Universidade Federal de Pernambuco
Centro de Informática

Graduação em Ciência da Computação

Caracterização de Comunidades Científicas usando Subgroup Discovery

Trabalho de Graduação

Autor: *Ângelo de Sant'Ana Santos Dias*
Orientador: *Prof. Renato Vimieiro*
Área: *Ciência dos Dados*

Recife
Novembro de 2018

A Cristo, Nosso Senhor, à Virgem Maria e à minha família que me apoiam desde sempre conforme as possibilidades e meu merecimento. Àqueles que de mim vierem a depender.

Agradecimentos

Sou imensamente grato ao Prof. Renato Vimieiro pelo auxílio, guia e paciência durante todo o tempo da minha orientação, por ter sido de fato um orientador e ter ensinado-me tanto.

Sou grato também aos diversos professores do Centro de Informática, cujos em algum momento transmitiram-me conhecimento em grau excelente, bem como a todos os que de algum modo são responsáveis por manter esse centro de excelência.

E finalmente, gostaria de agradecer a todos os meus amigos e às minhas diversas famílias em Recife, na Bahia, Brasil afora, e além dessas fronteiras, que com muito amor, conselhos, orações, companhia e convivência familiar e edificante me auxiliaram a concluir essa missão.

«Reach up as high as you can today, and God will reach down the rest of the way»

—

Resumo

Uma comunidade pode ser definida a partir de um conjunto de atores que interagem entre si frequentemente. No meio científico, são diversas as comunidades formadas e já existem métodos automatizados que realizam de modo eficiente essa descoberta. No entanto, ainda se investiga como realizar a caracterização das comunidades através dos seus documentos. Desses documentos, pode-se coletar ainda uma grande quantidade de variáveis (dimensões) que definem as comunidades. Por sua vez, estas podem ser reconhecidas através da associação dos autores dos documentos científicos produzidos, formando redes de comunidades de coautoria. Tendo essas redes representadas pelas muitas variáveis coletadas dos documentos, pode-se aplicar alguma área que utilize essas variáveis para encontrar descrições dessas comunidades. A *Subgroup Discovery (SD)* surge para solucionar o problema da caracterização, pois objetiva descobrir relações entre os valores do conjunto. Dentro dessa área há um algoritmo, o *Simple Search Discriminative Patterns (SSDP)*, que gera os resultados em tempo aceitável e com representação facilmente compreensível. O presente trabalho objetiva aplicar o SSDP para obter descrições de redes de coautoria para comunidades científicas.

Palavras-chave: Ciência dos Dados, Descoberta de subgrupos, Comunidades Científicas, Redes de coautoria, Problema de alta dimensionalidade, *Simple Search Discriminative Patterns*

Abstract

A community can be defined from a set of actors that interact with each other frequently. In the scientific environment, there are several formed communities and there are already automated methods that efficiently perform this discovery. However, it is still investigated how to carry out the characterization of communities through their documents. From these documents, a large number of variables (dimensions) can be collected that define the communities. In turn, these can be recognized through the association of the authors of the scientific documents produced, forming networks of co-authoring communities. Having these networks represented by the many variables collected from the documents, one can apply some area that uses these variables to find descriptions of these communities. Subgroup Discovery (SD) arises to solve the problem of characterization because it aims to discover relationships between the values of the set. Within this area there is an algorithm, the Simple Search Discriminative Patterns (SSDP), which generates the results in acceptable time and with easily understandable representation. The present work aims to apply the SSDP to get descriptions of co-authorship networks for scientific communities.

Keywords: Data Science, Subgroup Discovery, Scientific Communities, Co-authorship networks, High dimensional problem, Simple Search Discriminative Patterns

Sumário

Lista de Figuras

Lista de Tabelas

CAPÍTULO 1

Introdução

Uma comunidade pode ser considerada como um conjunto de atores que interagem entre si frequentemente. Muitos estudos têm sido realizados buscando descobrir como esses atores interagem, ou seja, como se conectam [28]. A partir disso, há o desafio da descoberta de comunidades científicas. O avanço da computação permitiu que o processo realizado para essa descoberta fosse automatizado, porém ainda se investiga como, e qual a melhor forma, para se caracterizar comunidades [12].

A caracterização de comunidades científicas exige que no processo de solução, descrições dessas comunidades sejam descobertas. Essas descrições do conjunto de dados, que são interpretáveis e ainda não foram completamente exploradas [23, 20], são comumente representadas através de variáveis (dimensões), normalmente muitas - por volta das quatro mil, considerando um arquivo com os termos dos documentos coletados e após a remoção de stopwords e geração de tokens -, que interagem entre si e descrevem as comunidades. Portanto, esse problema poderia ser solucionado com auxílio de diversas áreas que compartilham técnicas e princípios probabilísticos [30].

Uma dessas áreas que permite obter descrições interpretáveis dos exemplos de um conjunto de dados é *Subgroup Discovery (SD)*. SD objetiva descobrir relações entre os valores do conjunto com relação a uma propriedade específica visando uma variável alvo. No contexto de comunidades científicas, SD irá buscar as variáveis estatisticamente mais interessantes, ou seja, irá buscar descrições daquelas comunidades, definidas por essas variáveis, que apresentam tamanho de destaque e características incomuns.

Várias outras áreas, dentre elas medicina, bioinformática e segmentação de consumidores, têm utilizado SD [16, 22]. No entanto, mais recentemente surgiu um algoritmo de SD, *Simple Search Discriminative Patterns (SSDP)*, que busca resolver esse problema em tempo aceitável e com resultado facilmente compreensível. Resultado alcançado de tal modo por ser o SSDP um algoritmo que utiliza técnicas heurísticas baseando-se em Computação Evolucionária e *Beam Search* [23].

Como SSDP é uma solução geral para a obtenção de descrições, com grande número de variáveis, dos exemplos do conjunto de dados [23], pode-se aplicá-lo à caracterização de comunidades científicas. Então, considerando-se uma base de dados de redes de coautoria para comunidades científicas, aplicamos o SSDP a fim de descrever os relacionamentos das propriedades mais interessantes e assim caracterizar as comunidades científicas.

CAPÍTULO 2

Contexto e Trabalhos Relacionados

Neste capítulo será apresentando o contexto geral de pesquisa deste trabalho a fim de embasá-lo adequadamente e ao fim apresentar-se-ão trabalhos relevantes relacionados.

2.1 Subgroup Discovery

Existem meios automatizados para descoberta de comunidades, porém falta a geração de uma descrição interpretável dessas comunidades [3]. As técnicas de predição e descrição podem ser usadas para se compreender o relacionamento entre variáveis de um conjunto de exemplos, porém elas têm suas falhas. As técnicas de predição buscam mais a acurácia na classificação, enquanto que as técnicas de descrição buscam as características no conjunto de dados. Surge então *Subgroup Discovery*, que agrupa o melhor desses dois tipos de técnicas [15] e pode ser aplicada para gerar descrições interpretáveis de comunidades.

Subgroup Discovery é uma técnica de mineração de dados que gera um modelo simples com alto nível de interesse a partir de uma população de indivíduos (consumidores, grupos, ...) e uma propriedade desses indivíduos da qual se interessa [16]. Além disso, *Subgroup Discovery* não necessariamente objetiva alcançar exatidão no resultado, mas encontrar as regras independentes (padrões individuais) que sobressaem-se [1], devem ser representadas numa forma simbólica explícita e precisam ser relativamente simples de modo que possam ser contestadas/avaliadas por usuários que têm domínio sobre os dados [18] a fim de se obter conhecimento.

Ao fim do processo realizado pela técnica *Subgroup Discovery*, tendo a propriedade alvo (de interesse) como guia para saber-se quais são as regras independentes que desviam do conjunto de dados geral e são relevantes, obtém-se o conjunto das suas k descrições que são altamente interessantes conforme medidas escolhidas [19] (e.g. o grupo de características que mais vezes se repete entre os trabalhos dos autores de uma comunidade científica).

As descrições dos subgrupos são definidas a partir de: $T = \{a_1, a_2, ..., a_k\}$, onde T é um conjunto de k variáveis chamadas de atributos. Um atributo a_i é categórico se tem um conjunto previamente definido e limitado de valores possíveis $\{v_{i1}, v_{i2}, ..., v_{im}\}$ e é contínuo se assume qualquer valor no intervalo [min, max]. Características são da forma $a_i = v_{ij}$ para atributos categóricos e da forma $a_i > valor$ ou $a_i \leqslant valor$ para atributos contínuos. Com isso, as descrições dos subgrupos são as conjunções das características que são específicas de uma classe C selecionada de indivíduos, donde obtém-se a regra $X \rightarrow C$, onde X é o conjunto atributo-valor [17].

