XL CILAMCE
IBERO-LATIN AMERICAN CONGRESS ON COMPUTATIONAL METHODS IN ENGINEERING
NOVEMBER 11-14, 2019
Praiamar Natal Hotel & Convention
Natal, RN-BRAZIL

MINERAÇÃO E ANÁLISE EM BASE DE DADOS EDUCACIONAIS: UMA ABORDAGEM UTILIZANDO CLUSTERIZAÇÃO NOS DADOS DA PLATAFORMA LATTES

Guilherme Santos da Silveira
g.silveira.eng@gmail.com
Engenharia da Computação – Universidade do Estado de Minas Gerais(UEMG)
Avenida Paraná, 3001, Jardim Belvedere, 35501-170, Divinópolis, Minas Gerais, Brasil
Tiago Alves de Oliveira
Thiago Magela Rodrigues Dias
tiagofga@gmail.com
thiagomagela@gmail.com
Centro Federal de Educação Tecnológica de Minas Gerais (CEFET-MG)
Av. Amazonas, 7675 – Nova Gameleira – CEP: 30510-000 – Belo Horizonte – MG – Brasil

Abstract. Os recentes desenvolvimentos tecnológicos permitiram a geração de grandes volumes de dados que devido à sua complexidade, necessitam de auxilio computacional para gerar resultados precisos em um curto período de tempo. A Plataforma Lattes é um conjunto de sistemas web contendo bases de currículos, grupos de pesquisa e instituições, focados no campo da Ciência e Tecnologia e mantidos pelo CNPq. Devido a sua rica base de dados possui informações importantes no campo da Ciência e Tecnologia. Neste artigo, é demonstrada a viabilidade da aplicação dos métodos de mineração de dados no banco de dados da Plataforma Lattes externado seu conhecimento. Assim, como principais contribuições deste artigo são apresentadas formas de interpretação, desenvolvimento de padrões e geração de conhecimento a partir dessa base de dados onde foram aplicadas técnicas de clusterização.

Abstract. Recent technological developments have enabled the generation of large volumes of data that due to their complexity, require computational assistance to generate accurate results in a short period of time. The Lattes Platform is a set of Web systems containing databases of curricula, research groups and institutions, focused on the field of Science and Technology and maintained by CNPq. Due to its rich database it has important information in the field of Science & Technology. In this article, the feasibility of applying the data mining methods in the Lattes Platform database is demonstrated externalizing its knowledge. Thus, as main contributions of this article forms of interpretation, development of standards and generation of knowledge are presented from this data base, where clustering techniques were applied.

Keywords: Análise de dados, K-Means, Clusterização, Mineração de dados.

1 Introdução

O volume de dados gerado atualmente por pessoas e organizações dos mais variados tipos são de tamanhos gigantescos devido principalmente a evolução tecnológica e as facilidades que essa evolução trouxe.

As largas utilizações de diversos sistemas dão origem a grandes bases de dados, que por sua vez são complexas e de grandes volumes sem organização. Nesse contexto a mineração de dados auxilia na análise das informações contidas nessas bases. Contudo sua aplicação é algo complexo pois se realiza na busca e preparação dessas quantidades de dados, na aplicação de complexos algoritmos de inteligência artificial e a análise e interpretação corretas desses resultados.

Entre as diversas técnicas aplicadas na extração e tratamento de dados estão o processo chamado de Descoberta de Conhecimento em Banco de Dados do inglês KDD *(Knowledge Discovery in Database)*, que segundo seus próprios criadores [1] é o processo não trivial de extração de informações implícitas, previamente desconhecidas e potencialmente úteis a partir dos dados armazenados em um banco de dados. E também o algoritmo de classificação KNN *(K-Nearest Neighbor)*, que é um algoritmo que armazena todos os casos disponíveis e classifica novos casos com base em uma medida de similaridade.

Apesar de algoritmos serem capazes de descobrirem padrões valido não existe ainda uma solução eficiente para localizar padrões potencialmente úteis. Nesse caso, na mineração de dados ainda se faz necessário uma forte interação com analista humanos que são responsáveis principalmente por direcionar e determinar as direções que as explorações dos dados devem seguir.

A respeito de Ciência e Tecnologia (C&T),[2] afirmam que, ciência e tecnologia são elementos de transformação cruciais para uma nação, quando se trata de elevar o padrão de vida da população, consolidar uma economia moderna e participar com plenitude em um mundo cada vez mais globalizado. Portanto, para que uma nação consiga atingir adequados níveis de desenvolvimento tecnológico, social e cultural deve-se possibilitar as pessoas e as organizações competirem e interagirem nesse ambiente desenvolto e ágil considerando também a importância da geração de ciência para o desenvolvimento de um país. Nesse contexto, [3] disse que é fundamental que o conhecimento, os métodos e as descobertas em todas as áreas da ciência estejam disponíveis a todos na mais ampla escala.

A Plataforma Lattes, uma iniciativa do governo federal através do Ministério da Ciência e Tecnologia através do CNPq, constitui-se atualmente em um grande acervo de informações sobre os pesquisadores e sua produção científica e tecnológica. Através do Currículo Lattes, do formulário eletrônico do Ministério da Ciência e Tecnologia (MTC), do Conselho Nacional de Desenvolvimento Cientifico e Tecnológico (CNPq), da Financiadora de Estudos e Projetos do MTC (FINEP) e da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), sobre os grupos de pesquisa existentes nas universidades e nas empresas e outras instituições ligadas à pesquisa científica.

