UNIVERSIDADE NOVE DE JULHO - UNINOVE
PROGRAMA DE PÓS GRADUAÇÃO EM INFORMÁTICA E GESTÃO DO CONHECIMENTO - PPGIGC

SAULO DANIEL DOS SANTOS

MÉTODO DE AGRUPAMENTO HIERÁRQUICO A PARTIR DE CURRÍCULOS ACADÊMICOS

São Paulo
2017

SAULO DANIEL DOS SANTOS

MÉTODO DE AGRUPAMENTO HIERÁRQUICO A PARTIR DE CURRÍCULOS ACADÊMICOS

Dissertação apresentada ao Programa de Pós-Graduação em Informática e Gestão do Conhecimento da Universidade Nove de Julho - UNINOVE, como parte dos requisitos para a obtenção do título de Mestre em Informática e Gestão do Conhecimento.

Prof. Orientador: Dr. Wonder Alexandre Luz Alves

São Paulo
2017

Santos, Saulo Daniel dos.
Método de agrupamento hierárquico a partir de currículos acadêmicos. / Saulo Daniel dos Santos.2017.
76 f.
Dissertação (Mestrado) - Universidade Nove de Julho - UNINOVE, São Paulo, 2017.
Orientador (a): Prof. Dr. Wonder Alexandre Luz Alves.

1. Plataforma Lattes. 2. ScriptLattes. 3. MultiScriptLattes. 4. ScriptComp. 5. Agrupamento hierárquico.
I. Alves, Wonder Alexandre Luz. II. Titulo.

CDU 004

UNINOVE

Universidade Nove de Julho

PARECER – EXAME DE DEFESA

Parecer da Comissão Examinadora designada para o exame de defesa do Programa de Pós-Graduação em Informática e Gestão do Conhecimento a qual se submeteu o aluno **Saulo Daniel dos Santos**

Tendo examinado o trabalho apresentado pelo aluno **Saulo Daniel dos Santos**, aluno regularmente matriculado no Programa de Pós-Graduação em Informática e Gestão do Conhecimento da Universidade Nove de Julho - UNINOVE, para obtenção do título de "Mestre em Informática e Gestão do Conhecimento", com dissertação intitulada ": Análise de agrupamento hierárquico a partir de currículos acadêmicos ", e após ter ouvido a exposição do candidato, a Comissão o considerou:

(X) **Aprovado** ( ) **Aprovado condicionalmente**

( ) **Reprovado com direito a novo exame** ( ) **Reprovado**

EXAMINADORES

Prof. Dr. Wonder Alexandre Luz Alves

Prof. Dr. Jesús Pascual Mena Chalco

Prof. Dr. Renato Ribeiro Nogueira Ferraz

São Paulo, 04 de dezembro de 2017.

Dedicatória

Dedico esse trabalho a minha família,
que sempre esteve do meu lado me incentivando!

Agradecimentos

Primeiramente gostaria de agradecer a **DEUS**, por me proporcionar essa benção. Graças ao SR. que cheguei até aqui, e se não fosse tuas mãos jamais conseguiria. Muito Obrigado meu **DEUS**.

Também quero agradecer ao Prof e amigo Dr. Wonder A. L. Alvez, por acreditar em mim, e me incentivar desde o início de minha carreira, tanto profissional como acadêmica. Muito Obrigado agradeço a Deus pela vida do Sr.

Minha Esposa Patricia M. S. Santos, não tenho palavras para agradecer o quanto você foi importante em minha vida! obrigado por tudo!!! nos momentos que pensei em desistir, você sempre esteve do meu lado me ajudando, aconselhando e me confortando. Te amo!

Aos meu pais Theresa J. Mendonça e Deurival dos Santos, que sempre me apoiaram e me incentivaram, agradeço todos os ensinamentos e também todo amor. Amo vocês!!!

Também agradeço a todos os Professores do PPGI, que sempre estiveram prontos a me instruir e aconselhar. Muito obrigado por tudo, levarei por toda minha vida tudo que aprendi com vocês.

Aos meus amigos Charles, Cris e Kauê, que sempre estiveram ao meu lado me ensinando e me apoiando. Vocês são sem palavras!

E se o meu povo, que se chama pelo meu nome, se humilhar, e orar, e buscar a minha face e se converter dos seus maus caminhos, então eu ouvirei dos céus, e perdoarei os seus pecados, e sararei a sua terra .

2 Crônicas 7:14

Porque Deus amou o mundo de tal maneira que deu o seu Filho unigênito, para que todo aquele que nele crê não pereça, mas tenha a vida eterna.

João 3:16

Resumo

A plataforma Lattes passou a ser utilizada pelas universidades, agências de fomentos e grupos de pesquisa, como fonte principal de informações sobre o histórico acadêmico dos pesquisadores, influenciando na análise curricular. Devido a esse fato, é praticamente obrigatório no Brasil que todos os acadêmicos mantenham seus currículos devidamente atualizados nesta plataforma. Diante disso, muitos pesquisadores exploram essa base de currículos para descobrir novos conhecimentos sobre a construção da ciência no Brasil. Nesse sentido, a ferramenta ScriptLattes tem sido amplamente utilizada em todo o território nacional por uma gama de estudiosos, para extrair informações a partir dos currículos cadastrados na plataforma Lattes. Os resultados obtidos até então têm sido de grande valia para a extração de conhecimento desta base de currículos. Diante do exposto, esse trabalho faz o uso do ScriptLattes para extrair informações de currículos da plataforma Lattes. Assim, são aplicados conceitos de agrupamento hierárquico para construir um método para computação de hierarquias de agrupamentos. Com base nesse método proposto, foram construídas aplicações para: (*i*) produzir relatórios que auxiliam os preenchimentos das plataformas Sucupira e e-Mec; (*ii*) visualizar similaridades entre programas de pós-graduações por meio de indicadores de área.

Palavras-chave: Plataforma Lattes; Agrupamento Hierárquico; ScriptLattes; MultiScriptLattes; Agrupamento de dados.

Abstract

The Lattes platform has been used by universities, development agencies and research groups as the main source of information on researchers' academic history, influencing curriculum analysis. Due to this fact, it is practically obligatory in Brazil that all researchers keep their curricula duly updated in this platform. Faced with this, many researchers explore this basis of curricula to discover new knowledge about the construction of science in Brazil.In this sense, the ScriptLattes tool has been widely used throughout the national territory by a range of researchers, to extract information from the curricula registered on the Lattes platform. The results obtained so far have been of great value for the extraction of knowledge from this basis of curricula. Based on the above, this work makes use of ScriptLattes to extract curriculum information from the Lattes platform. Thus, hierarchical clustering concepts are applied to construct a method for constructing group hierarchies. Based on this method, applications were built to:(i) produce reports that help fill the Sucupira and e-Mec platforms; (ii) to visualize similarities among postgraduate programs through area indicators.

Keywords: Lattes Platform; Hierarchical Clustering; ScriptLattes; MultiScriptLattes; Grouping of data.

Sumário

Lista de Figuras

Lista de Tabelas

Lista de Abreviaturas

CAPES	Coordenação de Aperfeiçoamento de Pessoal de Nível Superior
MEC	Ministério da Educação e Cultura
CNPQ	Conselho Nacional de Desenvolvimento Científico e Tecnológico
URL	*Uniform Resource Locator*
DAML	*DARPA agent markup language*
OIL	*Ontology Inference Layer*
XML	*eXtensible Markup Language*
HTML	*HyperText Markup Language*
DOI	*Digital Object Identifier*
WEB	*World Wide Web*
CEP	Código de Endereço Postal
LIST	Extenção de arquivo de texto
CONFIG	Extenção de arquivo de texto
CSV	*Comma separated values*

Símbolos

$\mathcal{T}$	Representação de um agrupamento hierárquico
Cv	Lista de currículos extraídos da plataforma Lattes
$\mathbb{P}$	Um agrupamento de dados
$\mathbb{R}$	Conjunto dos números reais
$\emptyset$	Conjunto vazio
$\in$	É um elemento de
$\notin$	Não é um elemento de
$\cup$	União entre conjuntos
$\subseteq$	É um subconjunto de
$\subset$	É um subconjunto próprio de
$\min\{x : \star\}$	É o menor elemento do conjunto $\{x : \star\}$
$\max\{x : \star\}$	É o maior elemento do conjunto $\{x : \star\}$
$\underset{x \in X}{\operatorname{argmax}}\{f(x)\}$	Argumento $x \in X$ que maximiza $f(x)$
$\underset{x \in X}{\operatorname{argmin}}\{f(x)\}$	Argumento $x \in X$ que minimiza $f(x)$
$\forall$	Quantificador lógico universal

Capítulo 1

Introdução

1.1 Contextualização do tema

Com o crescimento da comunidade científica nas últimas décadas, novos problemas surgiram para quantificar a produção, qualidade e inovação dos resultados de determinadas pesquisas (LANE, 2010). Ao longo do tempo várias ações vêm sendo idealizadas pela comunidade científica para preencher essa lacuna, tais como: redes sociais (KADRIU, 2013), redes institucionais (LYNCH, 2003), plataforma curricular (FERNÁNDEZ-BREIS et al., 2012) entre outros recursos (BURNS; LANA; BUDD, 2013; EDGAR; WILLINSKY, 2010).

Idealizada pelo Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), a plataforma Lattes foi criada com o propósito de ser a base única de informações sobre a produção científica do Brasil (BRAS, 2003). Atualmente, tal plataforma concentra currículos de graduados, especialistas, mestres e doutores, distribuídos por todo o território nacional[1].

Por sua riqueza de informações, a plataforma Lattes passou a ser utilizada pelas universidades, agências de fomentos e grupos de pesquisa como fonte principal de informações sobre o histórico acadêmico dos pesquisadores, influenciando na análise curricular e na averiguação de mérito e competência dos pleitos de financiamentos. Devido a esse fato, é praticamente obrigatório no Brasil que todos os pesquisadores (ativos ou não) mantenham seus currículos devidamente atualizados na plataforma.