No presente trabalho assume-se que uma "população" de indivíduos (trabalhos acadêmicos,

autores, suas respectivas comunidades e etc.) será dada, bem como uma propriedade interessante desses indivíduos. A partir disso, a tarefa de caracterização das comunidades realiza-se ao descobrir os subgrupos da população que são estatisticamente "mais interessantes", i.e., são tão grandes quanto possível e têm as características estatísticas mais incomuns com respeito à propriedade de interesse [31].

Estando com os dados em termos dos subgrupos, este trabalho busca compreender as razões que levaram à formação das comunidades, por exemplo, se um conjunto de regras de tamanho n, previamente definido, com palavras dos trabalhos de autores dessas comunidades é suficiente para caracterizá-la e se há uma palavra que seja importante em várias comunidades. Portanto, o presente trabalho aplicou o processo de *Ciência dos Dados* sobre o resultado de um algoritmo que aplica a *Subgroup Discovery*, sobre uma base de dados que possui redes de coautoria para comunidades científicas, a fim de descrever os relacionamentos das suas propriedades mais interessantes, avaliar a qualidade do que foi estudado e apresentar de forma clara o conhecimento obtido.

2.1.1 Padrões Discriminativos (PDs)

Viu-se que a técnica *Subgroup Discovery* realiza a compreensão de um conjunto de dados com padrões dos subgrupos que apresentam maior destaque em comparação a uma propriedade de interesse. Esses padrões podem ser também chamados de *padrões discriminativos* e apresentam papel considerável para a descoberta dos subgrupos e obtenção de bom desempenho na classificação [10].

Os padrões discriminativos auxiliam a caracterizar as diferentes classes ao permitirem que se capture as diferenças entre conjuntos de dados. Alguns algoritmos que utilizam essa abordagem têm realizado a descoberta desses padrões usando estratégias exaustivas, ou heurísticas, de busca obtendo assim um conjunto de padrões com um resultado bom o suficiente [20].

2.1.1.1 Definição Formal de PDs

O problema dos *padrões discriminativos* pode ser definido a partir de um conjunto D que representa a base de dados onde D^+ representa os exemplos alvo da pesquisa e D^- os outros exemplos. Nesse sentido, os PDs buscam encontrar grupos onde a presença dos exemplos positivos é desproporcial em relação aos negativos. Um padrão discriminativo é formado por um ou mais itens. Cada item consiste de um par *(atributo, valor)*. O universo de todos os possíveis itens de D é dado por $I = \{i_1, ..., i_n\}$, onde $n = \mid I \mid$ [23]. Por exemplo, podemos observar que o padrão $A = \{i_1, i_2, i_3\}$ da Figura 2.1 é um padrão discriminativo, onde $A \subseteq I$.

2.1.1.2 Simple Search Discriminative Patterns (SSDP)

SSDP é uma abordagem evolucionária para mineração de padrões discriminativos que objetiva descobrir os k melhores padrões discriminativos, ou seja, o conjunto que apresenta maior relevância em relação a uma variável alvo, para descrição dos relacionamentos das propriedades mais interessantes.

Usando a abordagem de algoritmos de *Beam Search*, o algoritmo SSDP inicia a busca a

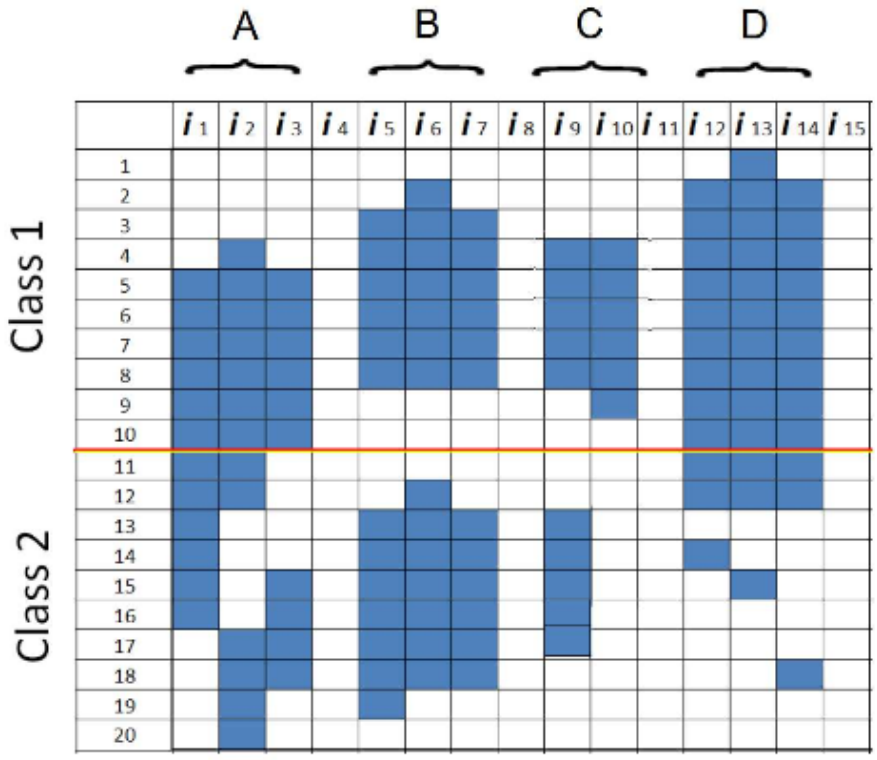

Figura 2.1 Fonte: [10]; Representação de um conjunto de dados que contêm *padrões discriminativos* (A, C e D) e um padrão que não é (B) [10]. Onde B não é um PD porque, de modo simples, não há uma predominância relevante dos padrões (i_5, i_6, i_7) do subconjunto positivo, *Class 1*, sobre o subconjunto negativo, *Class 2*, enquanto que para os demais padrões, isso ocorre.

partir de todos os padrões de tamanho unitário, onde cada padrão $i \in I$, garantindo assim que todos os itens no conjunto I sejam considerados, e em seguida evoluindo para dimensões mais altas [21].

Além disso, a abordagem SSDP agrega também o processo de algoritmo evolucionário focado em bases de dados com grande número de variáveis que descrevem os exemplos do seu conjunto de dados para geração de novas populações, onde a melhor população dentre três populações (a antiga, uma gerada pelo operador genético de crossover e outra pelo operador de mutação) é escolhida para ser a nova população [23].

Como o SSDP é um algoritmo que aplica a técnica de mineração de dados *Subgroup Discovery*, trabalhando para descobrir os padrões discriminativos e lida bem com grande número de variáveis que descrevem os exemplos do conjunto de dados, ele é uma boa opção para a caracterização de comunidades científicas, tema deste trabalho.

2.1.1.3 Definição do Algoritmo SSDP

O algoritmo SSDP trabalha com cinco populações. São elas: P (a população atual), P_c (população gerada através de crossover), P_m (população gerada através de mutação), P_* (população

com os melhores indivíduos de P, P_c e P_m) com tamanho $\mid I \mid$ e P_k (que tem os k indivíduos mais relevantes) com tamanho k [23].

A busca que o algoritmo realiza se inicia com a geração dos padrões unitários. Em seguida o SSDP usa os operadores genéticos para gerar novos candidatos. Após a geração dos candidatos, as medidas usadas no crossover e na mutação são alteradas de modo que os k melhores valores sejam propagados, tanto quando há evolução neles, aumentando o crossover e reduzindo a mutação a uma mesma taxa, como quando não ocorre evolução, reduzino o crossover e aumentando a mutação a uma mesma taxa.

O algoritmo SSDP para sua execução quando ocorre a estabilização do grupo dos k melhores padrões discriminativos após a população ter sido reiniciada duas vezes. O SSDP foi implementado em Java e está disponível num repositório de código online[1] [21].

2.2 Comunidades Científicas

É possível usar redes como meio útil para retratar comunidades [5]. Essas redes podem ser formadas a partir de interações entre indivíduos que compartilham interesses, ou seja, são formadas por conjuntos de atores que interagem entre si frequentemente. Além disso, podem estar presentes nesses conjuntos indivíduos que de algum modo realizaram atividades juntos. Elas demonstram que esses indivíduos uma vez agrupados apresentam potencial para se agregar a indivíduos de outros grupos aparentemente não relacionados.

Numa rede de co-autoria, as *comunidades científicas* desempenham papel importante pois definem a base de colaboração entre os autores. Vários estudos são realizados sobre a busca de novos relacionamentos entre colaboradores, no entanto pouco se investiga sobre como comunidades científicas em redes de co-autoria se formam, como elas são construídas, permitindo assim que se investigue as razões para a formação [12].