O trabalho consiste na utilização de técnicas de mineração de dados e clusterização, na identificação e extração de conhecimento possibilitando ações comparativas das principais características fornecidas pela base de doutores na Plataforma Lattes.

Este artigo é divido em 5 seções. Na seção 2 serão abordados os materiais e métodos utilizado no desenvolvimento do artigo. A seção 3 é composta pelas características dos dados, a seção 4 apresenta os resultados encontrados, e a seção 5 apresenta a conclusão.

2 Desenvolvimento

Para a realização dos objetivos o trabalho foi fundamentado em quatro etapas, primeiro o estudo da base de dados Plataforma Lattes, seguido do estudo a respeito da teoria de mineração de dados e das principais técnicas e métodos aplicados nesta tecnologia. Logo após foi realizado o tratamento dos dados e aplicação das técnicas e métodos de clusterização. E finalizado com a análise dos resultados obtidos. Na primeira etapa foi realizado o estudo da base de dados da Plataforma Lattes e realizado o levantamento dos dados relevantes ao trabalho como a instituição de ensino, o nível de formação, a área de formação, área de atuação, o número de artigos publicados de cada indivíduo, e definido então o

público alvo do trabalho.

A segunda etapa constituiu-se do estudo da área de mineração de dados como a inteligência artificial, as técnicas matemáticas e regras de associação, além dos diversos métodos, técnicas e ferramentas utilizadas que melhor se encaixaram no trabalho.

Na terceira etapa foram realizados o tratamento dos dados tendo em vista que os dados obtidos se encontravam ainda de forma desordenada e fora dos padrões necessários. Então foi implementado técnicas e algoritmos estudados e definidos na etapa anterior.

A quarta etapa constitui-se da apresentação dos resultados e das informações obtidas através das diversas técnicas de clusterização utilizadas, demonstrando as relações encontradas e os dados mais relevantes. A Figura 1 apresenta de forma visual as etapas de todo o processo de descoberta do conhecimento as quais foram aplicadas no desenvolvimento deste artigo.

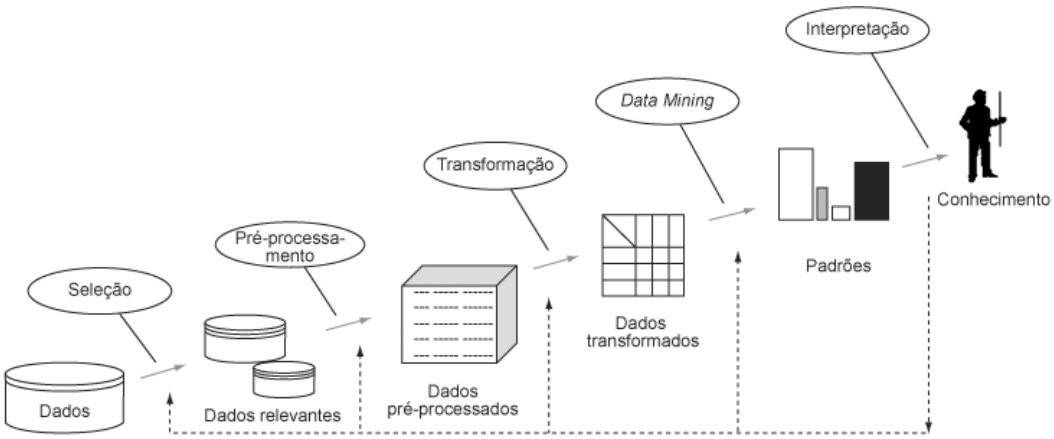

Figura 1. Etapas do Processo de KDD. Fonte: [1].

A mineração de dados (do inglês *data mining*) de acordo com [1] seria a etapa de aplicação de algoritmos específicos para extrair padrões dos dados. Já [4] dividem a mineração de dados em quatro tarefas: Modelos preditivos (classificação e regressão), Associação, Agrupamento e Detecção de anomalias.

Modelos preditivos dispõem-se a construção de um modelo para uma variável-alvo que consiga identificar padrões ocultos e prever o que poderá ocorrer. Associação é usada para descobrir padrões associados à características presentes nos dados, ou seja, elementos comuns em um determinado conjunto de dados. Agrupamento refere-se a tarefa inserir objetos dispersos em grupos e procurar características que façam os itens de um grupo serem mais similares em relação com as características apresentadas por outros grupos. Detecção de Anomalias procura identificar elementos e padrões que são significativamente diferentes ou apresentam comportamento inesperado do resto apresentado nos dados.

3 Caracterização do Repositório de Dados

Aqui é apresentado a caracterização do repositório de dados utilizado composta por doutores, pós-doutores e livre docentes contidos na Plataforma Lattes e publicados até o mês de junho de 2017. Livre docentes integram o quadro pois em seu conceito é o mais elevado estágio de formação universitária atingido por competência, podendo assim gerir seus projetos sem auxílio de um orientador. Como já mencionado a plataforma lattes conta com milhões de currículos cadastrados e em constante crescente impulsionados por agências de fomento e órgãos governamentais. A base de dados bem como todo o processo e extração dos dados foi realizado utilizando o extrator apresentado por [5].