A plataforma Lattes contém um número superior a 3.520.000 currículos cadastrados, sendo: $6,47\%$ de doutores, $10,54\%$ de mestres, $16,24\%$ de especialistas, 28% de graduados, $37,08\%$ de outros níveis e $1,67\%$ não informado. Os referidos dados, podem ser melhor visualizados na Figura 1.1a. Note que, dentre os currículos cadastrados, tem-se cerca de mais de 1.400.000 ainda em formação.

Apesar da disponibilização pública das informações, instituições de ensino e pesquisa solicitam ao CNPq acesso aos dados curriculares de seus professores, pesquisadores, alunos e colaboradores, com o objetivo de integrar os dados da plataforma Lattes aos seus sistemas de informação. Deste modo, a plataforma Lattes disponibiliza duas modalidades de extrações dos dados:

a) Espelhamento para fundações: é voltado às fundações estaduais de apoio à pesquisa. Basicamente consiste na disponibilização integral dos dados para réplica na base espelho.

b) Espelhamento para instituições: é destinado às instituições de ensino, tem como fi-

[1] <http://lattes.cnpq.br/, acessado em: 2017-10-20>

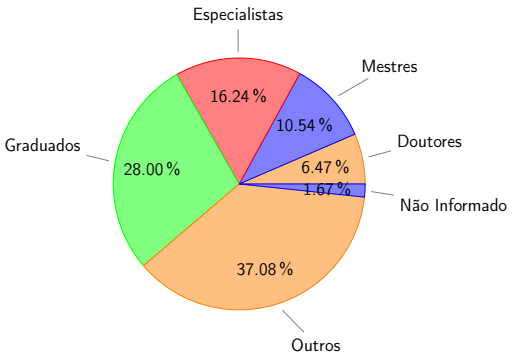


Figura 1.1: *Distribuição dos currículos cadastrados na plataforma Lattes.*
Fonte : Adaptado de pelo autor.

nalidade a extração das informações dos currículos e de grupos de pesquisa. Para as instituições utilizarem deste recurso é necessário encaminhar um ofício à presidência do CNPq, devidamente assinado pelo seu dirigente máximo.

Embora a plataforma disponibilize destas duas modalidades, surgiu a necessidade de realizar extrações destas informações de maneira ágil e automatizada (SILVA, 2007). Visando contornar essa necessidade, algumas ferramentas foram criadas pela comunidade científica, ao longo das últimas décadas, com o objetivo de auxiliar este processo. Dentre elas, destacam-se as seguintes: *OntoLattes* (BONIFACIO, 2002), a *Semantic Lattes* (COSTA; YAMATE, 2009) e o *ScriptLattes* (MENA-CHALCO; CESAR-JR, 2009). Vale destacar que o *ScriptLattes* tem ganhado atenção de diversos pesquisadores, como pode ser observado na literatura (MENA-CHALCO; CESAR-JR, 2013; FERNÁNDEZ-BREIS et al., 2012; FERRAZ; QUONIAM; MACCARI, 2014; NIGRO et al., 2015).

Em poucas palavras, o *ScriptLattes* é uma ferramenta computacional que recebe como entrada dois arquivos, um chamado *list*, que contém um lista de identificadores para a plataforma Lattes, e outro chamado *config*, contendo informações sobre o que se deseja extrair dos currículos Lattes. Após executar o *ScriptLattes*, é produzido um relatório em formato *HTML*, contendo descrições estatísticas sobre o conjunto de currículos, como por exemplo: número de artigos publicados, número de dissertações defendidas, entre outras mensurações acadêmicas. O *ScriptLattes* foi evoluindo com o passar dos anos e a partir dele novas pesquisas se iniciaram, como por exemplo: Em Giordano, Bruning e Bordin (2015) são analisadas redes de colaborações científicas, em NIGRO et al. (2016) são averiguados dados sobre as competências de pesquisa sobre a dengue no Brasil, e em NIGRO et al. (2015) são extraídas informações para o preenchimento da Plataforma Sucupira.

1.2 Problema de pesquisa

Como exposto anteriormente, o *ScriptLattes* produz um relatório estatístico a partir de um conjunto de currículos. Nesse sentido, uma das limitações do *ScriptLattes* está

em não prover relatórios com diferentes visões [2]do mesmo conjunto de currículo. Isso implica em dizer que, a partir de um conjunto de currículos, não é possível gerar relatórios categorizados por área de conhecimento, por curso ou até por linha de pesquisa. Para fazer isso, é necessário realizar a categorização (agrupamento) manual, seguida da execução do *ScriptLattes* para cada categoria (grupo), gerando uma tarefa árdua.

Não obstante, a categorização dos currículos, possibilita compará-los a partir de um determinado critério de similaridade, ou seja, é possível comparar determinadas categorias com outras por meio de uma medida de similaridade produzida, como por exemplo, através de indicadores de área.

Assim, este trabalho propõe a criação de um método que visa a utilização dos relatórios gerados pelo *ScriptLattes*, a fim de responder as seguintes questões:

a) Como criar e visualizar relatórios categorizados e gerados através de informações coletadas pelo *ScriptLattes*?

b) Como comparar programas de pós-graduação *stricto sensu*, utilizando indicadores de área computados a partir de relatórios categorizados?

Para responder estas questões, parte-se do princípio que o utilizador desta ferramenta defina as entradas dos grupos para a geração dos relatórios. Para o caso da comparação é necessário que os dados já estejam categorizados e agrupados. Seguindo essas premissas, foram definidas as seguintes hipóteses:

a) Técnicas de agrupamento de dados podem ser aplicadas na categorização de relatórios gerados pelo *ScriptLattes*.

b) Indicadores de áreas são informações que contribuem para um modelo de classificação de nota CAPES em programas de pós-graduação *stricto sensu*.

1.3 Objetivos

Com o intuito de validar as hipóteses e responder as questões de pesquisa, o objetivo deste trabalho está em criar um método para realizar à categorização em hierarquia os currículos extraídos da plataforma Lattes, utilizando informações de produção bibliográfica. Incluem-se também, como objetivos da pesquisa em questão, os seguintes:

a) Desenvolver uma método para categorizar relatórios gerados pelo *ScriptLattes*, utilizando conceitos de agrupamento hierárquico.

b) Aplicar este método proposto para extrair informações categorizadas para o preenchimento da plataforma Sucupira e e-Mec

[2]Denomina-se, neste trabalho, o termo visão do conjunto de currículos um relatório produzido pelo *ScriptLattes*

c) Aplicar este método proposto para comparar programas de pós-graduação *stricto sensu* por meio de indicadores de área.

d) Disponibilizar *script's* em linguagem de programação *python*, para a reprodução deste método e das aplicações propostas.

1.4 Estrutura do trabalho

O restante deste trabalho está organizado da seguinte forma: No Capítulo 2, são descritos os conceitos iniciais deste trabalho, ressaltando a importância dos *Web-Crawlers* na mineração dos dados, a utilização de ferramentas para extração e análise das informações da plataforma Lattes, e também os trabalhos que utilizaram a ferramenta *ScriptLattes*. No Capítulo 3, formalizado o conceito de agrupamento hierárquico de currículos extraídos da plataforma Lattes. No Capítulo 4, é descrito o método de hierarquia de currículo em um *script's* denominado *MultiScriptLattes*, assim como configurações específicas com foco na hierarquia de currículos para os programas de *stricto sensu* e em cursos de graduação. Além é apresentado a ferramenta *ScriptComp* que faz uma comparação através da similaridades dos programas de *stricto sensu*, assim como os resultados obtidos. Por último, o Capítulo 5 descreve a conclusão, assim como os trabalhos futuros.

Capítulo 2

Web-Crawlers para extração de dados na plataforma Lattes

Resumo do capítulo

O acesso as informações da plataforma Lattes tornou-se ação de grande importância para a comunidade científica, devido ao fato desta plataforma concentrar dados de diversos pesquisadores, e por ser a principal fonte de dados sobre evoluções de pesquisas no Brasil. Embora o CNPq disponibilize duas modalidades de extração em massa dos dados, utilizar ferramentas como Web-Crawlers*, passou a ter grande aceitação pela comunidade científica, pois além de realizar as extrações, estas ferramentas mineram as informações tornando-as mais objetivas e exploram relações entre os currículos extraídos.*

2.1 *Web-Crawlers*

Web-Crawlers é o nome dado às ferramentas de buscas e extrações de informações de determinadas páginas da *web*. Basicamente essas ferramentas partem de um conjunto de *URLs* (*Uniform Resource Locator*), com objetivo de extrair páginas da internet, ou apenas um conteúdo específico contido nelas. Essa necessidade surge geralmente quando a interface de uma página da *web* é complexa e não oferece opções para uma determinada extração, ou depende da entrada do usuário todas as vezes que é necessário recuperar um conteúdo (MIRTAHERI et al., 2013).

A história dos *Web-Crawlers* teve seu início em 1991, com a criação das ferramentas *World Wide Web Wanderer*, *Jump Station*, *World Wide Web Worm* e *RBSE Spider* (O.A, 1994). Esses *Web-Crawlers* buscavam informações estatísticas de algumas *URLs* que constantemente eram extraídas e armazenadas em um repositório próprio (MIRTAHERI et al., 2013). Em 1994, foram criados dois outros *Web-Crawlers*: *WebCrawler* e *MOMSpider*. Estas ferramentas além de buscarem informações da *web*, conseguiam recuperar e classificar a confiança de uma determinada página, inserindo essa informação em um lista-negra, para ser utilizada por outros *Web-Crawlers*. Um destaque para ferramenta *WebCrawler*, pois foi a percursora a permitir que seus usuários explorassem o conteúdo das páginas *web* e não apenas as palavras chaves das páginas (KAUSAR; DHAKA; SINGH, 2013).

Entre 1994 e 1998, alguns motores comerciais de busca surgiram na *web* utilizando o conceito e a inspiração dos *Web-Crawlers*, tais como: *infoseek*[1], *lycos* [2], *altavista* [3],

[1] http://www.infoseek.co.jp, acessado em 2017-010-29
[2] http://www.lycos.com, acessado em 2017-10-19
[3] http://www.altavista.com, acessado em 2017-10-24

dogpile [4] e *ask* [5].