2.2.1 Group Profiling

Os métodos de *group profiling*, ou de outro modo, que geram *perfis descritivos*, são usados para compreender as comunidades científicas, tanto implícitas, quanto explícitas, a partir de redes de co-autoria. Esses métodos objetivam construir um *perfil descritivo para comunidades em redes complexas*, perfil que é construído a partir da extração dos atributos descritivos de um grupo de pessoas (comunidade) [12] com base em suas interações. Nesse caso, os métodos para *group profiling*, ou no presente trabalho de modo semelhante, *Subgroup Discovery*, podem ser usados para se compreender as comunidades extraídas facilitando a análise da rede [29].

Dadas as comunidades, os perfis descritivos delas são construídos a partir da seleção automática das características mais descritivas dos membros de cada comunidade. A Figura 2.2 apresenta uma visão geral da abordagem do *group profiling*.

Através dessas características individuais mais destacadas é possível descrever as características compartilhadas entre as comunidades e assim compreender a formação das comunidades explícitas [11].

[1] https://github.com/tarcisiodpl/ssdp

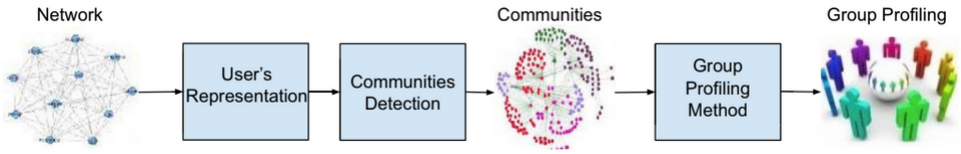

Figura 2.2 Fonte: [12]; **Estratégia do Group Profiling [12]** - *Representação dos usuários*: os atributos relevantes para a caracterização dos usuários, e formação da rede, são coletados no conjunto de dados; *detecção das comunidades*: com a rede estruturada, um algoritmo de agregação é aplicado para identificar as comunidades; *agrupamento dos perfis*: nessa fase seleciona-se as características que diferenciam o grupo dos demais na rede [11].

2.2.2 Aplicando Group Profiling Sobre Comunidades Científicas

As comunidades podem ser definidas em termos de grafos. O grafo G é definido pelo conjunto de vértices e arestas, onde $G = (V,E)$ com vértices $V = \{\mathbf{v}_1, ..., \mathbf{v}_n\}$ e arestas $E \subseteq V \times V$. Além disso, cada vértice é representado por um vetor de dimensão d, $\mathbf{A} \in \mathbb{D}^{\mathbf{d}}$, onde $\mathbb{D}$ inclui o domínio de atributo. Assume-se também que o grafo é não direcionado sem arestas de um vértice para si mesmo, isto é, $(v,u) \in E \iff (u,v) \in E$ e $(u,u) \notin E$.

Dessa definição formal dos elementos do grafo, pode-se chegar ao subgrafo que representa uma comunidade, tal que, seus vértices pertençam ao conjunto dos vértices do grafo G, bem como suas arestas. Desse modo, uma comunidade é representada pelo subgrafo $P = (V_P, E_P)$, onde $V_P \subseteq V$, $E_P \subseteq V_P \times V_P$ e $E_P \subseteq E$. Assim, a caracterização de uma dada comunidade é definida como a partição, o vetor de atributos, $\mathbf{c}_P \in \mathbb{D}^d$, $k \leq d$. Portanto, há um total de $\binom{n}{k}$ caracterizações distintas para cada comunidade. A saída esperada é uma lista dos k melhores atributos descritivos para cada partição $\mathbf{c}_P$ [12].

2.2.3 Subgroup Discovery Como Alternativa a Group Profiling

Na definição do problema de *group profiling* sobre as comunidades científicas, viu-se que o conjunto das caracterizações das comunidades pode superar e muito o conjunto amostral delas [9], podendo chegar a uma complexidade exponencial. Desse modo, onde p representa as caracterizações (variáveis) e n representa as amostras (*exemplos*), pode ocorrer o caso $p \gg n$.

Uma vez que a abordagem para *group profiling* apresenta um grande número de variáveis para as descrições sobre os exemplos do conjunto de dados, vê-se que de fato a área de *Subgroup Discovery* surge como alternativa para descobrir essas variáveis estatisticamente mais interessantes com relações incomuns [16].

2.3 Ciência dos Dados

Ciência dos Dados surge como a unificação entre os métodos matemático-estatísticos e a análise de dados - um conjunto de métodos que permitem que os relacionamentos nos dados sejam descobertos, bem como o verdadeiro significado e relevância deles [2] - com o propósito de desenvolvê-los extensiva e profundamente [14].

Ciência dos Dados é um processo que agrega diversas áreas, como: estatística; aprendizado de máquina; e outras relacionadas à coleta, processamento, avaliação e visualização de dados [14, 6]. De outro modo, o processo de Ciência dos Dados realiza um empreendimento sistemático que constrói e organiza conhecimento com previsões e explicações testáveis. O processo é abastecido por uma grande quantidade de dados normalmente heterogênea e não estruturada [8].

O processo de Ciência dos Dados é comumente compreendido sob a divisão em algumas fases. No início do processo é preciso compreender o que se deseja obter ao fim dele e se obtém conhecimento do domínio dos dados, buscando essa informação, possivelmente nas metas do negócio, que é a fase de *definição do objetivo*.

Após a definição dos objetivos, far-se-á a *seleção dos dados*, fase na qual realiza-se a busca dos dados nas diversas fontes disponíveis; o *processamento dos dados*, fase em que tendo os dados, realizam-se as primeiras investigações sobre eles, é o momento para se ficar "familiar" com os dados; a *transformação dos dados*, fase na qual ocorrem adaptações nos formatos dos dados, para um mesmo padrão, e preenchimento de dados faltantes, já que os dados normalmente não chegam prontos para análise imediata; a *mineração dos dados*, fase na qual começa-se a aplicação do algoritmo para se obter conhecimento; a *avaliação dos dados*, em cuja fase, com os dados minerados, é possível avaliar o que o resultado está dizendo, qual o conhecimento obtido e se chegou-se ao objetivo definido na primeira fase. Por fim, realiza-se a *entrega*, fase em que se realiza a apresentação e entrega dos resultados obtidos de modo que quem visualizá-los compreenda-os bem [7].

A Figura 2.3 apresenta as fases e como é possível ver nela, pode ser preciso voltar para alguma fase a fim de corrigir o andamento do processo. Mais precisamente, no decorrer do processo essas iterações entre as fases acabam ocorrendo por causa da natureza dos dados, por erros cometidos anteriormente ou por outro tipo de necessidade.

Uma vez que Ciência dos Dados tem um conjunto vasto de técnicas que são aplicadas no seu processo, ele foi aplicado para realizar a avaliação do resultado dado pelo algoritmo de *Subgroup Discovery*, SSDP. A partir do qual foi possível obter conhecimento de um modelo simples com propriedades da variável estudada. Em outras palavras, *Subgroup Discovery* foi usada para solucionar o problema da caracterização de *comunidades científicas* através de perfis descritivos (*group profiling*); problema que apresenta considerável complexidade para se solucionar.

2.4 Trabalhos Relacionados

A técnica *Subgroup Discovery* foi previamente usada para se obter a descrição de comunidades. No estudo [3], Atzmüeller et al. aplicaram *Subgroup Discovery* para a detecção de comunidade orientada a descrição. Neste estudo, foram usados grafos de conjuntos de dados de sistemas de mídias sociais. Esses grafos são gerados a partir das interações dos usuários e são usados para a busca local de comunidades. Com a base de dados definida, ela contém um registro para cada nó, onde somente os conjuntos de nós sem nós isolados são candidatos a serem comunidades.

