, Jan 2017. ISSN 2196-1115. Disponível em: <https://doi.org/10.1186/s40537-017-0063-x>.

POPESCUL, A.; UNGAR, L. H. *Automatic Labeling of Document Clusters*. 2000.

QUERCIA, D.; AIELLO, L. M.; SCHIFANELLA, R. The emotional and chromatic layers of urban smells. In: *ICWSM'16: Proceedings of the 10th AAAI International Conference on Weblogs and Social Media*. [S.l.]: AAAI, 2016.

RAMAGE, D.; HEYMANN, P.; MANNING, C. D.; GARCIA-MOLINA, H. Clustering the tagged web. In: *Second ACM International Conference on Web Search and Data Mining (WSDM 2009)*. [s.n.], 2009. Disponível em: <http://ilpubs.stanford.edu:8090/890/>.

RIDGWAY, J. Implications of the data revolution for statistics education. *International Statistical Review*, v. 84, n. 3, p. 528–549, 2016. ISSN 1751-5823. 10.1111/insr.12110. Disponível em: <http://dx.doi.org/10.1111/insr.12110>.

ROLE, F.; NADIF, M. Beyond cluster labeling: Semantic interpretation of clusters' contents using a graph representation. *Know.-Based Syst.*, Elsevier Science Publishers B. V., Amsterdam, The Netherlands, The Netherlands, v. 56, p. 141–155, jan. 2014. ISSN 0950-7051. Disponível em: <http://dx.doi.org/10.1016/j.knosys.2013.11.005>.

Sá, H. R. de; PRUDENCIO, R. B. C. Supervised link prediction in weighted networks. In: *Proceedings of the 2011 International Joint Conference on Neural Networks*. [S.l.: s.n.], 2011. (IJCNN '11), p. 2281–2288.

S., L. P. R. Community ties: patterns of attachment and social interaction in urban neighborhoods. In: . [S.l.: s.n.], 1981.

SALTON, G. *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 1989. ISBN 0-201-12227-8.

SAVIĆ, M.; IVANOVIĆ, M.; RADOVANOVIĆ, M.; OGNJANOVIĆ, Z.; PEJOVIĆ, A.; KRÜGER, T. J. Ict innovations 2014: World of data. In: _____. Cham: Springer International Publishing, 2015. cap. Exploratory Analysis of Communities in Co-authorship Networks: A Case Study, p. 55–64.

SCOTT, J. P.; CARRINGTON, P. J. *The SAGE Handbook of Social Network Analysis*. [S.l.]: Sage Publications Ltd., 2011. ISBN 1847873952, 9781847873958.

SILVA, A. B. O.; PARREIRAS, F. S.; MATHEUS, R. F.; BRANDãO, W. C. Redes de coautoria dos professores da ciência da informação: um retrato da colaboração científica dessa disciplina no brasil. *Encontros Bibli*, v. 7, p. 441–452, 2007.

SILVA, S. R. P.; PEREIRA, R. Aspectos da interação humano-computador na web social. In: *Proceedings of the VIII Brazilian Symposium on Human Factors in Computing Systems*. [S.l.: s.n.], 2008. p. 350–351.

SORIć, I.; DINJAR, D.; ŠTAJCER, M.; OREščANIN, D. Efficient social network analysis in big data architectures. In: *2017 40th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*. [S.l.: s.n.], 2017. p. 1397–1400.

STEMPEL, I. G. H.; HARGROVE, T.; BERNT, J. P. Relation of growth of use of the internet to changes in media use from 1995 to 1999. *Journalism and Mass Communication Quarterly*, v. 77, n. 1, p. 71–79, 2000.

SUN, J.; FALOUTSOS, C.; PAPADIMITRIOU, S.; YU, P. S. Graphscope: parameter-free mining of large time-evolving graphs. In: *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*. [S.l.: s.n.], 2007. p. 687–696.

TAN, W.; BLAKE, M. B.; SALEH, I.; DUSTDAR, S. Social-network-sourced big data analytics. *IEEE Internet Computing*, v. 17, n. 5, p. 62–69, set. 2013. ISSN 1089-7801.

TANG, L.; LIU, H. Bias analysis in text classification for highly skewed data. In: *Proceedings of the Fifth IEEE International Conference on Data Mining*. Washington, DC, USA: IEEE Computer Society, 2005. (ICDM '05), p. 781–784. ISBN 0-7695-2278-5. Disponível em: <http://dx.doi.org/10.1109/ICDM.2005.34>.

TANG, L.; LIU, H. *Community Detection and Mining in Social Media*. New York, USA: Morgan and Claypool, 2010.

TANG, L.; LIU, H. Understanding group structures and properties in social media. In: *Link Mining: Models, Algorithms, and Applications*. [S.l.]: Springer New York, 2010. p. 163–185.