Em 1998 o buscador Google foi lançado, e no mesmo ano teve uma grande adesão do mercado[6]. Com uma interface simples e bem organizada, trazia resultados mais precisos com um filtro bem definido, excluindo uma grande parte de resultados não relevantes, devido à utilização do algoritmo de *PageRank* (PAGE et al., 1999).

Atualmente o conceito de *Web-Crawlers*, possui aceitação relevante dos pesquisadores em diversas áreas do conhecimento, tais como: mineração de texto (SANTOS, 2014), análise de vulnerabilidades (ROCHA; KREUTZ; TURCHETTI, 2012; MACHADO et al., 2016), recuperação de informações (RODRIGUES et al., 2010) e mineração de dados (ELISHAR et al., 2012; SANTOS, 2010; MENA-CHALCO; CESAR-JR, 2009; ALVES; SANTOS; SCHIMIT, 2016).

2.2 Ferramentas para extração de dados na plataforma Lattes e *Web-Crawlers*

No Brasil, a plataforma Lattes é a base de dados com maior número de informações relacionadas a pesquisadores, sendo referência mundial no quesito organização da vida acadêmica e científica de um país. Como já dito no capítulo inicial, extrair informações da plataforma Lattes não é uma tarefa trivial. Sendo assim, diversas pesquisas foram iniciadas com o intuito de extrair informações e gerar dados estatísticos de forma mais amigável, dentre elas temos:

2.2.1 *OntoLattes*

Criado por Bonifacio (2002), a ferramenta *OntoLattes* tem por base a criação de uma ontologia do sistema de currículo da plataforma Lattes, com intuito de representá-lo em um modelo conceitual. Com este modelo criado, é possível disponibilizar os dados, advindos da plataforma Lattes, em um repositório utilizando uma linguagem semântica.

A utilização de uma linguagem semântica, veio para favorecer a consulta de informações dos currículos neste repositório. Por possibilitar a criação de conexões bem definidas dos dados, a linguagem *DAML+OIL* (linguagem de ontologia especificamente projetada para utilização na *web* (HORROCKS et al., 2002)), é possível realizar consultas mais produtivas e objetivas de um determinado assunto disposto em um conjuntos de currículos, através da equalização dos termos e os conceitos da ontologia (BONIFACIO, 2002).

A ferramenta também oferece a possibilidade do usuário traduzir seus currículos para o formato DAML+OIL, disponibilizando, quando necessário, na *web*. Na Figura 2.1, é possível ver o processo desta ferramenta.

[4] http://www.dogpile.com, acessado em 2017-10-09
[5] http://www.ask.com, acessado em 2017-10-09
[6] https://www.google.com/intl/en/about/our-story/, acessado em: 2017-10-09

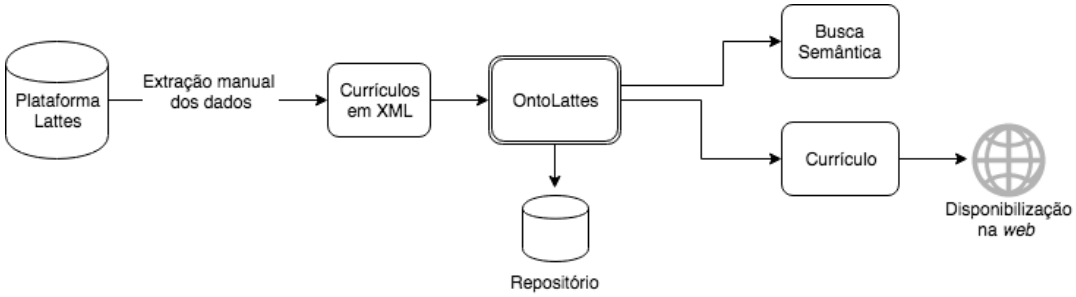


Figura 2.1: *Processo da ferramenta* OntoLattes

Fonte : Próprio autor

2.2.2 *GeraLattes*

O *GeraLattes* é uma ferramenta de mineração das informações existentes em um conjunto de currículos Lattes. A execução da mineração, verifica-se, inicialmente, por intermédio do agrupamento das informações, do seguinte modo: quantidades de publicações, teses, participações em bancas, entre outros (OLIVEIRA; BERMEJO; KERN, 2004). Após realizar as extrações das informações, é criado um relatório contendo as informações agrupadas deste conjunto de currículos.

Para iniciar este agrupamento o usuário do *GeraLattes* deve extrair as informações da plataforma Lattes, de forma manual, em formato *eXtensible Markup Language* (*XML*) Após a extração manual, deverá realizar o carregamento dos currículos para a ferramenta. Concluída a etapa anterior, os dados estarão preparados para iniciar a mineração.

2.2.3 *SemanticLattes*

A ferramenta *SemanticLattes*, foi criada por Costa e Yamate (2009), com a proposta de buscar informações de um conjunto de currículos cadastrados na plataforma Lattes, utilizando relações semânticas. Aplicação do conceito de *web* semântica sobre os dados extraídos da plataforma Lattes, possibilita a criação de ontologias sobre os domínios Lattes e Qualis (procedimentos para estratificação da qualidade da produção intelectual dos programas de pós-graduação).

As ontologias são criadas utilizando a linguagem *Web Ontology Language* (*OWL*), especifica para criação de ontologias (BECHHOFER, 2009), facilitando a criação dos domínios necessários para realizar buscas semânticas das informações existentes em um currículo Lattes.

Para criar as relações semânticas, é necessário que um conjunto de currículos em formato *XML* seja carregado para a ferramenta *SemanticLattes*. Esta ferramenta, por sua vez, não possui um módulo específico para extração das informações da plataforma Lattes, Deste modo, para que o usuário possa criar essa base de relações, deverá fazer uma

extração manual dos currículos que deseja carregar para a ferramenta. Na Figura 2.2, é descrito os passos necessários para utilização desta ferramenta.

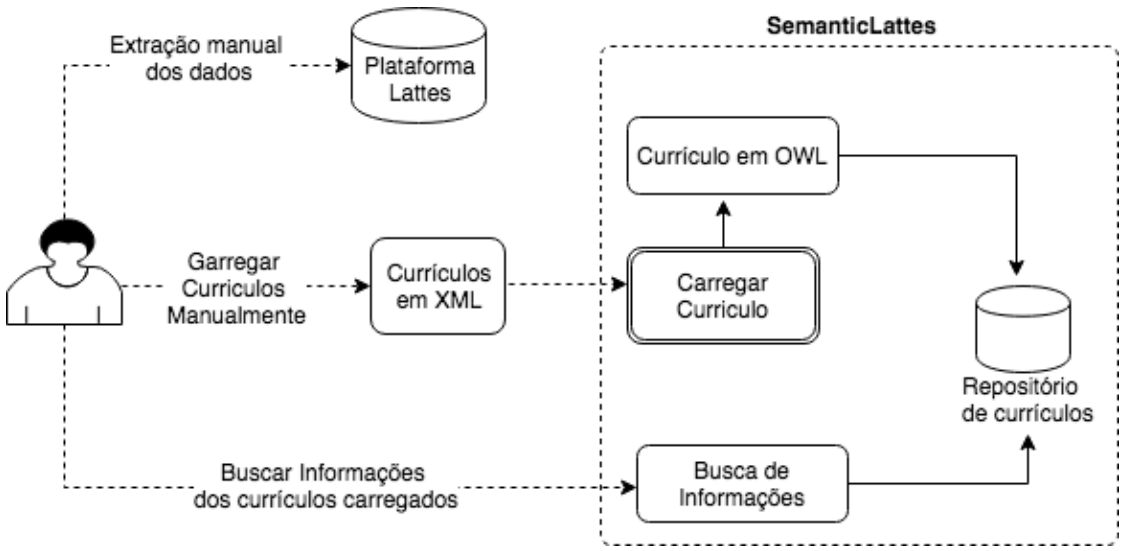


Figura 2.2: *Passos para a utilização da ferramenta* SemanticLattes

Fonte : Próprio autor

2.2.4 *LattesMiner*

O *LattesMiner* pode ser considerado um conjunto de bibliotecas, conhecidas como *Application Programming Interface* (API), para extração de dados da plataforma Lattes. Essa biblioteca foi desenvolvida em linguagem *Java* com conceito de Linguagem de Domínio Especifico (LDE), que se refere a criação de um padrão para utilização de suas funções (ALVES; YANASSE; SOMA, 2011).

Uma das facilidades do *LattesMiner* é a função para busca dos *IDs* (identificados dos currículos), essa opção facilita a busca de um grande volume de dados. Por ser construído com o conceito específico, é facilmente adaptável a uma ferramenta ou sistema próprio.

Uma das possibilidades é a implementação de regras para mineração dos dados. Os currículos são extraídos em formato *HTML*, o que permite, com base em suas características, a criação de combinações dos dados para obter resultados sumarizados, tais como: quantidade de artigos, quantidade de orientações, quantidade de teses, presença em bancas, entre outras informações acadêmicas. A Figura 2.3 demonstra o processo de utilização do *LattesMiner*.