A base era composta por vários arquivos CSV *(Comma Separated Values)*, inicialmente contendo 266.187 indivíduos separados por linhas e um total de 92 colunas indicados por diversos elementos como a titulação, áreas de atuação, proficiência, início, termino e local de doutorado. Porem todas essas colunas eram apresentadas de forma desordenadas, alguns campos compostos por textos dados e valores nulos.

Levando também em consideração que muitos dados são inseridos pelos próprios indivíduos e nem todos os currículos possuem informações padronizadas, porém é importante ressaltar que não é possível um indivíduo possuir mais de um currículo cadastrado assim excluindo a possibilidade de dados duplicados de um mesmo indivíduo na plataforma. A Tabela 1 apresenta um resumo dos dados iniciais recebido antes dos tratamentos realizados.

Tabela 1. Resumo dos principais dados da Plataforma Lattes (06/2017)

Características	Valores
Livre-Docentes	151 (0.056%)
Doutores	191.300 (72.14%)
Total de indivíduos	266.187 (100%)
Orientações Concluídas	5.872.924 (100%)
Anais em Congresso	9.051.257 (100%)
Texto em Jornais e Revistas	858.608 (100%)
Artigos em Periódicos	4.660.052 (100%)

Porém para o uso do algoritmo de clusterização foi necessário a realização da padronização dos dados e posteriormente a transformação de todos os dados de cada campo em valores inteiros além da exclusão dos campos vazios.

De acordo com o estudo da base e visando informações mais claras foi realizado uma redução da base filtrando e descartando indivíduos que não possuía nenhuma publicação cientifica, orientação em andamento e concluídas e não se encaixava em nenhuma área e/ou grande área, pois essas descaracterizações também podem vir a representar currículos desatualizados e não completos.

Ao final foi criado uma base que veio ser utilizada na criação dos *cluster*, esta base possui 236.598 indivíduos e for reduzida a 7 colunas sendo todos os campos transformadas em valores inteiros onde a:

- Coluna A apresenta a titulação sendo 1 Livre-Docentes, 2 Doutores e 3 Pós-Doutores;
- Coluna B apresenta a Grande Área de Atuação dividida de 1 a 8 sendo as grandes áreas padrões, 9 sendo outras grandes áreas e 0 os campos vazios;
- Coluna C apresenta de 1 a 99 as Área de Atuação e o 100 sendo os campos vazios;
- Coluna D apresenta o Total das Orientações em Andamento;
- Coluna E apresenta o Total das Orientações concluídas;
- Coluna F apresenta a soma das Publicações em anais de congresso, periódicos e jornais e revistas;
- Coluna G apresenta de 1 a 27 os estados do Brasil, e o 0 como outros países ou campos vazios;

Com base nos dados obtidos foi possível observar uma média bem próxima referente ao número de orientações em andamento em todos os estados do Brasil com uma média geral de 2,8 orientações em andamento. O estado de Alagoas apresenta a maior média com 3,8, seguido pelo estado de Goiás com 3,6 e Bahia com 3,4. O Distrito federal apresentou o menor índice com média de 2,1 orientações em andamento, São Paulo e Tocantins com 2,5 e 2,4 respectivamente.

É importante observar também a média de publicações por estados, Rio Grande do Sul possui a maior média entre todos os estados brasileiros apresentando um índice com média de 77,1 publicações científicas, seguido por São Paulo com 75,6. Ceará com 67,9, Paraná com 67,5 e Minas Gerais com 66,1. O menor índice válido apresentado pelos estados brasileiros veio do Acre com média de 41,5 e Amapá com 42,0.

A Tabela 2 apresenta outro fator importante, a distribuição geográfica dos indivíduos pelas regiões do Brasil. A região Sudeste concentra 100.557 indivíduos presentes na base, o Sul 37.234, a região do

Nordeste 33.012, o Centro-Oeste 15.075, e a região Norte apenas 7.827 indivíduos. A alta concentração no Sul e Sudeste e baixa concentração de indivíduos no Norte do país tem influência nos índices e por isso são importantes fatores estatísticos.

Tabela 2. Média de Orientações e Publicações por Região do Brasil

Estado	Ori. Andamento	Ori. Concluídas	Public. Científicas	Numero de indivíduos
Norte	3,0	24,7	48,8	7.827
Nordeste	3,2	27,3	58,3	33.012
Centro-Oeste	2,7	26,2	54,1	15.075
Sudeste	2,6	25,7	68,1	100.557
Sul	2,9	31,2	70,9	37.234

3.1 K-means

Segundo [6], o algoritmo k-means leva o parâmetro de entrada k, e as partições um conjunto de n objetos em k clusters de modo que a similaridade intracluster resultante é alta, mas a semelhança do intercluster é baixa. A similaridade do cluster é medida em relação ao valor médio dos objeto em um cluster que pode ser visto como cluster ou centro de gravidade do cluster. Para [7], normalmente referido simplesmente como k-means, o algoritmo de Lloyd começa com k "centros"arbitrários, tipicamente escolhidos uniformemente e aleatoriamente dos pontos de dados.

Então cada ponto é atribuído ao centro mais próximo e cada centro é recalculado como o centro de massa de todos os pontos atribuídos a ele. Ou seja, ele seleciona aleatoriamente k objetos, onde cada um destes objetos inicialmente representam um meio ou um centro do cluster. Para cada um dos objetos restantes um objeto é atribuído ao cluster ao qual é o mais parecido com base na distância entre o objeto e o cluster, em seguida, calcula a nova média para cada cluster. É a velocidade e a simplicidade do método k-means que o tornam atraente, e não a sua precisão.