TANG, L.; LIU, H.; ZHANG, J.; AGARWAL, N.; SALERNO, J. J. Topic taxonomy adaptation for group profiling. *ACM Trans. Knowl. Discov. Data*, ACM, New York, NY, USA, v. 1, n. 4, p. 1:1–1:28, fev. 2008. ISSN 1556-4681. Disponível em: <http://doi.acm.org/10.1145/1324172.1324173>.

TANG, L.; WANG, X.; LIU, H. *Understanding Emerging Social Structures - A Group Profiling Approach*. 2010.

TANG, L.; WANG, X.; LIU, H. Group profiling for understanding social structures. *ACM Trans. Intell. Syst. Technol.*, v. 3, p. 15:1–15:25, 2011.

TREERATPITUK, P.; CALLAN, J. Automatically labeling hierarchical clusters. In: *Proceedings of the 2006 International Conference on Digital Government Research*. Digital Government Society of North America, 2006. (dg.o '06), p. 167–176. Disponível em: <http://dx.doi.org/10.1145/1146598.1146650>.

WALTHER, J. Group and interpersonal effects in international computer-mediated collaboration. In: . [S.l.: s.n.], 1997. p. 342–369.

WANG, X.; TANG, L.; GAO, H.; LIU, H. Discovering overlapping groups in social media. In: *Proceedings of the 2010 IEEE International Conference on Data Mining*. [S.l.: s.n.], 2010. p. 569–578.

Wang, Z.; Zhang, D.; Zhou, X.; Yang, D.; Yu, Z.; Yu, Z. Discovering and profiling overlapping communities in location-based social networks. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, v. 44, n. 4, p. 499–509, April 2014. ISSN 2168-2216.

WASSERMAN, S.; FAUST, K. *Social Network Analysis: methods and applications (structural analysis in the social sciences)*. New York, USA: Cambridge University Press, 1994.

WATTS, D. J.; STROGATZ, S. H. Collective dynamics of small-world networks. *Nature*, v. 393, n. 6684, p. 440–442, 1998.

WEI, T.; LU, Y.; CHANG, H.; ZHOU, Q.; BAO, X. A semantic approach for text clustering using wordnet and lexical chains. *Expert Systems with Applications*, v. 42, n. 4, p. 2264 – 2275, 2015. ISSN 0957-4174. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0957417414006472>.

WELLMAN, B. The influence of intelligence on the selection of associates. *Journal of Educational Research 14*, p. 126–132, 1926.

WITTEN, I. H.; FRANK, E. *Data Mining: Practical Machine Learning Tools and Techniques, Second Edition (Morgan Kaufmann Series in Data Management Systems)*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2005. ISBN 0120884070.

XIE, J.; KELLEY, S.; SZYMANSKI, B. K. Overlapping community detection in networks: the state of the art and comparative study. *CoRR*, abs/1110.5813, 2011. Disponível em: <http://arxiv.org/abs/1110.5813>.

YANG, Y.; PEDERSEN, J. O. A comparative study on feature selection in text categorization. In: *Proceedings of the Fourteenth International Conference on Machine Learning*. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 1997. (ICML '97), p. 412–420. ISBN 1-55860-486-3. Disponível em: <http://dl.acm.org/citation.cfm?id=645526.657137>.

YAQOOB, I.; HASHEM, I. A. T.; GANI, A.; MOKHTAR, S.; AHMED, E.; ANUAR, N. B.; VASILAKOS, A. V. Big data. *Int. J. Inf. Manag.*, Elsevier Science Publishers B. V., Amsterdam, The Netherlands, The Netherlands, v. 36, n. 6, p. 1231–1247, dez. 2016. ISSN 0268-4012. Disponível em: <https://doi.org/10.1016/j.ijinfomgt.2016.07.009>.

YUAN, Y. C.; GAY, G. Homophily of network ties and bonding and bridging social capital in computer-mediated distributed teams. *J. Computer-Mediated Communication*, v. 11, n. 4, p. 1062–1084, 2006.

YURUK, N.; XU, X.; SCHWEIGER, T. A. J. Structure-connected as functional groups in metabolic networks. In: *Proceedings of the International Workshop and Conference on Network Science*. [S.l.: s.n.], 2007. (NetSci '07).

ZAFARANI, R.; COLE, W.; LIU, H. Sentiment propagation in social networks: A case study in livejournal. In: CHAI, S.-K.; SALERNO, J.; MABRY, P. (Ed.). *Advances in Social Computing*. Springer Berlin Heidelberg, 2010, (Lecture Notes in Computer Science, v. 6007). p. 413–420. ISBN 978-3-642-12078-7. Disponível em: <http://dx.doi.org/10.1007/978-3-642-12079-4_52>.