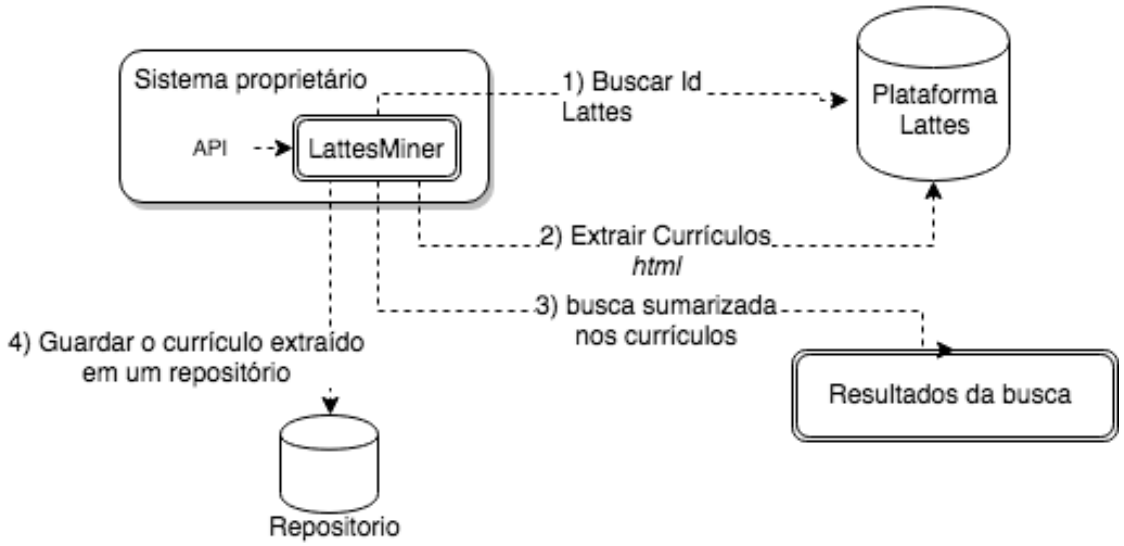


Figura 2.3: *Passos para utilização do* LattesMiner

Fonte : Próprio autor

2.2.5 *Lattes Extrator*

O *Lattes Extrator* é uma ferramenta *online* disponibilizada pelo CNPq para realizar extrações em massa dos currículos, está destinada às instituições e usuários credenciados juntos ao CNPq (STELA, 2002).

O CNPq não disponibiliza nenhuma ferramenta apta a possibilitar a análise dos dados extraídos. O mecanismo online em questão, tão somente realiza o download dos currículos no formato *XML*.

Tal ferramenta, é utilizada por instituições que já possuem sistemas para importação destes currículos.

Cumpre acrescentar por fim, que a grande desvantagem do *Lattes Extrator*, é ser um sistema fechado, posto que necessita de uma licença prévia para que possa ser usado.

2.2.6 *ScriptLattes*

A ferramenta *ScriptLattes* foi desenvolvida por Mena-Chalco e Cesar-Jr (2009), Atualmente, é a de maior utilização pela comunidade científica, para extração de informações da plataforma Lattes. Dentre as ferramentas apresentadas, o ScriptLattes, é a única que faz extrações dos currículos de forma automática utilizando o conceito dos *Web-Crawlers*. Primeiramente o ScriptLattes foi construído em linguagem *perl* e, posteriormente, migrado para linguagem *python*, que facilitou as novas atualizações no extrator dos currículos da

plataforma Lattes.

O *ScriptLattes* é uma ferramenta que faz as extrações dos currículos e a mineração dos dados, gerando relatórios como saída do seu processamento, contendo informações especificadas de cada currículo extraído, assim como relatórios de todo agrupamento. O *ScriptLattes* possui seis módulos, sendo eles:

a) Seleção dos dados:

   Este módulo é responsável por fazer a extração de currículos da plataforma Lattes em formato *html*. Durante o processo de extração, os currículos são normalizados em uma única codificação de caracteres, para facilitar a execução dos outros módulos.

   A extração é realizada utilizando o ID (identificar Lattes) de cada currículo. Composto por um número de 16 algarismos. Cada pessoa registrada na plataforma Lattes, têm seu registro associados a um ID único. Com isso, este módulo possui uma opção automática e semi-automática, que realiza a extração de uma lista de currículos com base do ID Lattes.

   Em 29 de abril de 2015 a plataforma Lattes passou a utilizar o conceito de *Captchas*, para evitar que os *Web-Crawlers* fizessem extrações em massa, sem a intervenção humana. Desta forma, a opção de extração automática ficou indisponível. No entanto, a última versão do *ScriptLattes* disponibiliza uma opção onde o usuário visualiza a imagem com o código *Captchas* e solicita que o código seja digitado manualmente. Em seguida, o currículo é extraído normalmente, de forma semiautomática.

   Um abaixo-assinado idealizado pelos criadores do ScriptLattes, através do site *on-change.org*, teve adesão de mais de 3.000 peticionários, que solicitavam a retirada dos *Captchas* dos currículos da plataforma Lattes. No entanto, a reivindicação veiculada pelo referido documento, não foi atendida pelo CNPq.

   Além da opção automática e semi-automática, o usuário pode realizar uma extração manual dos currículos em formato *html*, e disponibiliza-lo em um diretório para execução dos próximos módulos. Dependendo da quantidade de currículos a serem baixados, esta opção se torna uma tarefa demorada e cansativa.

b) Pré-processamento dos Dados

   Neste módulo é realizado uma análise no *HTML* de cada currículo extraído, a fim de encontrar as informações referentes aos dados básicos da estrutura de um currículo, sendo: informações pessoais, produções bibliográficas, produções técnicas, produções artísticas e supervisões em andamento e concluídas.

   Este módulo é totalmente dependente da estrutura dos dados disposta em um currículo Lattes, sendo necessário uma atualização no processo de análise a cada modificação realizada pelo CNPq na estrutura do currículo. Estas mudanças não são frequentes.

Todavia, caso seja necessário os mantenedores do ScriptLattes disponibilizarão uma nova versão da ferramenta com as modificações necessárias em sua página online [7].

c) Tratamento de redundâncias

Realizar um tratamento de redundância sobre os dados de currículo Lattes, é fundamental para o real entendimento das informações dispostas. Neste módulo é realizado um tratamento sobre os dados de produções realizadas em colaboração entre os pesquisadores.

No âmbito da pesquisa, é muito comum que os acadêmicos realizem seus trabalhos em colaboração (MENA-CHALCO; DIGIAMPIETRI; CESAR-JR, 2012). Desta maneira, as informações sobre um determinado artigo publicado em congresso, revistas, eventos ou por outros meios de divulgação de trabalhos, podem existir com relações a diversos autores e coautores, sendo assim, as informações podem aparecer duplicadas nos relatórios dos currículos. Para evitar tal ocorrência. Desta forma, este módulo detecta e elimina as duplicidades obtidas após o pré-processamento do currículo.

Para realizar esse tratamento, algumas características dos currículos são analisadas, tais como: título da produção, tipo de publicação e ano. Dentre essas características o título da produção é considerado o padrão para comparar as produções científicas elaboradas por um determinado grupo.

d) Grafo de colaboração

Este módulo é responsável em gerar um grafo para representar a colaboração entre um grupo de pesquisadores que tiveram relação em uma publicação. Cada membro deste grupo é representado por um vértice e as ligações são feitas pelas arestas deste grafo. A Figura 2.4 demonstra o grafo de colaboração entre pesquisadores.

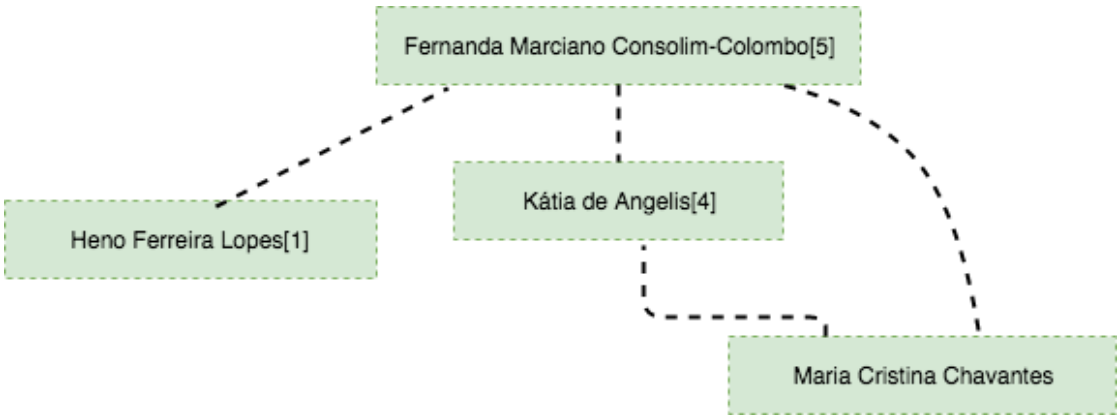


Figura 2.4: *Grafo de colaboração do ScriptLattes.*

Fonte : Adaptado de pelo autor.

[7] http://scriptlattes.sourceforge.net/, acessado em: 2017-11-07

e) Geração de mapas

Em algumas situações, mostra-se relevante saber saber a localização geográfica de cada pesquisador ou grupo de pesquisa. Neste caso, o módulo do ScriptLattes gera um mapa, que representa a localização geográfica de cada membro. Este mapa não se limita a um determinado país ou região.

Para gerar esse mapa, são utilizadas as seguintes informações: nome do país, nome da cidade e o CEP disponível no currículo Lattes de cada pesquisador. O *ScriptLattes* utiliza da ferramenta de mapa disponibilizada pelo *Google* (*Google Maps*) para capturar longitude, latitude e gerar os mapas.

f) Geração de Relatórios

Os relatórios gerados neste módulo concentram informações específicas dos currículos extraídos e processados. Os dados são separados por tipo e apresentam informações quantitativas, divididas e ordenadas por anos. É neste módulo que são gerados os gráficos de barras, no qual o comprimento destas são proporcionais às produções científicas realizadas pelo grupo de currículos ordenado por ano. Com isso, é possível analisar a quantidade de produções científicas em uma linha do tempo. Outra possibilidade é adicionar o Qualis específico do grupo. Dessa forma cada artigo publicado virá classificado com seu Qualis.

Atualmente, estes relatórios são gerados em formato *HTML*, e incluem *links* específicos para os mecanismos de busca *web* (Google e o Google Acadêmico, entre outros), para obter as citações de uma produção.

O *ScriptLattes* passou a ser adotado por diversas universidades e centros de pesquisa em todo território nacional. Ao todo são 56 cadastrados na pagina oficial do ScriptLattes. No mapa da Figura 2.5, é possível visualizar com maior clareza, em quais regiões do território brasileiro a ferramenta é usada.