Em outras palavras, o algoritmo funciona de forma simples e rápida, é um algoritmo amplamente utilizado classificando de forma automática sem a necessidade de supervisão humana, eficiente em tratar grandes conjuntos de dados. A Figura 2 a seguir demonstra de forma visual o funcionamento do algoritmo.

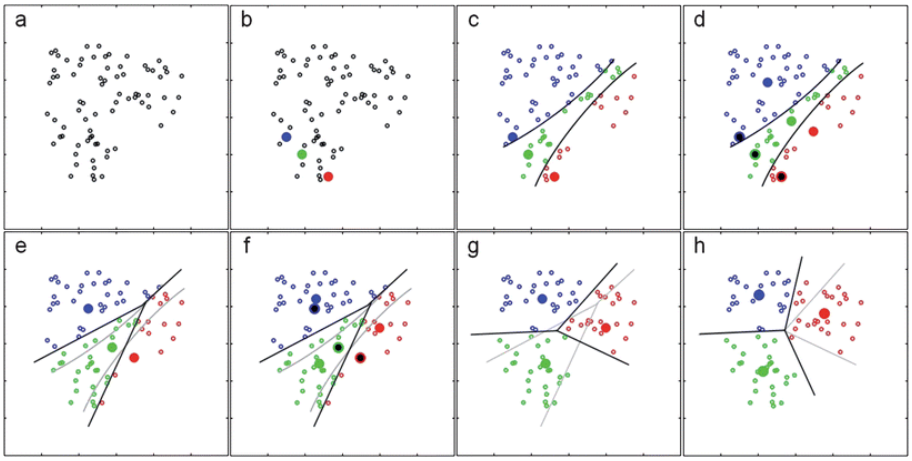

Figura 2. Exemplo de clusterização utilizando o K-Means. Fonte: Elaborado pelo autor.

O desenvolvimento utilizando os métodos do k-means se deu através da utilização da biblioteca

Scikit-learn[1] de aprendizado de máquina *open-source* para a linguagem de programação Python, a biblioteca interage com outras bibliotecas de Python, a biblioteca numérica NumPy e a cientifica SciPy, além de outras, como a biblioteca de plotagem gráfica matplotlib e a biblioteca pandas responsável por manipulação e análise de dados, todas bibliotecas escritas na linguagem de programação Python.

Inicialmente foram utilizadas as bibliotecas Numpy, SciPy e pandas que permitem a conversão e manipulação da base de dados em arquivo CSV *(Comma Separated Values)* para *arrays* e matrizes parâmetros esse que são necessários para a utilização do k-means. Então logo após a conversão dos dados o algoritmo k-means realiza os treinamentos sobre os dados da base, assim encontrando os centros e criando os clusters, e com base nos resultados do algoritmo a biblioteca de plotagem matplotlib gera o gráfico conforme os parâmetros solicitados.

Como explicado anteriormente, cada parâmetro foi separado em colunas de valores que variam de 0 até 6 passando o valor referente a coluna escolhida sendo no máximo três colunas por vez, então informando também no código a quantidade de cluster solicitado, por fim executando o código gerando figuras como os resultados apresentados a seguir.

4 Resultados

Com o uso de métodos de cluterização sobre a base de doutores da Plataforma Lattes é possível observar importantes características de difícil identificação que são demonstradas através de gráficos facilitando e simplificando o processo de identificação de padrões. Padrões e características estas que podem ser de grande importância vindo a fomentar pesquisas relacionadas a base estudada.

Os resultados são apresentados em forma de gráficos apresentando os clusters de acordo com características determinadas considerando sempre os indivíduos em sua totalidade. Através do uso de k-means é possível observar interessantes relações como a distribuição entre número de orientações concluídas por estados e por grande área de atuação como apresentado nas Figuras 3 e 4 , regiões como a região Norte apresenta índices bem próximos entre seus estados. O gráfico também apresenta um equilíbrio em relação as grandes áreas de atuação por seus respectivos estados.