Estudos também foram realizados sobre um conjunto de dados de movimentos de quartetos de cordas, compostos por Haydn e Mozart, usando *Subgroup Discovery* [26]. No processo de

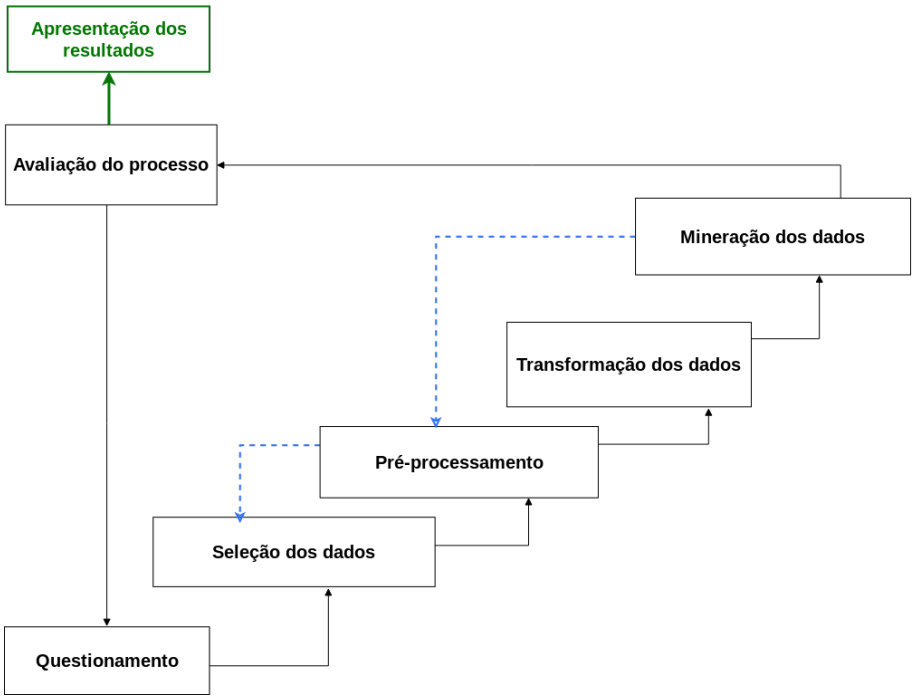

Figura 2.3 Baseada em: [7]; **Representação gráfica em fases do processo de Ciência dos Dados**

análise das bases de dados, as regras resultantes foram investigadas a partir de metadados, que foram adicionalmente coletados, e usadas para classificar os dados. Além disso, usou-se uma combinação das regras para realizar uma análise preditiva.

Estudou-se também o campo científico brasileiro para realizar a caracterização de comunidades científicas, porém sem aplicação de *Subgroup Discovery* [4]. Neste estudo, os pesquisadores analisaram as principais áreas de conhecimento avaliando o local de estudo, a competência para treinamento de novos pesquisadores e a capacidade para produção acadêmica. Estes aspectos foram usados para gerar o conjunto de variáveis usado para o mapeamento da composição do campo científico nas principais áreas do conhecimento, assim chegando à caracterização das diversas comunidades.

CAPÍTULO 3

Metodologia

Neste capítulo será apresentada a metodologia de trabalho realizada para se compreender as razões para formação das comunidades científicas e se fará detalhamento e avaliação do Processo de Ciência dos Dados aplicado sobre a base usada.

3.1 Visão Geral do Processo

Imagens que descrevem de forma geral o procedimento usado para realizar a análise do problema de caracterização de comunidades científicas serão apresentadas a seguir. Isso se dará ainda sem muitos detalhes, cujos serão fornecidos em seções seguintes. As fases apresentadas nas figuras representam como ocorreu a aplicação no presente trabalho, portanto são características dele, não necessariamente de um modelo formal.

Na Figura 3.1 as duas fases iniciais são ampliadas. Elas têm acréscimo de duas atividades realizadas. Na fase de *Questionamento* é possível ver que foram realizadas *Reuniões* seguidas do *Entendimento* da base de dados. As reuniões ocorreram com o pesquisador detentor do conhecimento dos dados, e do problema de *Group Profiling*, a fim de que os objetivos do trabalho fossem alinhados à realidade e ao que deveria ser de fato investigado. O entendimento ocorreu através do que foi informado pelo pesquisador e da pesquisa do estado da arte sobre Group Profiling [25, 29, 13, 27, 28, 11, 12], *Data Science* [8, 14, 7], *alta-dimensionalidade* [30, 23, 32, 24] e *Subgroup Discovery* [23, 21, 16, 10, 20, 22], que envolvem o problema e como solucioná-lo. Essas duas atividades foram realizadas repetidas vezes conforme o conhecimento do problema crescia.

Ainda na mesma figura é possível ver que a fase de *Seleção dos Dados* tem as atividades *Fonte dos dados* e *Coleta dos dados*. A atividade "Fonte dos dados" deu-se pela definição de qual seria a fonte provedora de dados. Isso foi feito seguindo a decisão tomada pelo pesquisador quanto à base de dados que ele formou para sua pesquisa. Por sua vez, a atividade de "Coleta dos dados" deu-se pela criação de scripts que coletassem os arquivos da fonte e consolidassem-nos para a próxima fase. Foram coletados arquivos com o mesmo conteúdo que o pesquisador selecionou.

Na Figura 3.2 é possível ver a ampliação das duas fases seguintes. Na fase de *Pré-processamento* vê-se a atividade de *Conversão PDF para TXT*. Nela, após a coleta dos arquivos em PDF, a partir da fonte de dados escolhida, realizou-se a cópia do texto desses arquivos e a consolidação deles em arquivos de formato TXT a fim de se facilitar manuseios futuros, além de armazenar em memória para poder-se realizar a atividade seguinte. Tendo o conteúdo acessível em memó-

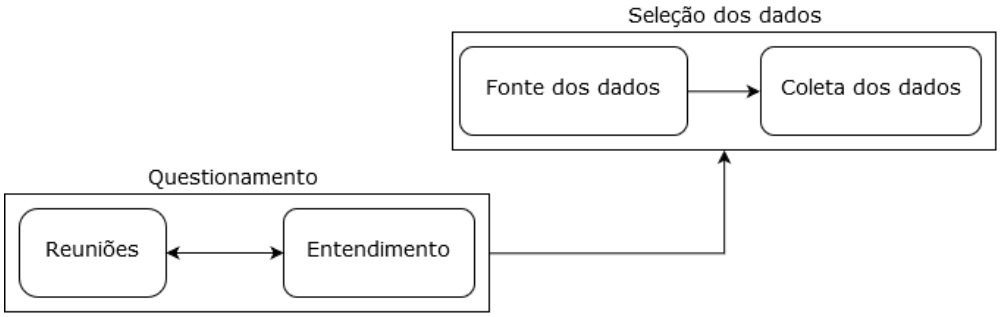

Figura 3.1 Representação gráfica das fases de Questionamento e Seleção dos Dados

ria, a atividade *Tokenizer* pode ser realizada. Nessa fase aplica-se uma adaptação de um script[1] para remoção de stopwords e geração de tokens.

Ainda na Figura 3.2, vê-se a fase de *Transformação dos dados* com as atividades *Matriz TF-IDF* e *Matriz de Frequência* que culminaram na atividade *Transformação e discretização*. Essas duas primeiras atividades geraram duas matrizes que tiveram suas colunas e linhas transformadas e seus registros discretizados gerando planilhas que foram aplicadas ao SSDP a fim de se verificar qual a melhor opção.

Uma matriz é da Frequência dos termos no documento, enquanto a outra é de TF-IDF. Para a geração delas foram usados os métodos *TfidfVectorizer*, para TF-IDF, e *CountVectorizer*, para Frequência dos termos, do sub-módulo, em *Python*, de extração de características *sklearn.feature_extraction.text*[2]. Os parâmetros passados foram os mesmos nos dois métodos, com exceção do parâmetro que define uso do IDF no método *TfidfVectorizer*.

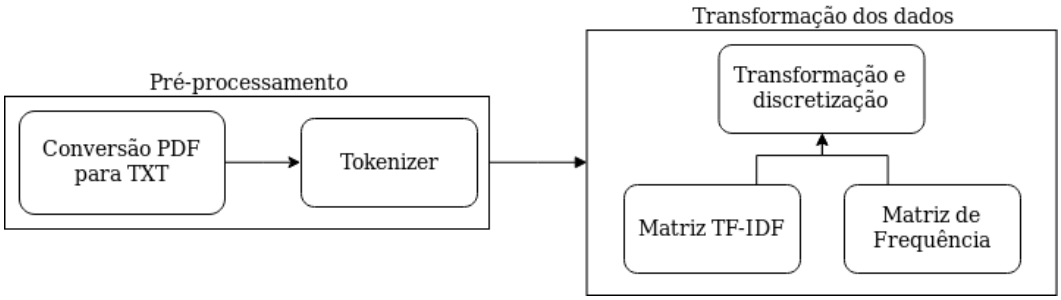

Figura 3.2 Representação gráfica das fases de Pré-processamento e Transformação dos Dados

Por fim, a Figura 3.3 apresenta as três últimas fases do processo. A fase de *Mineração dos dados* deu-se pela aplicação ao SSDP das planilhas geradas na fase anterior. Na fase de *Avaliação do processo* usando o resultado desse algoritmo separadamente para cada planilha, fez-se então uma avaliação das regras geradas para descrever a variável alvo. Disso resultaram dois arquivos com avaliações, uma para a matriz de Frequência e outra para a de *TF-IDF*, que foram fornecidas à fase de *Apresentação dos resultados*. Nesta fase, por sua vez, dois tipos de arquivos foram gerados com intuito de melhor apresentar o que foi compreendido.