Figura 2.5: *Utilização do* ScritLattes *por região.*
Fonte : Adaptado de pelo autor

2.3 Utilizando *ScriptLattes* para a extração de currículos e mineração de informações

A ferramenta *ScriptLattes* é bem completa no que tange a analise de dados de currículos Lattes. Desta análise deriva relatórios que auxiliam o pesquisador a responder questões relacionadas a vida acadêmica deste grupo, tais como:

a) Quantas produções bibliográficas, técnicas ou artísticas, foram elaboradas?

b) Quais são os diferentes tipos de produções bibliográficas?

c) Como é a evolução e a relação das publicações ao longo dos anos?

d) Como é a colaboração e a cooperação entre os pesquisadores?

e) Quantas teses e dissertações foram concluídas?

f) Qual é a distribuição geográfica dos pesquisadores?

g) Qual a formação acadêmica de um pesquisador?

As respostas às perguntas formuladas acima, são informações muito úteis, que auxiliam pesquisas em diversas áreas (FERRAZ; QUONIAM, 2017). Seu processo de utilização, embora possua módulos complexos, é bem simples. Inicialmente, é necessário ter em mãos uma lista de IDs Lattes, e preencher dois arquivos lista de IDs e configurações.

Deve-se incluir, separando com virgula, a lista de IDs com o nome do pesquisador e o período que gostaria de analisar no arquivo lista de IDs.

O arquivo de configuração, contém as configurações de cada módulo do *ScriptLattes*, e deverá ser preenchido de acordo com o nível de detalhamento que se faz necessário. Neste arquivo que será inserido o caminho do *Cache* (diretório onde será gerado os arquivos *html* dos currículos extraídos), caminho do arquivo lista de IDs e o Qualis.

Após o preenchimento dos arquivos, bastará apenas executar o arquivo *scriptLattes.py* passando o caminho do arquivo configuração) como parâmetro. No término da execução, os relatórios serão gerados no diretório informado. A Figura 2.6 demostra o processo automático para extração dos currículos da plataforma Lattes.

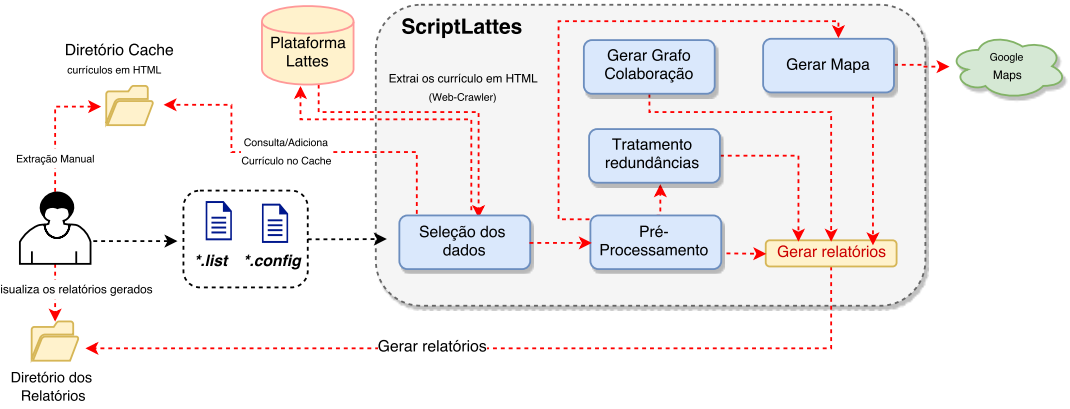


Figura 2.6: *Processo de utilização do* ScriptLattes.

Fonte : Próprio autor

2.3.1 Trabalhos que utilizam o *ScriptLattes*

Apresenta-se por este trabalho Fernández-Breis et al. (2012) o desenvolvimento de um sistema para visualizar, através de gráficos, as redes de colaboração entre um conjunto de pesquisadores. A ferramenta *ScriptLattes* foi utilizada na extração e disponibilização *on-line* de toda a produção acadêmica gerada por uma lista de IDs pré-determinada realizada manualmente (para avaliação da produção acadêmica de um Programa de pós-graduação), ou por uma lista criada automaticamente pela plataforma Lattes (que neste caso referiu-se à lista de pesquisadores sobre o tema nanotecnologia no Brasil).

Para a análise da rede de colaboração científica, Giordano, Bruning e Bordin (2015), utilizaram, em conjunto, as ferramentas *Scriptlattes* e *Gephi*. Com finalidade de demonstrar o uso destas ferramentas, foi realizado um estudo da rede de colaboração científica de professores e alunos de uma universidade. A partir dos resultados obtidos, observou-se um grande número de componentes, que, comparado com o total de nós da rede, indicava que um percentual considerável de pesquisadores, não produziu conjuntamente com outros na instituição avaliada.

No trabalho de, Mena-Chalco e Cesar-Jr (2013) detalharam a utilização da ferramenta *ScriptLattes*, para prospectar dados acadêmicos. A prospecção de dados de grupos de médio a grande porte, utilizando a plataforma Lattes, não é uma tarefa trivial, e fazê-la de forma manual acarreta brechas para falhas, no que se deve-se à contabilização das informações existentes em cada currículo. Portanto, utilizar os relatórios gerados pelo *ScriptLattes*, facilita esse processo, pois, como dito anteriormente, esta ferramenta dispõe de módulos específicos para processar as informações extraídas, gerando assim, relatórios com informações fidedignas dispostas na plataforma Lattes.

A fim de identificar nas regiões brasileiras, as produções acadêmicas, produtividade de programas *stricto sensu* e colaborações entre pesquisadores, que se relacionam com a área da Ciência da Informação, Andretta et al. (2012) utilizaram a ferramenta *ScriptLattes* para realizar extrações em massa de currículos relacionados a este tema na plataforma Lattes. Para isso, esta pesquisa foi dividida em quatro etapas, sendo: mapeamento dos programas de pós-graduação em Ciência da Informação; criação de listas de docentes e colaboradores; produção de relatórios e tabulação; sintetização dos dados. Na etapa de produção de relatórios, foram extraídos currículos de docentes e colaboradores dos programas de pós-graduação em Ciência da Informação, do triênio de 2017 a 2009. Com isso, foi possível descobrir uma tendência de maior produtividade de programas nas regiões Nordeste e Sudeste.

Neste trabalho NIGRO et al. (2016) utilizaram as ferramentas *ScriptLattes*, *ScriptGP* e *Patent2net*, para extrair informações sobre o desenvolvimento de pesquisas sobre dengue no Brasil, e a evolução de seus grupos de pesquisas sobre esse tema.

Visando analisar redes de coautoria a partir de registros bibliográficos, Oliveira, Silva e Hayashi (2014), através dessa pesquisa extraíram dados de programas de pós-graduação *stricto sensu* em Educação, utilizando a ferramenta *ScriptLattes*. Devido às inconsistências encontradas nos dados cadastrados na plataforma Lattes, foi realizado um trabalho manual complementar, sobre os relatórios gerados pelo *ScriptLattes*. Com intuito de relacionar os dados dos autores em uma matriz, utilizou-se o software VantagePoint [8], e para visualizar redes de colaboração foi utilizado os softwares NetDrawSoftware [9] e UCINet-Software [10].

[8] https://www.thevantagepoint.com/, acessado em: 2017-10-28
[9] https://sites.google.com/site/netdrawsoftware/home, acessado em: 2017-10-28
[10] https://sites.google.com/site/ucinetsoftware/home, acessado em: 2017-10-28

Capítulo 3

Formalizando a definição de agrupamentos hierárquicos de currículos extraídos da plataforma Lattes

Resumo do capítulo

Neste Capítulo, será abordado o conceitos de agrupamento de dados para definir agrupamentos de currículos extraídos da plataforma Lattes. O agrupamento hierárquico é uma técnica de aprendizagem não supervisionada fortemente consolidada na literatura. Nesta abordagem, grupos de objetos são agrupados de acordo com uma medida de similaridade, de forma a ordenar os objetos por meio de um refinamento. Os objetos que se quer agrupar são currículos extraídos da plataforma Lattes. Em adição, categorizar os currículos extraídos da plataforma Lattes revelam informações importantes sobre a pesquisa no país.

3.1 Agrupamento de currículo Lattes

Nesta serão abordados o conceito de agrupamento de dados para definir agrupamentos de currículos extraídos da plataforma Lattes. Assim, denota-se por $\mathcal{C}_v$ o conjunto de Currículos Lattes. Neste sentido, é dito que $\mathcal{H}$ é um agrupamento (ou partição) de currículos sobre o conjunto $\mathcal{C}_v$, se somente se, $\mathcal{H}$ conter n grupos (ou ainda n subconjuntos disjuntos de $\mathcal{C}_v$) $\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_n$ de $\mathcal{C}_v$, que satisfaçam as seguintes condições:

1. O agrupamento $\mathcal{H}$ não pode conter um conjunto vazio, ou seja,

$$\emptyset \notin \mathcal{H}. \tag{3.1}$$

2. A união dos grupos em $\mathcal{H}$ é igual a $\mathcal{C}_v$, isto é,

$$\mathcal{C}_v = \bigcup_{\mathcal{S}_i \in \mathcal{H}} \mathcal{S}_i. \tag{3.2}$$

3. A intersecção dos dois grupos disjuntos em $\mathcal{H}$ deve ser vazia, ou seja,

$$\forall \mathcal{S}_i, \mathcal{S}_j \in \mathcal{H}, \mathcal{S}_i \neq \mathcal{S}_j \Rightarrow \mathcal{S}_i \cap \mathcal{S}_j = \emptyset. \tag{3.3}$$

3.2 Hierarquia de Currículos

Uma hierarquia de currículos $\mathcal{T}$ é um conjunto de agrupamento $\{\mathcal{H}_1, \mathcal{H}_2, \ldots, \mathcal{H}_N\}$ indexados por uma relação de ordem sobre os grupos dos agrupamentos. Primeiramente, os agrupamentos estão ordenados por uma relação de ordem, denominado como *refinamento*, definido sobre o conjunto de currículos $\mathcal{C}_v$ da seguinte forma: para quaisquer dois agrupamentos $\mathcal{H}_i$ e $\mathcal{H}_j$ (com $i < j$) sobre $\mathcal{C}_v$, pode-se dizer que $\mathcal{H}_i$ é um refinamento de $\mathcal{H}_j$, se e somente se, cada grupo de $\mathcal{H}_i$ é um subconjunto de algum grupo de $\mathcal{H}_j$, ou seja,

$$\mathcal{H}_i \text{ é um } \textit{refinamento} \text{ de } \mathcal{H}_j \Leftrightarrow \forall C \in \mathcal{C}_v, \mathcal{H}_i(C) \subseteq \mathcal{H}_j(C), \tag{3.4}$$

em que $\mathcal{H}(C)$ é o grupo $\mathcal{S} \in \mathcal{H}$ contendo o currículo Lattes C, isto é, $\mathcal{H}(C) = \mathcal{S}$, se e somente se, $C \in \mathcal{S}$[1].