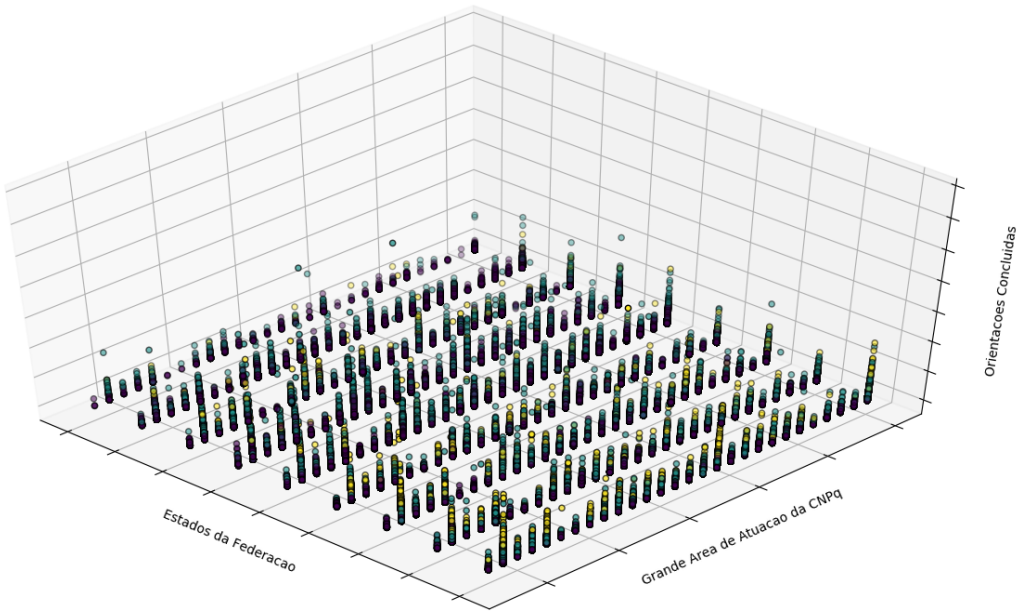

Figura 3. Relação entre Estados, Orientações Concluídas e Grande Área de Atuação. Fonte: Elaborado pelo autor.

Importante notar na Figura 4 uma diferença considerável de orientações concluídas existentes entre

[1]Disponível em http://scikit-learn.org/

alguns estados como Acre na posição 1, Minas Gerais na posição 11, São Paulo na posição 26 e Tocantins na posição 27. Um importante fator que deve ser levado em consideração é o fato de estados da região Norte do Brasil possuírem um número reduzido de indivíduos presentes na base de dados em relação a regiões como o Sul e o Sudeste. E com menor número de individuo é natural que venha a apresentar menores índices em comparação as outras regiões. Também é possivel notar em todos os *cluster* de todos os estádos brasileiros a grande concentração que se dá na marge de até 0 à 200 orientações, importante apontar a presença de *Outlier* apresentando cada indivíduo mais de 1000 orientações concluídas cada.

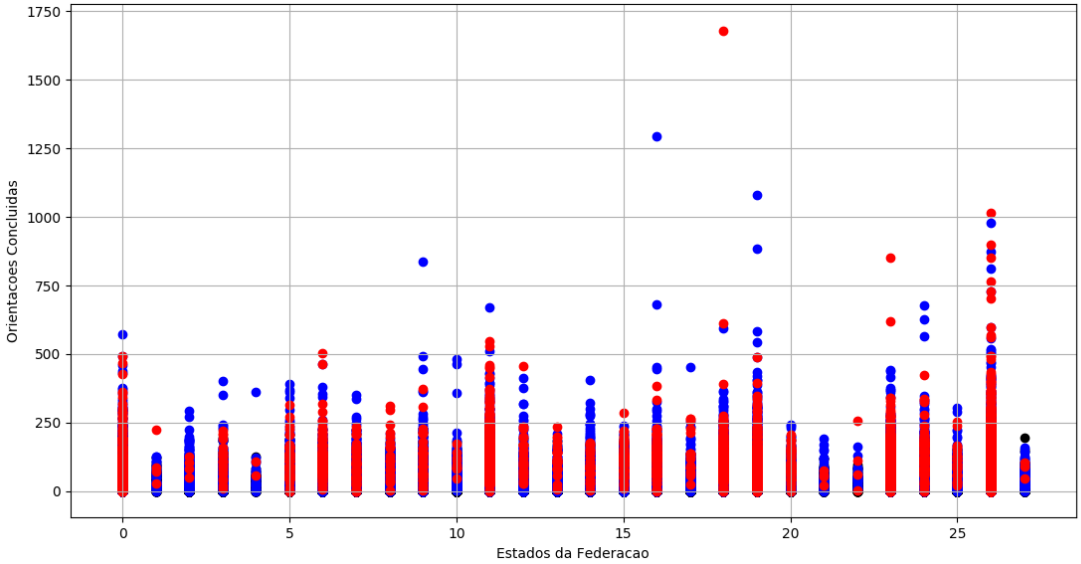

Figura 4. Relação entre Estados com Orientações Concluídas. Fonte: Elaborado pelo autor.

A Figura 5 apresenta um equilíbrio no número de Publicações Cientificas dividido entre os estados com exceção de Minas Gerais na posição 11, Rio de Janeiro na posição 19 e São Paulo na posição 26 com maiores volumes de publicações, é natural estados da região Sudeste apresentarem valores elevados em termos de números tendo em vista que é a região do país onde se concentra o maior numero de indivíduos conforme dados apresentados anteriormente na tabela 2 . Assim também como Amapá na posição 4 e Rondônia na posição 21 apresentam níveis mais baixos em valores totais em relação aos demais estados brasileiros, poís em numeros totais a região Norte apresenta o menor número de indivíduos dos dados utilizados das regiões do Brasil. É possível notar é a presença de um indivíduo *Outlier* com mais de 11.000 publicações científicas no estado de São Paulo, o que se pode relatar é o fato já dito antes de que não existe um controle de adição de informações na Plataforma Lattes, e tendo em vista que os outros dados extraidos estão dentro de valores padrões, por tanto não é possível dizer precisamente qual seria o verdadeiro número de publicações válidas deste indivíduo.