[1] http://brandonrose.org/clustering#Stopwords,-stemming,-and-tokenizing
[2] http://scikit-learn.org/stable/modules/feature_extraction.html

Um tipo de arquivo (no formato CSV) apresenta mais numérica e visualmente os subgrupos compreendidos, enquanto que um texto apresenta a noção do que o processo revelou sobre os dados.

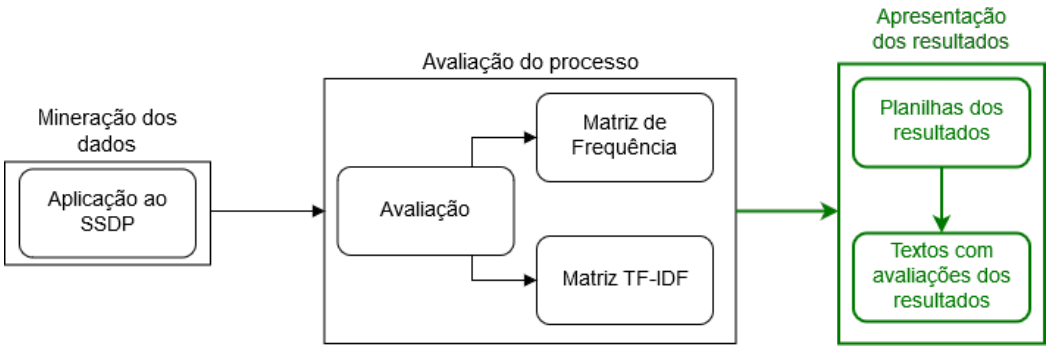

Figura 3.3 Representação gráfica das fases de Mineração dos Dados, Avaliação do processo e Apresentação dos resultados

3.2 Detalhando o Processo

3.2.1 Base de Dados

A base de dados utilizada neste trabalho foi consolidada a partir dos artigos definidos pelos pesquisadores do trabalho [13], no qual eles decidiram usar o subconjunto de Inteligência Artificial coletados do repositório arXiv[3] (a fonte dos dados) sobre o qual aplicaram um algoritmo de detecção de comunidade, resultando em 10 comunidades, apresentadas na Tabela 3.1.

Para se conseguir os mesmos artigos de IA, fez-se *data scraping* no arXiv através da API[4] tendo como partida o conjunto total dos artigos (572 ao todo), usados na pesquisa citada acima, de onde pegou-se o nome de cada artigo e fez-se uma consulta para cada um deles[5]. Nesta atividade surgiram problemas com alguns arquivos que não estavam com seus IDs acessíveis pelo resultado retornado pela consulta, então aprofundou-se nela para conseguir encontrar o ID. Finalmente com os IDs dos artigos, outras consultas foram feitas para baixar os arquivos, para a *coleta dos dados*, no formato PDF e seguir com as atividades de Ciência dos Dados.

Os arquivos em PDF coletados foram salvos numa pasta e seus conteúdos copiados para arquivos do formato TXT a fim de serem mais facilmente manejados futuramente. Nessa atividade alguns PDFs tinham uma formatação não resolvida pelo script de transformação para TXT e por isso precisaram ser tratados individualmente.

A partir dos PDFs dois conjuntos de arquivos TXT foram gerados. Um conjunto tem todo

[3] https://arxiv.org/

[4] https://arxiv.org/help/api/index

[5] Código das consultas pelos títulos no Jupyter Notebook: https://github.com/rvimieiro/SD_community/blob/master/coleta_id_arxiv.ipynb

Grupo	Tamanho
6	41
80	28
104	68
116	28
134	60
145	14
151	18
153	53
156	29
256	33

Tabela 3.1 Comunidades detectadas: *Grupo* é o identificador da comunidade e *Tamanho* é a quantidade de nós na comunidade (um nó é um autor, com acréscimo de informações de quais artigos de IA no arXiv foram de sua autoria)

o conteúdo dos PDFs[6], enquanto que o outro tem esse mesmo conteúdo sem as 20 primeiras linhas, cujas potencialmente corresponderiam ao cabeçalho do artigo. Decidiu-se por retirar o cabeçalho porque ele traz informações que são mais características dos pesquisadores predominantes num determinado subgrupo que características específicas do tema pesquisado. Esse fato, no entanto, pode ser algo determinante para a comunidade, e portanto bom, já que um autor pode ter escrita muito profícua na área de IA, dominando a comunidade a que pertence.

Os textos escolhidos no processo realizado foram os que estão com o todo o conteúdo do artigo, inclusive seu cabeçalho. Com esses dados em mãos, realizou-se então a criação da base de dados, primeiro do conjunto total, ou seja, com todos os artigos que foram coletados. Em seguida, fez-se o mesmo para o conjunto dos artigos filtrados quando os perfis foram gerados pela abordagem de centralidade *PageRank* [13]. Esses foram os dois conjuntos investigados que compuseram a base de dados.

3.2.2 Aplicação do SSDP

Para obter-se conhecimento da base de dados, testando os dois conjuntos, aplicou-se o algoritmo SSDP. Ele recebeu como entrada um arquivo do formato CSV, de uma vez foi a matriz de Frequência e noutra a de *TF-IDF*. Os arquivos têm em suas colunas os tokens gerados após remoção de stopwords, além de uma coluna com os identificadores de cada artigo pertencente à base. Os demais registros (linhas) são preenchidos com o valor discreto equivalente.

Dois conjuntos (*PageRank* e Conjunto Total) foram aplicados ao SSDP separadamente, com combinação de geradores de termos (Frequência dos termos ou *TF-IDF*), onde para o presente estudo foi escolhido o conjunto do *PageRank* com termos gerados por *TF-IDF*. Para o algoritmo, foram passados os parâmetros de qual operador entre características deveria ser usado (AND), qual a métrica de avaliação (Qg) e quantas (20) seriam as melhores palavras e

[6]Arquivos dos TXTs dos artigos completos: https://github.com/rvimieiro/SD_community/tree/master/txt

subgrupos (regras) selecionados. O algoritmo por sua vez dá como retorno um arquivo com essas informações, bem como as palavras mais relevantes e os melhores subgrupos.

A partir dos arquivos gerados para os conjuntos e matrizes, realizou-se, para cada Grupo, a agregação e simplificação dos textos, previamente à avaliação do processo[7]. A agregação foi feita ao dispor as regras geradas do conjunto total dos dados seguidas das regras do subconjunto do *PageRank*. A simplificação foi realizada através da remoção do conteúdo que representava formatação e dos valores das métricas estatísticas retornados para cada subgrupo e característica.

[7]Agregação e simplificação prévias: https://github.com/AngeloDias/SD_community/blob/master/testingPageRank/analise_doc_term_matriz_freq_with_header_discrete_AND.txt

CAPÍTULO 4

Resultados e Discussão

Neste capítulo serão apresentados os resultados do algoritmo SSDP, ou seja, as regras geradas. Dessas regras foram obtidas informações sobre o tema pesquisado naquela comunidade. Isso foi apresentado como sendo o que foi compreendido da caracterização de comunidades científicas através de perfis descritivos (que no contexto de *Subgroup Discovery* são chamados de subgrupos, ou regras). Disso, saber-se-á quais foram as razões que levaram à formação das comunidades e se o conjunto de 20 regras proposto é suficiente para descrever cada uma.

4.1 Comunidades Compreendidas

A análise foi realizada procurando as regras que descrevem algo que indique o tema pesquisado em cada comunidade. Então, para cada uma delas há uma Tabela com as 20 regras geradas pelo SSDP, com a(s) que representa(m) o tema pesquisado em negrito. A Tabela 4.1 apresenta as regras da *Comunidade 6*, cujo tema de pesquisa compreendido foi "Reforço por aprendizado" a partir dos termos "Políticas ótimas" e "Littman", que indica que o pesquisador "Michael L. Littman" tem influência nessa comunidade. Além disso, das regras pode-se compreender que o pesquisador "David Poole" também está presente na comunidade e que os trabalhos mais influentes foram desenvolvidos na Universidade da Colúmbia Britânica.

A Tabela 4.2 apresenta o tema "Redes Bayesianas", pesquisado na *Comunidade 80*. Das regras ainda é possível compreender que o pesquisador "Judea Pearl" contribuiu para a comunidade, além de "Marek J. Druzdzel" e "Adnan Darwiche". Na Tabela 4.3, a da *Comunidade 104*, vê-se que o tema pesquisado foi mais complicado de compreender já que as regras estavam compostas de características relacionadas aos autores (cujos mais destacados são "Daphne Koller" e "Craig Boutilier") ou eram termos que não expressam um tema específico. Apesar disso, foi possível compreender que o tema pesquisado nessa comunidade é "Redes Bayesianas" porque a regra em destaque apresenta o termo "network" acompanhado do nome da pesquisadora "Daphne Koller" que tem muitos trabalhos desenvolvidos sobre o tema.

