UFPE

Universidade Federal De Pernambuco
Centro De Informática

Graduação Em Ciência Da Computação

Caracterização de Comunidades Científicas usando Subgroup Discovery

Proposta De Trabalho De Graduação

Aluno: Ângelo De Sant'Ana Santos Dias (assd@cin.ufpe.br)
Orientador: Renato Vimieiro (rv2@cin.ufpe.br)

Recife, 28 de agosto de 2018

RESUMO

Uma comunidade pode ser definida a partir da interação frequente de um conjunto de atores. No meio científico, são diversas as comunidades formadas, porém não há um método automatizado eficiente de descoberta delas, sendo preciso realizar um agrupamento manual a partir do que é apresentado em documentos científicos. 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 dimensões, pode-se utilizar alguma área que solucione o problema de alta dimensionalidade. A *Subgroup Discovery (SD)* surge como provável melhor opção para solucioná-lo. Dentro dessa área há um algoritmo, o *Simple Search Discriminative Patterns (SSDP)*, que gera os resultados em tempo aceitável e com representação facilmente compreensível. A partir desse panorama de dados e técnicas, o presente trabalho objetiva aplicar o processo da *Ciência dos Dados* onde se utilizará o SSDP sobre uma base de dados que possui redes de coautoria para comunidades científicas. Sobre a saída do algoritmo se verificará se as comunidades caracterizadas por ele concordam com as apresentadas na base e as razões para formação delas.

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 the frequent interaction of a set of actors. In the scientific environment, there are several formed communities, but there is no efficient automated method of discovery of them, so a manual grouping is necessary from what is presented in scientific documents. From these documents one can still collect a large number of variables (dimensions) 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-authorship communities. Having these networks represented by many dimensions, one can use some area that solves the problem of high dimensionality. *Subgroup Discovery (SD)* is likely to be the best solution. 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. From this panorama of data and techniques, the present work aims to apply the Data Science process where the SSDP will be used on a database that has co-authorship networks for scientific communities. On the output of the algorithm will verify if the communities characterized by it agree with those presented in the base and the reasons for their formation.

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

SUMÁRIO

INTRODUÇÃO

Uma comunidade pode ser considerada 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. Porém, pouco têm-se estudado sobre as razões da formação de uma comunidade [1]. A partir disso, há o desafio da descoberta de comunidades científicas [2]. Apesar do avanço geral da computação, o processo realizado para essa descoberta ainda não alcançou um grau automatizado e de fácil entendimento. Ou seja, ainda é necessário fazer a seleção de forma manual.

A descoberta de comunidades científicas exige que no processo de solução se descubra variáveis (dimensões), normalmente muitas, que interagem entre si e que descrevem as comunidades. Portanto, esse problema poderia ser solucionado com auxílio de diversas áreas que compartilham técnicas e princípios probabilísticos [3]. Além disso, por apresentar um conjunto com muitas dimensões, o problema pode ser chamado também de alta dimensionalidade, que ainda não foi completamente explorado [4].

Uma dessas áreas que lida com problemas de alta dimensionalidade 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 aquelas comunidades, definidas por essas variáveis, que apresentam tamanho de destaque e características incomuns. Por buscar as características interessantes, SD obtém relações não necessariamente completas, mas parciais [5].

Várias outras áreas, dentre elas medicina, bioinformática e segmentação de consumidores, têm utilizado SD [6]. Em bioinformática, por exemplo, a alta dimensionalidade é um problema presente e que de forma geral tem tido pouco esforço dedicado [7]. No entanto, mais recentemente surgiu um algoritmo de SD, *Simple Search Discriminative Patterns (SSDP)*, que tenta 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* [4].

Dado que pouco tem-se estudado sobre a formação de comunidades científicas [1] e problemas de alta dimensionalidade não foram completamente explorados [4], é possível reunir esses dois problemas e aplicar um algoritmo de SD seguindo o processo de Ciência dos Dados a fim de obter conhecimento.

Como SSDP é uma solução geral para o problema de alta dimensionalidade [4], pode-se aplicá-lo à descoberta de comunidades científicas. Então, considerando-se uma base de dados de redes de coautoria para comunidades científicas, aplica-se o SSDP a fim de descrever os relacionamentos das propriedades mais interessantes e assim caracterizar as comunidades científicas. Ainda aplicando o processo de Ciência dos Dados, seria preciso usar métricas, em parte estatísticas, de comparação dos resultados obtidos pelo algoritmo com relação à base de dados e em seguida exibir de forma clara o conhecimento obtido.