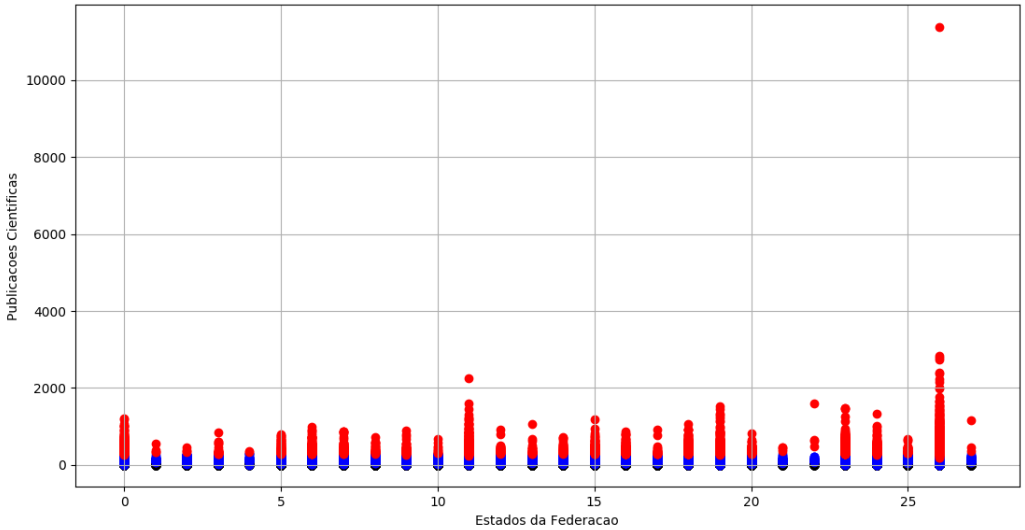

Figura 5. Relação entre Publicações e Estados da Federação. Fonte: Elaborado pelo autor.

Ja ná Figura 6 demostra bem a deficiência em números totais de Orientações Conluíadas por algumas Áreas de Atuação. Áreas como Ciências Humanas e Ciências Sociais, Engenharia Cartográfica, Engenharia de Agrimensura e Estudos Sociais. Algumas destas áreas estão presentes na plataforma Lattes porem ja são áreas consideradas descontinuadas, áreas como Carreira Militar na posição 15 e Ciências Atuarias na posição 22 são exemplo disso e explicam o nível nulo de informações. Novamente é possível notar a presença dos *outlier* como demonstrado antériormente com mais de 1000 orientações cada.

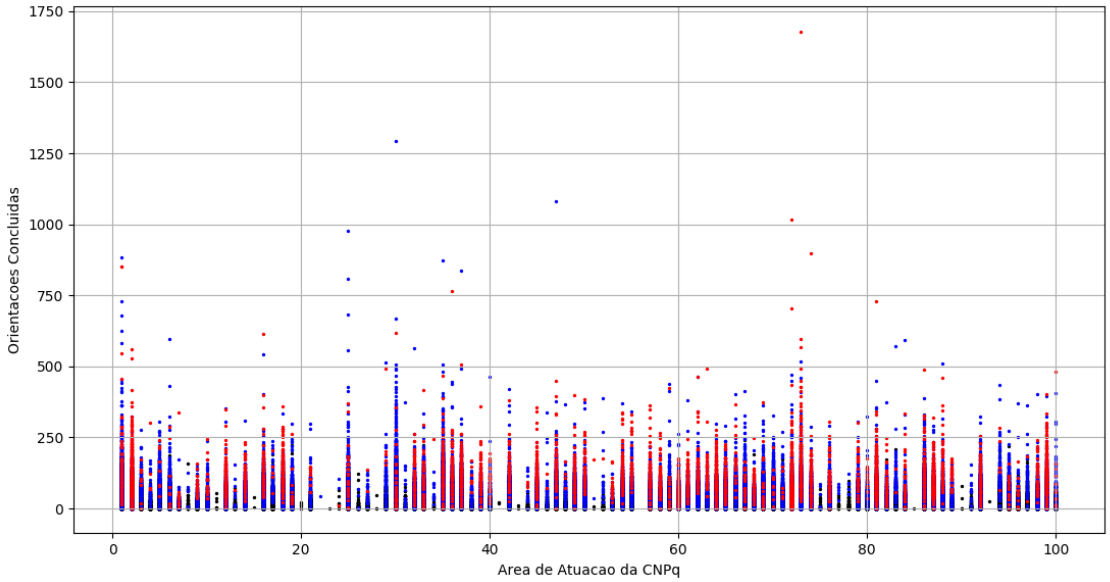

Figura 6. Relação entre Orientações Concluídas e Área de Atuação. Fonte: Elaborado pelo autor.

As Figuras 7 e 8 repetem os padrões apresentados anteriormente pela figuras 6, porém desta vez demonstrando as Orientações em Andamento e não mais a Concluídas. As mesmas questões podem ser levantadas no que diz a respeito de áreas descontinuadas, outro ponto que se pode considerar é o fato de que alguns cursos são cursos mais específicos como Engenharia Cartográfica na posição 41, Programação Fetal na posição 85 e não são oferecidos amplamente como Direito na posição 30 e Letras na posição 69.