A Tabela 4.4 destaca o tema "Linguagem Natural" pesquisado na *Comunidade 116*. Além disso, pode-se compreender da última regra que o pesquisador "Didier Dubois" contribui para a comunidade. A Tabela 4.5 da *Comunidade 134* apresenta o tema "Cadeias de Markov" e da quinta e sétima regras pode-se inferir que essa comunidade lida também com complexidade computacional. Como visto na Tabela 4.6, compreende-se a partir das regras apresentadas que o tema pesquisado na *Comunidade 145* é "Aprendizado de Máquina", bem como que a comunidade lida com modelos estocásticos e os trabalhos mais relevantes foram desenvolvidos na Universidade de Alberta.

#	**Regras**
1	pool=alto,similar=baixo
2	conclus=baixo,pool=alto
3	carri=alto,zhang=alto
4	vancouv canada=medio
5	choos=alto,littman=alto
6	columbia=alto,univers british=alto
7	**comput scienc univers=alto,pool=alto**
8	pool=alto,vancouv=medio
9	approxim=baixo,lang=alto
10	**littman=alto,research=baixo**
11	state=alto,zhang=alto
12	british columbia=alto,pool=alto
13	british=alto,pool=alto
14	structur=baixo,vancouv=medio
15	group=baixo,pool=alto
16	british columbia=alto,lead=medio
17	columbia=medio,univers british columbia=medio
18	structur=baixo,univers british columbia=medio
19	british columbia=alto,upper bound=baixo
20	**optim polici=medio,zhang=alto**

Tabela 4.1 Tabela das regras da *Comunidade 6*

Da Tabela 4.7 compreende-se que o pesquisado na *Comunidade 151* envolve o uso de modelo gráfico e, da última regra, tem-se que nela também se pesquisa sobre *Bucket Elimination*, temas que são pesquisados por "Rina Dechter" a pesquisadora mais destacada nesta comunidade. Na Tabela 4.8 vê-se que os temas pesquisados na *Comunidade 153* envolvem "Redes de Crença de Inferência Probabilística" e "Diagrama de Influência de Matheson". A partir das regras dela também foi possível compreender que os seus pesquisadores mais influentes são "Eric J. Horvitz" e "Ross D. Shachter", há trabalhos em associação com Medicina e as pesquisas são desenvolvidas sob o financiamento dos laboratórios de pesquisa da *Microsoft* em *Redmond* e a partir da Universidade Stanford.

Na Tabela 4.9 compreende-se que o tema pesquisado na *Comunidade 156* é "Modelo Causal de Raciocínio", o pesquisador "Judea Pearl" é bastante influente nela e os trabalhos foram desenvolvidos a partir da Universidade da Califórnia em *Los Angeles*. Por fim, na Tabela 4.10 vê-se que na *Comunidade 256* o tema de pesquisa envolve "Algoritmo Junction Tree", "Probabilidade Condicional" e "Rede Probabilística", os pesquisadores predominantes são "Steffen L. Lauritzen", "Finn Verner Jensen" e "Frank Jensen" e eles estão concentrados na Universidade de Aalborg.

#	Regras
1	**kaufmann publish=alto,pearl=medio**
2	**darwich=alto,moreov=alto**
3	moreov=alto,scienc depart univers=baixo
4	morgan kaufmann publish=alto,pearl=medio
5	pittsburgh=alto,veri=alto
6	**darwich=alto,moreov=alto,network=alto**
7	network=alto,scienc depart univers=baixo
8	**druzdzel=alto,publish=alto**
9	univers pittsburgh=alto,veri=alto
10	inc.=alto,pearl=medio
11	**bayesian=alto,darwich=alto**
12	kaufmann publish=alto,reduc=medio
13	druzdzel=alto,scienc=alto
14	morgan kaufmann publish=alto,reduc=medio
15	inc.=alto,reduc=medio,studi=alto
16	**bayesian=alto,chang=alto,kaufmann publish=alto**
17	bayesian=alto,chang=alto,pearl=medio
18	druzdzel=alto,morgan kaufmann=alto
19	inc.=alto,laboratori=medio
20	chang=alto,network node=medio

Tabela 4.2 Tabela das regras da *Comunidade 80*

4.2 Análise das Comunidades

Aplicamos o algoritmo SSDP e obtivemos os subgrupos que devem indicar qual o tema de pesquisa, com as propriedades mais interessantes de cada comunidade. O conjunto das 20 regras trazem em si características extraídas dos documentos dos artigos científicos coletados. Essas características indicam que as comunidades foram definidas por alguns autores dominantes. De fato, quando os métodos para descoberta das comunidades geraram os grupos, um autor mais relevante foi escolhido para representar a comunidade juntamente com seus principais colaboradores. Essa informação foi compreendida porque o cabeçalho dos documentos foi mantido, tendo em si os nomes dos autores e universidades onde estão instalados.

Além disso, a partir de todos os conjuntos de regras foi possível compreender o tema pesquisado na respectiva comunidade. As regras que foram melhor posicionadas pelo resultado do SSDP nem sempre foram as que definiram o tema. Em sua maioria as primeiras foram dominadas pelos dados dos autores.

A comunidade com mais regras apresentando o tema pesquisado foi a *Comunidade 256*. Esta comunidade foi também a que mais informações trouxe em suas regras, ou seja, mais temas de pesquisa foram compreendidos e informações sobre os autores. Por sua vez, as regras da *Comunidade 145* apresentaram descrição do tema pesquisado a partir de um termo que fica aparentemente no cabeçalho do artigo, reforçando a influência dele. Além desta comunidade,

#	**Regras**
1	toronto=alto
2	**boutili=alto,certain=alto**
3	daphn=baixo,random=baixo
4	boutili=alto,element=baixo
5	allow=alto,discuss abov=medio
6	daphn koller=baixo,size=medio
7	dean=baixo,final=medio
8	allow=alto,daphn koller=baixo
9	daphn koller=baixo,koller=alto
10	boutili=alto,depart comput=alto
11	**craig=alto,techniqu=alto**
12	daphn koller=alto,prior=medio
13	**daphn koller=baixo,network=medio**
14	avail=baixo,boutili=alto
15	boutili=alto,knowledg=baixo,optim=alto
16	acknowledg=baixo,darpa=baixo
17	boutili=alto,final=medio,function=alto,optim=alto
18	boutili=alto,dynam=alto
19	assign=medio,daphn=baixo
20	daphn koller=baixo,let=medio

Tabela 4.3 Tabela das regras da *Comunidade 104*

as comunidades 6, 104, 116, 134 e 156 apresentaram somente uma regra descritiva do tema e outras com informações sobre os autores e local de pesquisa. A *Comunidade 116* teve predominância de um único termo, que deixou de aparecer em somente uma regra, demonstrando sua grande importância nesta comunidade.

#	**Regras**
1	fuzzi=alto,theori=alto
2	fuzzi=alto,interpret=alto
3	fuzzi=alto,inform=alto
4	fuzzi=alto,order=medio
5	fuzzi=alto,present=alto
6	fuzzi set=alto
7	fuzzi=alto,languag=alto
8	fuzzi=alto,word=alto
9	fuzzi=alto,uncertain=alto
10	data=medio,fuzzi=alto
11	fuzzi=alto,match=alto
12	fuzzi=alto,user=alto
13	fuzzi=alto,represent=alto
14	fuzzi=alto
15	fuzzi=alto,like=baixo
16	**fuzzi=alto,natur languag=alto**
17	fuzzi=alto,studi=medio
18	fuzzi=alto,uncertainti=alto
19	approxim=medio,fuzzi=alto
20	**duboi=alto,ed=alto**

Tabela 4.4 Tabela das regras da *Comunidade 116*

#	Regras
1	expect valu=baixo,partial=alto
2	earli=baixo,use ani=baixo
3	bound=alto,use ani=baixo
4	usa=baixo,use ani=baixo
5	**limit=alto,optim solut=alto,studi=baixo**
6	limit=alto,studi=baixo,use ani=baixo
7	**bound=alto,fewer=baixo,limit=alto**
8	journal artifici=alto,limit=alto,solv=alto
9	develop=medio,guid=baixo,implement=alto
10	develop=medio,elimin=baixo,guid=baixo
11	develop=medio,implement=alto,limit=alto
12	limit=alto,solv=alto,use ani=baixo
13	mechan=baixo,solut=alto,solv=alto
14	artifici intellig=medio,implement=alto,solv=alto
15	conclus=baixo,earli=baixo,solv=alto
16	mechan=baixo,solv=alto,studi=baixo
17	guid=baixo,use ani=baixo
18	**local=alto,observ markov=medio**
19	journal artifici=alto,use ani=baixo
20	artifici intellig=medio,defin equat=baixo