Assim, a hierarquia de currículos é um subconjunto $\mathcal{T} \subseteq \{\mathcal{S} : \mathcal{S} \in \bigcup_{i=1}^{N} \mathcal{H}_i\}$ dos grupos dos agrupamentos sobre $\mathcal{C}_v$ parcialmente ordenados pela relação de inclusão, isto é, $(\mathcal{T}, \subseteq)$. Vale salientar, que um conjunto parcialmente ordenado (poset - do inglês, *partially ordered set*) pode ser representado por um grafo não-direcionado conhecido por *Diagrama de Hasse* e desenhado de forma hierárquica.

Para ilustrar essa ideia, considere o seguinte exemplo: seja $\mathcal{H}_{\text{AreaDeConhecimento}}$, $\mathcal{H}_{\text{Curso}}$ e $\mathcal{H}_{\text{LinhaDePesquisa}}$ agrupamentos sobre $\mathcal{C}_v$ de tal forma que os currículos estão agrupados por área de conhecimento, curso e linha de pesquisa, respectivamente, e satisfazendo : $\mathcal{C}_v \preceq \mathcal{H}_{\text{LinhaDePesquisa}} \preceq \mathcal{H}_{\text{Curso}} \preceq \mathcal{H}_{\text{AreaDeConhecimento}} \preceq \{\mathcal{C}_v\}$. Assim, tem-se:

a) $\mathcal{S}_i \in \mathcal{H}_{\text{AreaDeConhecimento}}$ é um grupo de $\mathcal{C}_v$ pertencentes área de conhecimento i;

b) $\mathcal{S}_{ij} \in \mathcal{H}_{\text{Curso}}$ é um grupo de $\mathcal{S}_i \subseteq \mathcal{C}_v$ pertencente ao curso j da área de conhecimento i;

c) $\mathcal{S}_{ijk} \in \mathcal{H}_{\text{LinhaDePesquisa}}$ é um grupo de $\mathcal{S}_{ij} \subseteq \mathcal{S}_i \subseteq \mathcal{C}_v$ pertencentes à linha de pesquisa k do curso j da área de conhecimento i.

Portanto, $\mathcal{T} = \{\mathcal{S} \in \mathcal{H}_{\text{Area}}\} \cup \{\mathcal{S} \in \mathcal{H}_{\text{Curso}}\} \cup \{\mathcal{S} \in \mathcal{H}_{\text{LinhaDePesquisa}}\} \cup \{\mathcal{S} \in \mathcal{C}_v\}$. A Figura 3.1 mostra as ramificações do diagrama Hasse $(\mathcal{T}, \preceq)$, desta situação.

Nesta dissertação interessados em construir hierarquias de currículos a partir de uma relação de ordem denominada *prefixo*. Com essa relação torna-se fácil representar manualmente esta hierarquia de currículos. Assim, representa-se esta hierarquia por meio de uma codificação associada aos vértices de tal modo a satisfazer a relação prefixo definida como: $\mathcal{S}_i, \mathcal{S}_j \in \mathcal{T}$, $\mathcal{S}_i$ é prefixo de $\mathcal{S}_j$, se e somente se, código de $\mathcal{S}_i$ é prefixo do código de $\mathcal{S}_j$. Assim, a hierarquia de currículos $\mathcal{T}$ e a relação de ordem prefixo constituem um *poset*. Note que, uma possível codificação de $\mathcal{T}$ pode ser vista na Figura 3.1.

[1]Para mais detalhes verificar em Brualdi (2012) e Newman (1992).

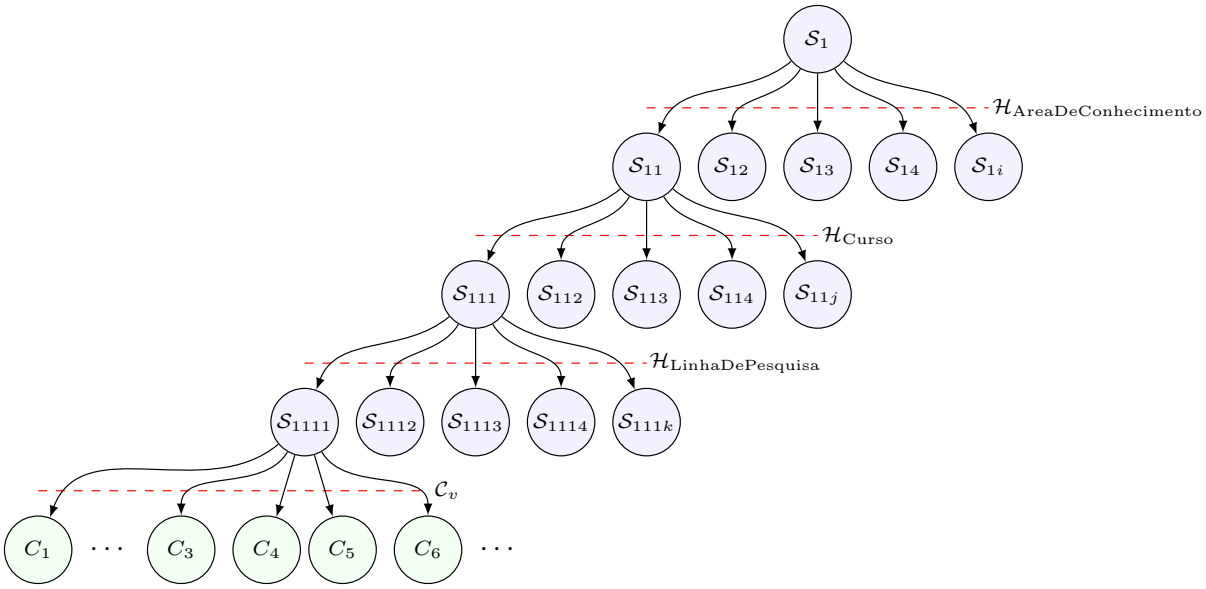


Figura 3.1: *Exemplo: ramificações do diagrama Hasse* $(\mathcal{T}, \preceq)$.
Fonte : Próprio autor

3.3 Agrupamento Hierárquico

Nesta dissertação, além do interesse em construir hierarquias de currículos manualmente, pretende-se também construir hierarquias por meio de medidas de similaridades que podem ser extraídas a partir do conjunto de currículos. Neste sentido é apresentado uma técnica de agrupamento hierárquico (DUDA et al., 2001; FRIEDMAN; HASTIE; TIBSHIRANI, 2001) que permite a criação de uma árvore de agrupamentos, como mencionado anteriormente. Na literatura, há duas abordagens tradicionais que podem ser aplicadas para a construção de agrupamento hierárquico (DUDA et al., 2001; FRIEDMAN; HASTIE; TIBSHIRANI, 2001):

a) *Divisivo*: neste método é atribuído todos os objetos para um único grupo, em seguida, este grupo é subdividido em dois grupos. Este processo de subdividir se repete até que haja apenas grupos unitários.

b) *Aglomerativo*: inicialmente cada currículo é considerado um grupo unitário. Após, cria-se um novo grupo por meio da fusão de dois grupos similares. Este processo de fusão se repete até que haja apenas um único grupo.

Seguindo nesta linha, para construir o agrupamento hierárquico $\mathcal{T}$, é necessário determinar a similaridade entre cada par de grupos por meio de uma função de distância. Os principais métodos utilizados para realizar as similaridades entre grupos são (DUDA et al., 2001; FRIEDMAN; HASTIE; TIBSHIRANI, 2001):

1. *Ligações simples*: a distância entre dois grupos $\mathcal{S}_i$ e $\mathcal{S}_j$ é definida como a menor

distância entre dois objetos, isto é

$$D_{min}(\mathcal{S}_i, \mathcal{S}_j) = \min\{dist(C_i, C_j) : C_i \in \mathcal{S}_i, C_j \in \mathcal{S}_j\}. \tag{3.5}$$

2. *Ligações completas*: a distância entre os dois grupos $\mathcal{S}_i$ e $\mathcal{S}_j$ é definida como a maior distância entre dois objetos, ou seja

$$D_{max}(\mathcal{S}_i, \mathcal{S}_j) = \max\{dist(C_i, C_j) : C_i \in \mathcal{S}_i, C_j \in \mathcal{S}_j\}. \tag{3.6}$$

3. *Ligações por médias*: a distância entre os grupos $\mathcal{S}_i$ e $\mathcal{S}_j$ é definida pela distância média entre os objetos, isto é

$$D_{avg}(\mathcal{S}_i, \mathcal{S}_j) = \frac{1}{|\mathcal{S}_i| \times |\mathcal{S}_j|} \sum_{C_i \in \mathcal{S}_i} \sum_{C_j \in \mathcal{S}_j} dist(C_i, C_j). \tag{3.7}$$

Para ilustrar o funcionamento de um agrupamento hierárquico $\mathcal{T}$ considere os seguintes objetos $\{A = 25, B = 30, C = 80, D = 90, E = 60, F = 56, G = 50, H = 40\}$ e a construção usando ligações simples. Dessa forma, pelo método aglomerativo, tem-se inicialmente que cada objeto é um grupo, isto é $\{\{A\}, \{B\}, \{C\}, \{D\}, \{E\}, \{F\}, \{G\}, \{H\}\}$. Depois, repetimos o processo de fusão pelos grupos mais similares e assim temos que:

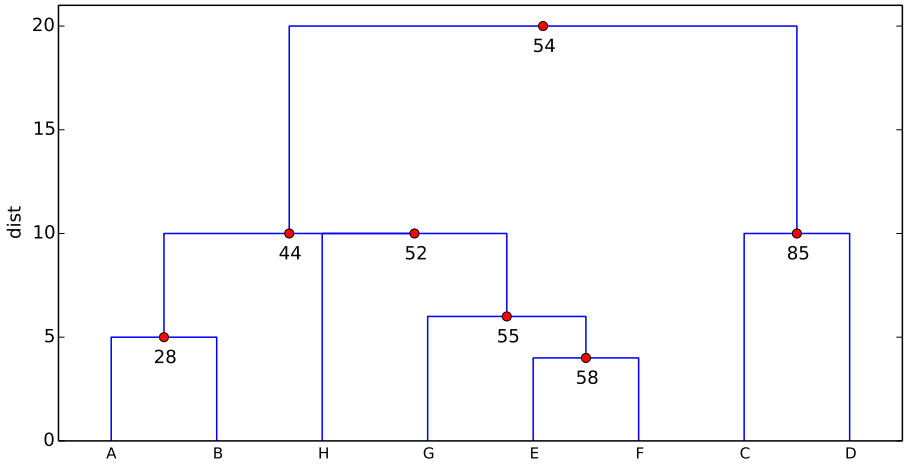
Inicialização:$\big\{\{A\}, \{B\}, \{C\}, \{D\}, \{E\}, \{F\}, \{G\}, \{H\}\big\}$

Fusão 1:$\big\{\{A\}, \{B\}, \{C\}, \{D\}, \{E, F\}, \{G\}, \{H\}\}$	$D_{min}(\{E\}, \{F\}) = 4$
Fusão 2:$\big\{\{A, B\}, \{C\}, \{D\}, \{E, F\}, \{G\}, \{H\}\}$	$D_{min}(\{A\}, \{B\}) = 5$
Fusão 3:$\big\{\{A, B\}, \{C\}, \{D\}, \{E, F, G\}, \{H\}\}$	$D_{min}(\{E, F\}, \{G\}) = 6$
Fusão 4:$\big\{\{A, B\}, \{C, D\}, \{E, F, G\}, \{H\}\}$	$D_{min}(\{C\}, \{D\}) = 10$
Fusão 5:$\big\{\{A, B\}, \{C, D\}, \{E, F, G, H\}\}$	$D_{min}(\{E, F, G\}, \{H\}) = 10$
Fusão 6:$\big\{\{A, B, E, F, G, H\}, \{C, D\}, \}$	$D_{min}(\{A, B\}, \{E, F, G, H\}) = 10$
Fusão 7:$\big\{\{A, B, E, F, G, H, C, D\}\}$	

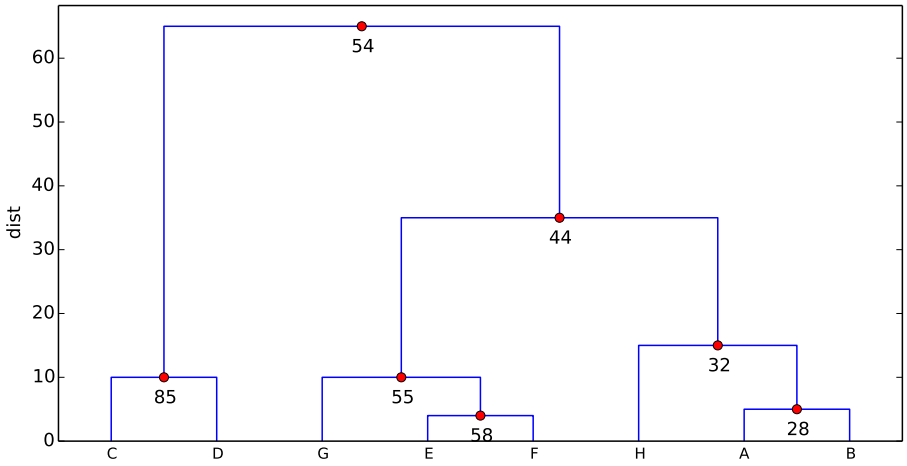
A maneira tradicional de visualizar um agrupamento hierárquico é por meio de um diagrama, chamado dendrograma, que apresenta os objetos no eixo horizontal e as distâncias entres os grupos no eixo vertical. Na Figura 3.2 são apresentados os dendrogramas deste agrupamento hierárquico construídos por ligações simples, completa e por média.

Um *corte* sobre um agrupamento hierárquico $\mathcal{T}$ é uma operação que intercepta qualquer caminho da base para o topo de $\mathcal{T}$ apenas uma vez. Mais especificamente, um corte é qualquer subconjunto $\mathcal{H}$ de $\mathcal{T}$ que tenha como resultado uma partição sobre $\mathcal{C}_v$. Neste sentido, denota-se por $\mathcal{H}_d$ uma partição sobre $\mathcal{C}_v$ obtida por um corte dado por um parâmetro de distância $d \in \mathbb{R}^+$. Na Figura 3.3 são apresentadas as partições $\mathcal{H}_{12}$, $\mathcal{H}_{17}$ e $\mathcal{H}_{30}$

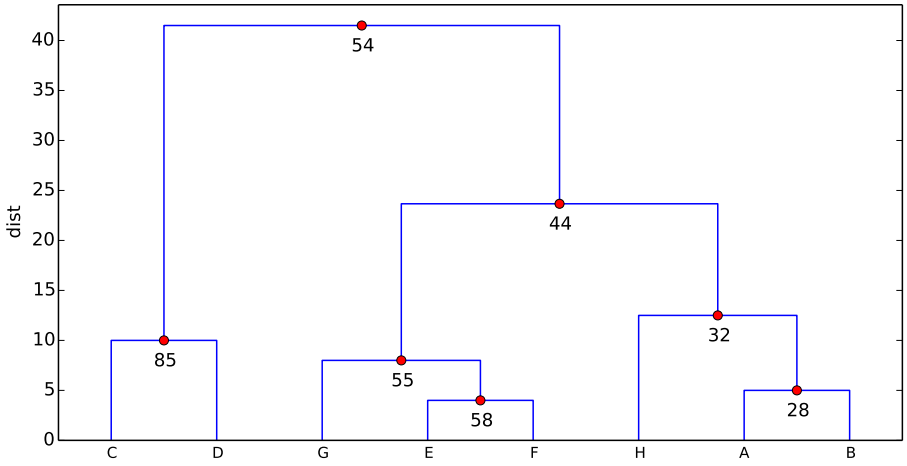
destacando-se os locais dos cortes e os agrupamentos formados. Note-se que, $\mathcal{H}_{30}$ é um refinamento de $\mathcal{H}_{17}$ e $\mathcal{H}_{17}$ é um refinamento $\mathcal{H}_{12}$. Obviamente, $\mathcal{H}_{30}$ é um refinamento de $\mathcal{H}_{12}$ pela transitividade.



(a) *Ligações simples*



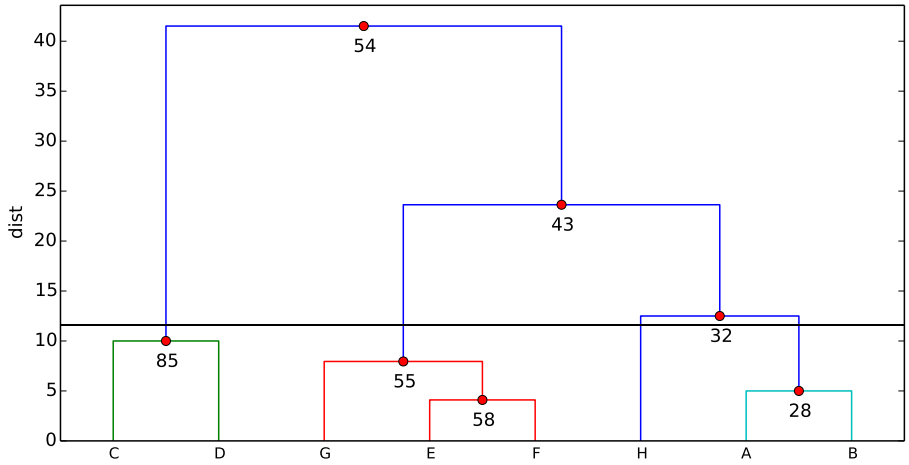
(b) *Ligações completas*



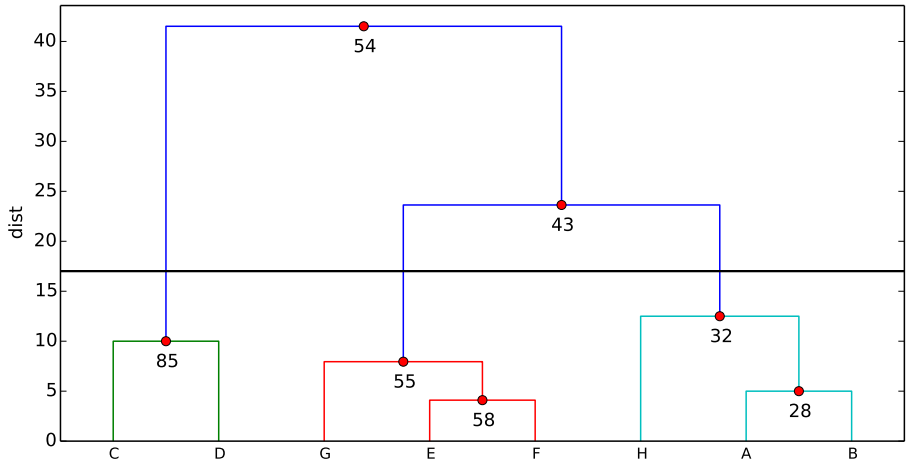
(c) *Ligações por média*

Figura 3.2: *Dendrogramas dos agrupamentos hierárquicos dos objetos* $\{A = 25, B = 30, C = 80, D = 90, E = 60, F = 56, G = 50, H = 40\}$

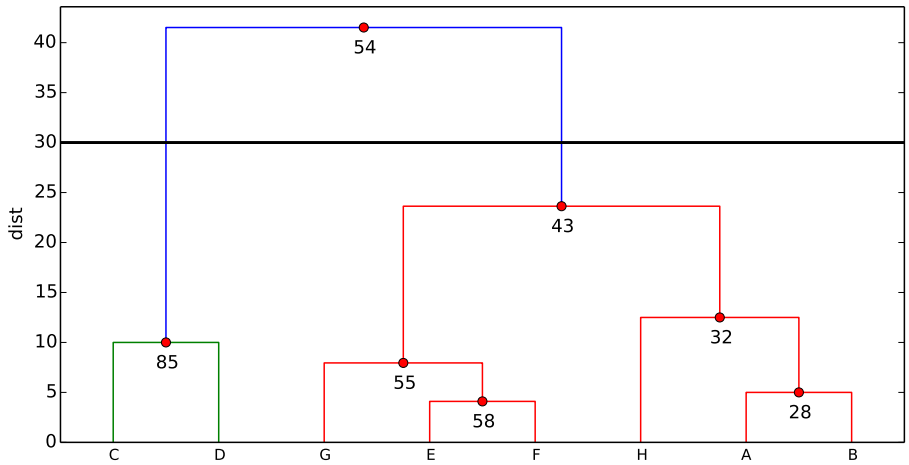
Fonte : Próprio autor



(a) *Corte em $\mathcal{T}$ que tem como resultado* $\mathcal{H}_{12} = \{\{C, D\}, \{G, E, F\}, \{H\}, \{A, B\}\}$.