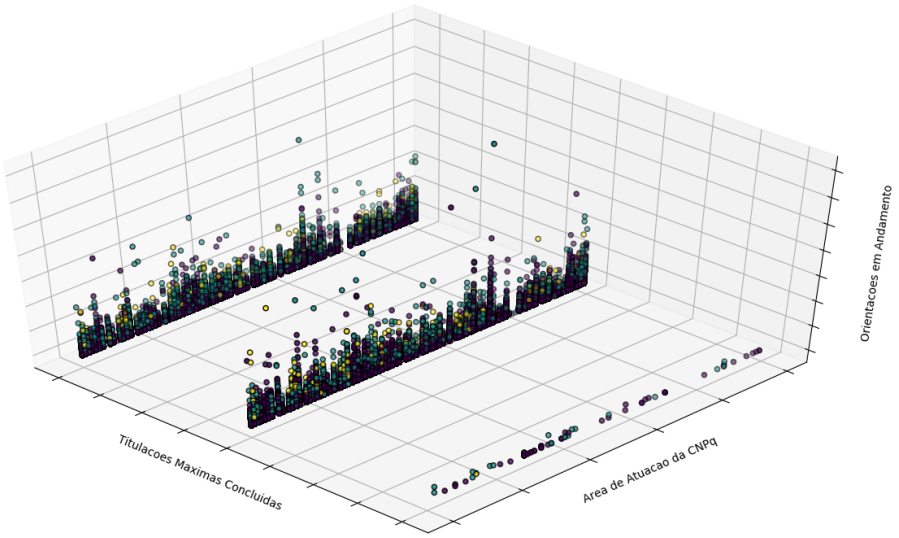

Figura 7. Relação entre Titulação Máxima Concluídas, Área de Atuação e Orientações em andamento. Fonte: Elaborado pelo autor.

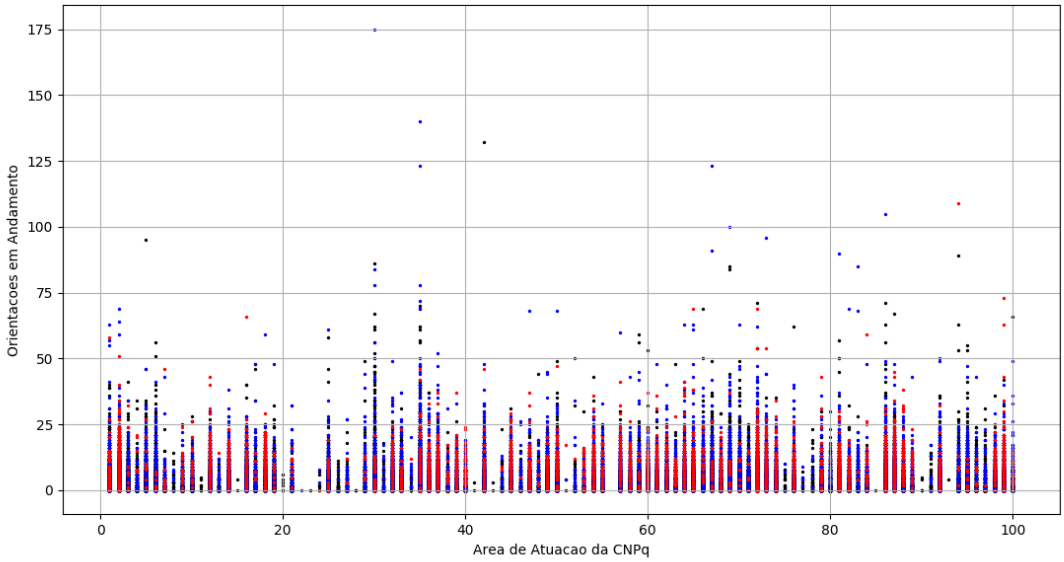

Figura 8. Relação entre Orientações em Andamento e Áreas de atuação. Fonte: Elaborado pelo autor.

A Figura 9 apresenta a relação das Orientações Concluídas com o número de Publicações Cientificas separado pela titulação mostrando um maior número de concentração em Doutores e Pós-Doutores, já a Figura 10 apresenta uma alta concentração justamente na menor quantidade de Publicações Cientificas em relação as Orientações Concluídas. Mesmo após os indivíduos com zero publicações serem removidos exite uma grande concentração nos índices mais baixos. Este gráfico apresenta uma visão geral dos baixos índices de publicações no Brasil se tratando de Doutores salvo alguns indivíduos isoladamente. Novamente é possível notar a presença nos extremos de indivíduos com índices muito elevados em relação aos demais, como já explicado não é ha um controle e nem como validar os reais dados nesta situação destes indivíduos. Em ambas as figuras é importante salientar a presença dos *outliers* ja sinalisados anteriormente.