Tabela 4.5 Tabela das regras da *Comunidade 134*

#	**Regras**
1	alberta=alto
2	**univers alberta=alto**
3	**learn=alto,proof=alto,stochast=alto**
4	alberta=medio,learn=alto
5	let denot=alto,univers alberta=baixo
6	empir=alto,nserc=alto
7	**confer machin learn=medio,proceed intern=medio**
8	advanc neural inform=alto,proof=alto
9	inf=alto,learn=alto
10	background=medio,nserc=alto
11	advanc neural inform=alto,proceed intern=alto
12	real-valu=alto,result section=alto
13	averag=medio,univers alberta=baixo
14	loss=alto,univers alberta=baixo
15	real-valu=alto,stochast=alto
16	result section=alto,stochast=alto
17	proof=alto,univers alberta=baixo
18	element=medio,real-valu=alto
19	loss=alto,real-valu=alto
20	nserc=alto,return=medio

Tabela 4.6 Tabela das regras da *Comunidade 145*

#	Regras
1	**dechter=alto,graphic=alto**
2	**dechter=alto,graphic model=alto**
3	**and/or=alto,graphic model=alto**
4	dechter=alto,weight=alto
5	dechter=alto,idea=alto
6	dechter=alto,simpl=baixo
7	dechter=alto,tupl=alto
8	dechter=alto,random=alto
9	condit probabl=medio,dechter=alto
10	dechter=alto,output=alto
11	**assign=alto,bucket=alto**
12	dechter=alto,ident=alto
13	elimin=alto,graphic=alto,scope=alto
14	dechter=alto,describ=medio
15	assign=alto,dechter=alto,time=medio
16	and/or=alto,ident=alto
17	condit probabl=medio,tupl=alto
18	dechter=alto,let=medio
19	dechter=alto,given figur=alto
20	**bucket=alto,dechter=alto**

Tabela 4.7 Tabela das regras da *Comunidade 151*

#	Regras
1	stanford=alto,workshop uncertainti=alto
2	artifici intellig=alto,microsoft=alto
3	probabilist=alto,workshop uncertainti artifici=alto
4	probabilist=alto,shachter=alto,stanford=alto
5	stanford=alto,workshop uncertainti artifici=alto
6	**artifici intellig=alto,microsoft research=alto**
7	infer use=alto,medic=alto
8	meek=alto
9	**belief network=alto,probabilist infer belief=alto**
10	**medic informat=baixo**
11	**redmond=alto**
12	**medic=alto,shachter=alto**
13	nonetheless=alto,transform=alto
14	**matheson influenc diagram=baixo,multipl=alto**
15	**shachter=alto,univers stanford=baixo**
16	medic=alto,workshop uncertainti artifici=alto
17	**horvitz=alto,medic=alto**
18	drawn=alto,shachter=alto
19	**matheson influenc diagram=baixo,ross=baixo**
20	ross=baixo,shachter=alto

Tabela 4.8 Tabela das regras da *Comunidade 153*

#	Regras
1	california los=alto,causal=alto
2	angel=alto,causal=alto,identifi=alto
3	california los angel=alto,identifi=alto
4	**california los angel=alto,judea=alto**
5	california los=alto,effect=alto
6	suffici=alto,ucla=alto
7	character=alto,write=alto
8	cambridg univers=alto,judea=alto
9	**california=alto,causal model=alto**
10	angel=alto,identifi=alto,suffici=alto
11	independ=alto,ucla=alto
12	**california los=alto,judea pearl=medio**
13	california los=alto,conclus=alto
14	**graphic=alto,model reason=alto**
15	judea=alto,judea pearl=medio
16	effect=alto,judea pearl=medio
17	angel=alto,california=alto,shown=alto
18	california los=alto,state=medio
19	character=alto,effect=alto,graphic=alto
20	california=alto,los angel=alto,scienc depart=alto

Tabela 4.9 Tabela das regras da *Comunidade 156*

#	Regras
1	conclud=alto,der=alto
2	**distribut=baixo,jensen=alto,junction tree=alto**
3	aalborg=alto,tree=alto
4	**junction tree=alto,mani=baixo,tree=alto**
5	**aalborg univers=alto,junction=alto**
6	**aalborg=alto,lauritzen=alto**
7	basic=alto,der=alto
8	aalborg=alto,olesen=alto
9	aalborg=alto,mani=baixo
10	**aalborg=alto,condit probabl=medio**
11	aalborg=alto,distribut=baixo
12	**correspond=baixo,junction tree=alto**
13	**accord=medio,jensen=alto**
14	aalborg=alto,introduc=medio
15	distribut=baixo,jensen=alto,shall=alto
16	distribut=baixo,introduc=medio,jensen=alto
17	jensen=alto,shall=alto,tree=alto
18	henc=medio,jensen=alto
19	jensen=alto,let denot=medio
20	**probabilist network=alto,therebi=alto**

Tabela 4.10 Tabela das regras da *Comunidade 256*

CAPÍTULO 5

Conclusões

Este trabalho objetiva principalmente compreender como as comunidades científicas, de um subconjunto de pesquisadores de Inteligência Artificial que têm artigos no arXiv, são formadas e quantas regras são capazes de caracterizar a comunidade. Para isso, realizou-se o processo de Ciência dos Dados sobre os resultados do algoritmo SSDP, cujo aplica técnicas de mineração de dados de *Subgroup Discovery* para descrever os relacionamentos das propriedades interessantes da base de dados.

Dos resultados do SSDP analisados foi possível compreender que as regras geradas não foram muito interessantes na descrição dos relacionamentos das comunidades já que trouxeram em sua maioria somente informações relacionadas aos autores, informações genéricas sobre o tema pesquisado e em pouca quantidade. Apesar disso, as características resultantes foram diversas entre as comunidades, permitindo que se visse que elas apresentam focos distintos. Por fim, ainda com o empecilho do cabeçalho, o SSDP conseguiu caracterizar as comunidades descrevendo-as a partir das 20 regras resultantes.

Partindo do exposto no presente trabalho é possível investigar futuramente qual o efeito da remoção do cabeçalho dos textos dos artigos. Disso poderia-se ver se regras com mais características relacionadas ao tema de pesquisa surgiriam. Pode-se também verificar se um conjunto de regras ainda menor é suficiente para caracterizar as comunidades com boa qualidade após a remoção do cabeçalho. E, por fim, dentro das regras, seria ainda possível verificar qual a quantidade mínima de características que preserva esse resultado.

Referências Bibliográficas

[1] T. Abudawood and P. Flach. Evaluation measures for multi-class subgroup discovery. In *Proceedings of the 2009th European Conference on Machine Learning and Knowledge Discovery in Databases - Volume Part I*, ECMLPKDD'09, pages 35–50, Berlin, Heidelberg, 2009. Springer-Verlag. ISBN 3-642-04179-5, 978-3-642-04179-2. doi: 10.1007/978-3-642-04180-8_20. URL https://doi.org/10.1007/978-3-642-04180-8_20.

[2] M. J. Albers. Quantitative data analysis—in the graduate curriculum. *Journal of Technical Writing and Communication*, 47(2):215–233, 2017. doi: 10.1177/0047281617692067. URL https://doi.org/10.1177/0047281617692067.

[3] M. Atzmüller, S. Doerfel, and F. Mitzlaff. Description-oriented community detection using exhaustive subgroup discovery. *Information Sciences*, 329:965–984, 2016. ISSN 0020-0255. doi: https://doi.org/10.1016/j.ins.2015.05.008. URL http://www.sciencedirect.com/science/article/pii/S0020025515003667. Special issue on Discovery Science.

[4] R. B. BARATA, E. ARAGÃO, L. E. F. D. SOUSA, T. M. SANTANA, and M. L. BARRETO. The configuration of the Brazilian scientific field. *Anais da Academia Brasileira de Ciências*, 86:505 – 521, 03 2014. ISSN 0001-3765. doi: 10.1590/0001-3765201420130023. URL http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0001-37652014000100505&nrm=iso.

[5] A. Clauset, M. E. J. Newman, and C. Moore. Finding community structure in very large networks. *Phys. Rev. E*, 70:066111, Dec 2004. doi: 10.1103/PhysRevE.70.066111. URL https://link.aps.org/doi/10.1103/PhysRevE.70.066111.