OBJETIVOS

O objetivo geral deste trabalho é aplicar o processo de Ciência dos dados à caracterização de comunidades científicas. Esse processo será realizado sobre uma base de dados que descreve comunidades científicas. Na fase de análise da base será utilizado o algoritmo SSDP para obter regras que definam os grupos, em seguida se fará a análise dessas regras com relação à base, ou seja, se verificará se o que foi retornado é consistente em relação aos dados. Em se tratando do processo natural da Ciência dos dados, poderá ser preciso voltar a uma determinada fase a fim de ajustar parâmetros. Chegando-se a uma avaliação consistente e bem documentada das comunidades caracterizadas, então se realizará a apresentação do conhecimento obtido.

ESTRUTURA DO TRABALHO

O trabalho será composto pelos seguintes capítulos:

1. **Introdução** - Apresentará uma caracterização de grupo, o contexto geral das comunidades científicas, do problema de alta dimensionalidade, um algoritmo que obtém resultado para tal problema e breve descrição de um modelo do processo de Ciência dos Dados que será aplicado;
2. **Contexto e trabalhos relacionados** - Neste capítulo se apresentará o contexto das comunidades científicas, problemas de alta dimensionalidade, o algoritmo SSDP, descrição de modelo de Ciência dos dados e trabalhos relacionados;
3. **Metodologia** - Apresentar-se-ão os dados usados para a análise e o processo de Ciência dos Dados para caracterização das comunidades com aplicação do algoritmo SSDP;
4. **Resultados** - Serão apresentados os resultados obtidos a partir dos dados utilizando-se análise qualitativa em conformidade com o modelo de Ciência dos Dados;
5. **Conclusão** - Neste capítulo serão apresentadas as conclusões obtidas a partir do trabalho desenvolvido, bem como, limitações no decorrer do trabalho e possibilidades de desenvolvimento de trabalhos futuros;
6. **Referências Bibliográficas** - Serão apresentadas as referências utilizadas para construção do trabalho.

CRONOGRAMA

	Agosto				*Setembro*				*Outubro*			
Revisão da literatura												
Montagem dos resultados												
Escrita do relatório												
Preparação da defesa												
Defesa												

As fases acima listadas são melhor explicadas abaixo:

1. **Revisão da literatura**: nessa fase o trabalho realizado terá por resultado uma coleção de fontes (artigos) que servirão para embasar a escrita do TG;
2. **Montagem dos resultados**: nesta fase espera-se ter alcançado os objetivos do trabalho e com os dados necessários, far-se-á a montagem dos resultados;
3. **Escrita do relatório**: aqui dar-se-á a escrita propriamente dita do texto do Trabalho de Graduação tendo por fim o processo bem documentado no artefato final do TG, que deverá estar devidamente formatado em conformidade aos padrões adequados. Nesta fase também será realizada a entrega do artefato aos professores, o orientador e o avaliador, para que façam as devidas considerações próprias desse tipo de trabalho;
4. **Preparação da defesa**: nesta fase será montada a apresentação da defesa do TG;
5. **Defesa**: nesta fase se realizará a defesa do trabalho realizado da fase 1 a 3.

POSSÍVEIS AVALIADORES

São possíveis avaliadores do trabalho a ser produzido conforme as especificações nesta proposta:

1. Paulo Salgado Gomes de Mattos Neto
2. Ricardo Bastos Cavalcante Prudêncio

ASSINATURAS

_____

Ângelo de Sant'Ana Santos Dias
(Aluno)

_____

Renato Vimieiro
(Orientador)

REFERÊNCIAS BIBLIOGRÁFICAS

[1] L. Tang, X. Wang, e H. Liu, “Understanding Emerging Social Structures - A Group Profiling Approach”. School of Computing, Informatics, and Decision Systems Engineering, Arizona State University, Tech. Rep. TR-10-002. 2010.

[2] J. Gomes, R. Prudêncio, A. Nascimento. “A Comparative Study of Group Profiling Techniques in Co-Authorship Networks”. Brazilian Conference on Intelligent Systems, 5ª, IEEE, Recife, 2016, pp. 373-378.

[3] R. Handel. “Probability in High Dimension”. Lecture Notes. 2014, Princeton University

[4] T. Pontes, R. Vimieiro e T. Ludermir, “SSDP: A Simple Evolutionary Approach for Top-K Discriminative Patterns in High Dimensional Databases”. Brazilian Conference on Intelligent Systems, 5ª, IEEE, Recife, 2016, pp. 361-366.

[5] Herrera, F., Carmona, C. J., González, P., & Del Jesus, M. J. (2011). An overview on subgroup discovery: Foundations and applications. Knowledge and Information Systems, 29, 495–525.

[6] Kralj-Novak P, Lavrac N, Webb GI (2009) Supervised descriptive rule discovery: a unifying survey of contrast set, emerging pattern and subgroup mining. J Mach Learn Res 10:377–403.

[7] X. Liu, J. Wu, F. Gu, J. Wang, and Z. He, “Discriminative pattern mining and its applications in bioinformatics,” Briefings Bioinform., vol. 16, no. 5, pp. 884–900, 2015