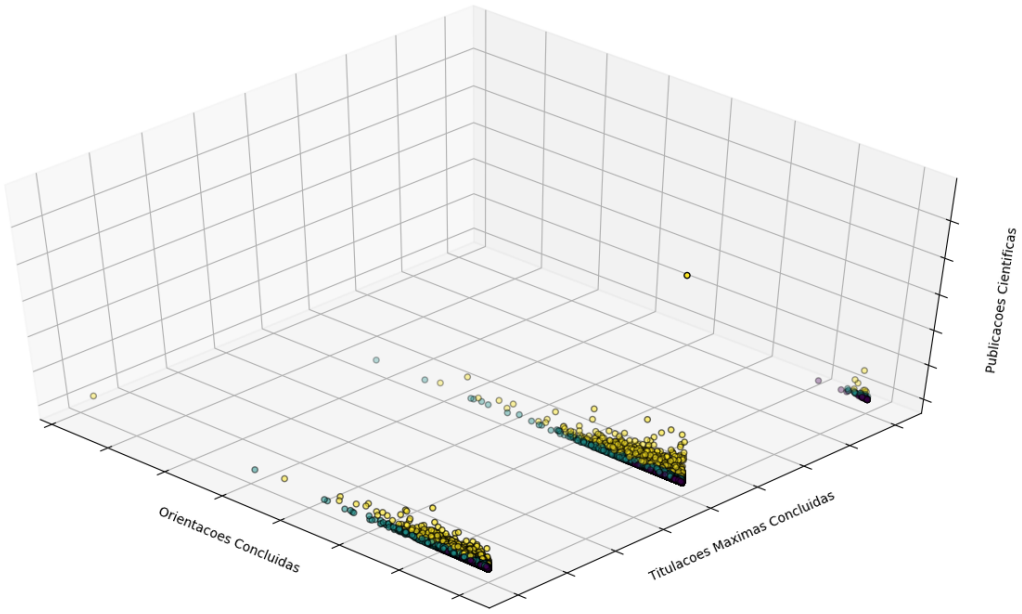

Figura 9. Relação entre Orientações Concluídas, Titulação Máxima e Publicações Científicas. Fonte: Elaborado pelo autor.

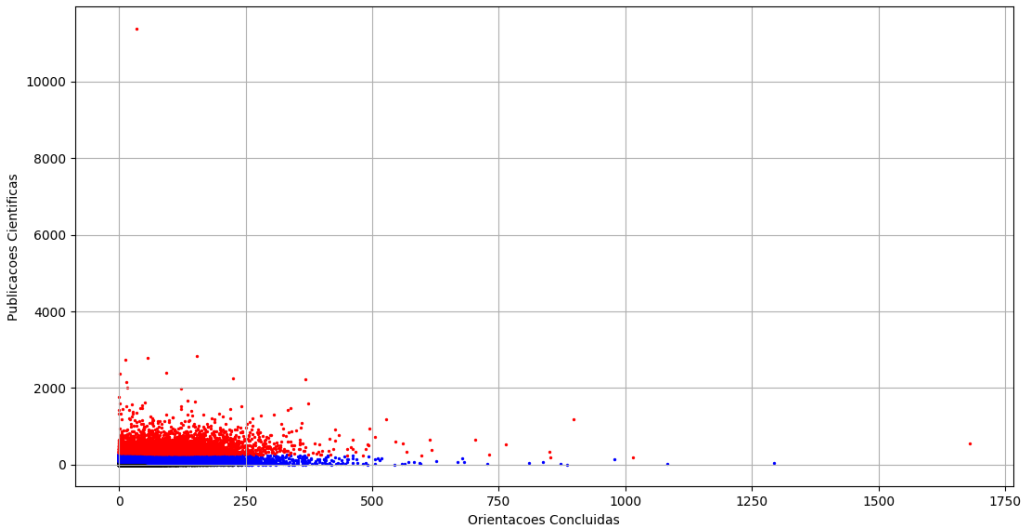

Figura 10. Relação entre Publicações Científicas e Orientações Concluídas. Fonte: Elaborado pelo autor.

Uma vez que o processo tem sido utilizado com sucesso apresentando resultados satisfatórios os dados obtidos são de grande importância tendo em vista também a escassez de informações confiáveis relacionados a dados educacionais.

5 Conclusão

A extração e análise de grande base de dados demonstra ser uma importante ferramenta de auxílio e apresentação de dados complexos, a utilização de métodos de clusterização demonstrou resultados de forma simples com importantes relações presente nos dados da Plataforma Lattes. Foi possível identificar a relação científica no que tange as Publicações e Orientações por região do Brasil demonstrando a defasagem de algumas Áreas que podem ser mais bem desenvolvidas, e também um baixo índice no geral de publicações, mesmo depois de filtrados os indivíduos com publicações inexistentes. A presença de *outliers* é vista como normal, tendo em vista que a Plataforma Lattes é aberta e não apresenta formas de restringir ou validar todos os dados ali depositados. O trabalho demonstra também uma possível escalabilidade na base de dados, podendo apresentar uma análise de toda a base da Plataforma Lattes contribuindo ainda mais com diversas áreas de pesquisas.

References

[1] FAYYAD, Usama; PIATETSKY-SHAPIRO, G. S. P., 1996. *From data mining to knowledge discovery: An overview. In: Advances in Knowledge Discovery and Data Mining*. AAAI Press / The MIT Press, MIT, Cambridge, Massachusetts, and London, England.

[2] SCHWARTZMAN, Simon; KRIEGER, E. G. F. G. E. A. B. C. O., 1992. Ciência e tecnologia no brasil: Uma nova política para um mundo global. in: Ciência e tecnologia no brasil: Política industrial, mercado de trabalho e instituições de apoio. *Rio de Janeiro: Editora da Fundação Getúlio Vargas*, vol. v. 2.

[3] Sagan, C., 2006. *O mundo assombrado por demônios. A ciência vista como uma vela na escuridão.* São Paulo: Editora Cia das Letras.

[4] STEINBACH, M.; TAN, P. K. V., 2006. *Introduction to Data Mining*. Pearson Education.

[5] DIAS, T. M. R. e. a., 2013. Modelagem e caracterização de redes científicas: Um estudo sobre a plataforma lattes. *In: Brazilian Workshop on Social Network Analysis and Mining.*

[6] HAN, J.; KAMBER, M., 2006. *Data Mining: Concepts and Techniques*. Elsevier.

[7] Arthur, D. & Vassilvitskii, S., 2007. Proceedings of the eighteenth annual acm-siam symposium on discrete algorithms. *Society for Industrial and Applied Mathematics*.