[6] B. K. Daniel. Reimaging research methodology as data science. *Big Data and Cognitive Computing*, 2(1):4, 02 2018. ISSN 2504-2289. doi: 10.3390/bdcc2010004. URL http://www.mdpi.com/2504-2289/2/1/4.

[7] D. Davis. The data science process. on the blog Data Science Exploration, Dec. 2015. URL https://datascienceexploration.com/2015/12/21/the-data-science-process/.

[8] V. Dhar. Data science and prediction. *Commun. ACM*, 56(12):64–73, Dec. 2013. ISSN 0001-0782. doi: 10.1145/2500499. URL http://doi.acm.org/10.1145/2500499.

[9] J. Fan and R. Li. Statistical challenges with high dimensionality: Feature selection in knowledge discovery. In *Proc. Madrid Int. Congress of Mathematicians*, volume 3, pages 595–622, 08 2006.

[10] G. Fang, W. Wang, B. Oatley, B. V. Ness, M. Steinbach, and V. Kumar. Characterizing discriminative patterns. *CoRR*, abs/1102.4104, 2011. URL https://arxiv.org/pdf/1102.4104.pdf.

[11] J. Gomes, R. Prudêncio, L. Meira, A. A. Filho, A. Nascimento, and H. Oliveira. Group profiling for understanding educational social networking. In *25th International Conference on Software Engineering and Knowledge Engineering*, volume 25, pages 101–106, 2013. URL http://www.cin.ufpe.br/~jeag/papers/seke_2014.pdf.

[12] J. E. A. Gomes, R. B. Prudêncio, and A. C. Nascimento. A comparative study of group profiling techniques in co-authorship networks. In *5th Brazilian Conference on Intelligent Systems*. IEEE, Oct. 2016.

[13] J. E. A. Gomes, R. B. C. Prudêncio, and A. C. A. Nascimento. Centrality-based group profiling: A comparative study in co-authorship networks. *New Generation Computing*, 36(1):59–89, Jan 2018. ISSN 1882-7055. doi: 10.1007/s00354-017-0028-9. URL https://doi.org/10.1007/s00354-017-0028-9.

[14] C. Hayashi. What is data science? fundamental concepts and a heuristic example. In C. Hayashi, K. Yajima, H.-H. Bock, N. Ohsumi, Y. Tanaka, and Y. Baba, editors, *Data Science, Classification, and Related Methods*, pages 40–51, Tokyo, 1998. Springer Japan. ISBN 978-4-431-65950-1.

[15] S. Helal, J. Li, L. Liu, E. Ebrahimie, S. Dawson, and D. J. Murray. Identifying key factors of student academic performance by subgroup discovery. *International Journal of Data Science and Analytics*, Jun 2018. ISSN 2364-4168. doi: 10.1007/s41060-018-0141-y. URL https://doi.org/10.1007/s41060-018-0141-y.

[16] F. Herrera, C. J. Carmona, P. González, and M. J. del Jesus. An overview on subgroup discovery: foundations and applications. *Knowledge and Information Systems*, 29(3): 495–525, Dec 2011. ISSN 0219-3116. doi: 10.1007/s10115-010-0356-2. URL https://doi.org/10.1007/s10115-010-0356-2.

[17] P. Kralj Novak, N. Lavrac, D. Gamberger, and A. Krstacic. Csm-sd: Methodology for contrast set mining through subgroup discovery. *Journal of biomedical informatics*, 42: 113–22, 09 2008. doi: 10.1016/j.jbi.2008.08.007.

[18] N. Lavrač, B. Cestnik, D. Gamberger, and P. Flach. Decision support through subgroup discovery: Three case studies and the lessons learned. *Machine Learning*, 57(1):115–143, Oct 2004. ISSN 1573-0565. doi: 10.1023/B:MACH.0000035474.48771.cd. URL https://doi.org/10.1023/B:MACH.0000035474.48771.cd.

[19] F. Lemmerich, M. Atzmueller, and F. Puppe. Fast exhaustive subgroup discovery with numerical target concepts. *Data Min. Knowl. Discov.*, 30(3):711–762, May 2016. ISSN 1384-5810. doi: 10.1007/s10618-015-0436-8. URL http://dx.doi.org/10.1007/s10618-015-0436-8.

[20] X. Liu, J. Wu, F. Gu, J. Wang, and Z. He. Discriminative pattern mining and its applications in bioinformatics. *Briefings in Bioinformatics*, 16(5):884–900, 2015. doi: 10.1093/bib/bbu042. URL http://dx.doi.org/10.1093/bib/bbu042.

[21] T. Lucas, T. C. P. B. Silva, R. Vimieiro, and T. B. Ludermir. A new evolutionary algorithm for mining top-k discriminative patterns in high dimensional data. *Appl. Soft Comput.*, 59:487–499, 2017. doi: 10.1016/j.asoc.2017.05.048. URL https://doi.org/10.1016/j.asoc.2017.05.048.

[22] P. K. Novak, N. Lavrač, and G. I. Webb. Supervised descriptive rule discovery: A unifying survey of contrast set, emerging pattern and subgroup mining. *J. Mach. Learn. Res.*, 10:377–403, June 2009. ISSN 1532-4435. URL http://dl.acm.org/citation.cfm?id=1577069.1577083.

[23] T. Pontes, R. Vimieiro, and T. B. Ludermir. SSDP: A simple evolutionary approach for top-k discriminative patterns in high dimensional databases. In *5th Brazilian Conference on Intelligent Systems, BRACIS 2016, Recife, Brazil, October 9-12, 2016*, pages 361–366, 2016. doi: 10.1109/BRACIS.2016.072. URL https://doi.org/10.1109/BRACIS.2016.072.

[24] P. Rigollet and J.-C. Hütter. High dimensional statistics. On Professional Page in MIT, Feb. 2018. URL http://www-math.mit.edu/~rigollet/PDFs/RigNotes17.pdf. Lecture Notes.

[25] C. Senot, D. Kostadinov, M. Bouzid, J. Picault, and A. Aghasaryan. Evaluation of group profiling strategies. In *IJCAI*, volume 2011, pages 2728–2733, 2011.

[26] J. Taminau, R. Hillewaere, S. Meganck, D. Conklin, A. Nowé, and B. Manderick. Applying subgroup discovery for the analysis of string quartet movements. In *Proceedings of 3rd International Workshop on Machine Learning and Music*, MML '10, pages 29–32, New York, NY, USA, 2010. ACM. ISBN 978-1-4503-0161-9. doi: 10.1145/1878003.1878014. URL http://doi.acm.org/10.1145/1878003.1878014.

[27] L. Tang, H. Liu, J. Zhang, N. Agarwal, and J. J. Salerno. Topic taxonomy adaptation for group profiling. *ACM Trans. Knowl. Discov. Data*, 1(4):1:1–1:28, Feb. 2008. ISSN 1556-4681. doi: 10.1145/1324172.1324173. URL http://doi.acm.org/10.1145/1324172.1324173.

[28] L. Tang, X. Wang, and H. Liu. Understanding emerging social structures - a group profiling approach. Technical Report TR-10-002, School of Computing, Informatics, and Decision Systems Engineering, Arizona State University, Tempe, AZ 85287, USA, 2010.

[29] L. Tang, X. Wang, and H. Liu. Group profiling for understanding social structures. *ACM Transactions on Intelligent Systems and Technology*, 3(1):15:1–15:25, Oct. 2011. ISSN 2157-6904. doi: 10.1145/2036264.2036279. URL http://doi.acm.org/10.1145/2036264.2036279.

[30] R. van Handel. Probability in high dimension, Dec. 2016. URL https://www.princeton.edu/~rvan/APC550.pdf. APC 550 Lecture Notes in Princeton University.

[31] S. Wrobel. *Inductive Logic Programming for Knowledge Discovery in Databases*, pages 74–101. Springer Berlin Heidelberg, Berlin, Heidelberg, 2001. ISBN 978-3-662-04599-2. doi: 10.1007/978-3-662-04599-2_4. URL https://doi.org/10.1007/978-3-662-04599-2_4.

[32] L. Yu and H. Liu. Feature selection for high-dimensional data: A fast correlation-based filter solution. In *Proceedings of the Twentieth International Conference on International Conference on Machine Learning*, ICML'03, pages 856–863. AAAI Press, 2003. ISBN 1-57735-189-4. URL http://dl.acm.org/citation.cfm?id=3041838.3041946.

Este volume foi tipografado em LaTeX na classe UFPEThesis (www.cin.ufpe.br/~paguso/ufpethesis) pelo autor (assd@cin.ufpe.br).