(b) *Corte em $\mathcal{T}$ que tem como resultado* $\mathcal{H}_{17} = \{\{C, D\}, \{G, E, F\}, \{H, A, B\}\}$.



(c) *Corte em $\mathcal{T}$ que tem como resultado* $\mathcal{H}_{30} = \{\{C, D\}, \{G, E, F, H, A, B\}\}$.

Figura 3.3: *Agrupamentos formados por meio de cortes na hierarquia $\mathcal{T}$ construída utilizando ligações por média.*
Fonte : Próprio autor

Capítulo 4

Método de hierarquia de currículo para geração de relatórios de produção acadêmica

Resumo do capítulo

Neste Capítulo, será abordado a implementação dos métodos propostos, através da criação de script's*, assim como apresentação dos resultados obtidos. Inicialmente será apresentado uma implementação base da hierarquia de currículos. A seguir será apresentado adaptações com foco no levantamento de informações para cadastramento dos dados na plataforma Sucupira e e-Mec. Assim como uma implementação de agrupamentos hierárquicos, sobre os dados dos programas de pós-graduação* stricto sensu*, da câmera temática Engenharia/Tecnologia/Gestão da área Interdisciplinar CAPES.*

4.1 Método de hierarquia de currículos

Para geração de relatórios em diferentes visões dos dados, na Figura 4.1 é apresentado o método precursor. Este método, pode ser categorizado como referência na implementação de uma hierarquia de currículos. Por intermédio dele outras implementações surgiram, com a finalidade de resolver necessidades específicas, como: criação de hierarquia para gestão de dados de cursos superiores ou de programas de pós-graduação *stricto sensu*. Basicamente, a implementação aqui discutida, se dá em 3 etapas, a primeira, parametrização da hierarquia; a segunda, geração de *script's*; e a terceira, geração dos relatórios.

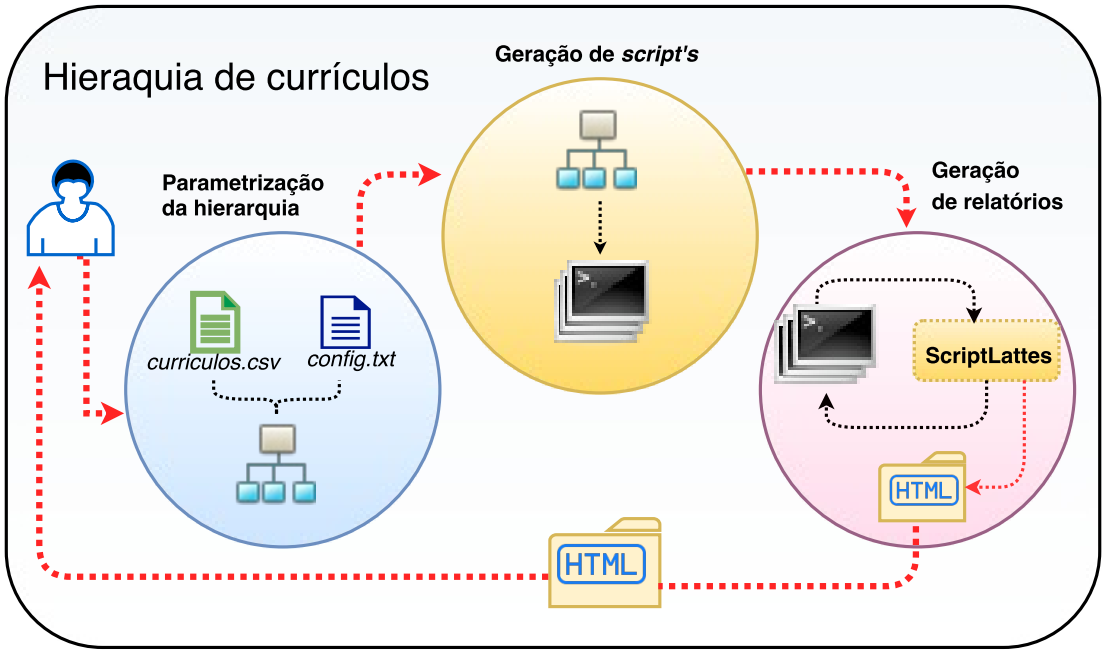


Figura 4.1: *Processo para construir hierarquia de currículos.*

Fonte : Próprio autor

4.1.1 Parametrização da hierarquia

Este módulo de parametrização é composto pela configuração manual de dois arquivos centrais, *config.txt* e *curriculos.csv*. O arquivo *config.txt* é responsável em parametrizar os dados das hierarquias, assim como as configurações exigidas para extrair os relatórios do *ScriptLattes*. No arquivo *curriculos.csv* deve conter os dados dos currículos, assim como as informações dos grupos contidos na hierarquia. Ambos os arquivos são necessários, e devem ter suas configurações sincronizadas.

O arquivo *config.txt* é basicamente composto por diversos parâmetros com seus valores, ou seja, deve conter todas as informações que serão utilizadas para construção da hierarquia e dos grupos. Os parâmetros relacionados à criação da hierarquia, podem ser configurados livremente. Entretanto, algumas configurações específicas são obrigatórias. Para simplificar, este arquivo foi separado em 3 sessões, sendo: configurações do *ScriptLattes*; configurações do arquivo *curriculos.csv*, e; criação da hierarquia. Abaixo, segue o detalhamento delas:

a) Configurações do *ScriptLattes*: para executar o *ScriptLattes* de cada agrupamento, é necessário informar o nível de detalhamento em que os relatórios serão gerados. Deste modo, faz-se imprescindível a configuração do arquivo com a extensão *config*. Nesta sessão deverão ser informados os parâmetros de configuração base do *ScriptLattes*, que serão adotados pela hierarquia. Para que assim ocorra, deve-se informar de modo obrigatório os parâmetros *diretorio_de_cache_dos_cvs* e *diretorio_de_cache_dos_doi*.

O parâmetro *diretorio_de_cache_dos_cvs* é responsável por informar em qual diretório os currículos extraídos pelo *ScriptLattes* serão armazenados, já o parâmetro *diretorio_de_cache_dos_doi* informa o diretório de armazenamento do código DOI (Digital Object Identifier [1]) dos artigos pertencentes aos currículos.

b) Criação da hierarquia: para construção da hierarquia, deve ser informado os grupos, utilizando os parâmetros:

   b.1) *Node_Nome_Coluna_n*: representa o nome da coluna no arquivo *curriculo.csv*.

   b.2) *Node_Valor_Celula_Coluna_completo_n*: representa o valor da coluna no arquivo *curriculo.csv*. É a partir deste valor que é feita associação do currículo informado com o grupo da hierarquia.

   b.3) *Node_Descricao_n*: descreve o nome do agrupamento.

   b.4) *Node_Modelo_Config_Scriptlattes_n*: informa o caminho com o modelo préconfigurado do arquivo *config* do *ScriptLattes*. Os únicos parâmetros que não devem ter seus valores preenchidos no arquivo, são *global-arquivo_de_entrada* e *global-diretorio_de_saida*, pois ambos serão informados com o processamento dos grupos na etapa de geração dos relatórios.

   O sufixo *n* representa a própria hierarquia, ou seja, para cada grupo deve ser acrescentado um caractere numérico em sequência, deste modo: primeiro grupo 1, segundo grupo 11, terceiro grupo 111, segundo registro do segundo grupo 12 e terceiro registro do segundo grupo 121.

c) Configurações do arquivo *curriculos.csv*: para o parâmetro *dataset_curriculos*, deve ser informado o apontamento para o arquivo *curriculos.csv*. Assim como o separador, o parâmetro *separador_csv*, é utilizado para delimitar as colunas deste arquivo de extensão *csv*. O arquivo *curriculos.csv*, deverá conter algumas colunas com cabeçalhos e valores informados nos parâmetros *Node_Nome_Coluna_n* e *Node_Valor_Celula_Coluna_n* de cada grupo da hierarquia, descritos no arquivo *config.txt*. São por esses valores que serão feitas as associações dos currículos, com os grupos na hierarquia. A seguir detalharemos mais sobre essa associação, assim como um algoritmo que percorre a hierarquia fazendo estas associações e gerando as entradas do *ScriptLattes* de cada grupo.

A sequência descrita na sessão criação da hierarquia, poderá ter qualquer tamanho, assim como na metodologia proposta. Abaixo na Figura 4.2, segue um exemplo das sessões do *config.txt* com as 3 etapas configuradas.

[1] https://sites.google.com/site/netdrawsoftware/home, acessado em: 2017-10-28

```
# Configurações do ScriptLattes
diretorio_de_cache_dos_cvs: ./cache/uninove/ppgi/cache
diretorio_de_cache_dos_doi: ./cache/uninove/ppgi/doi
modelo_programa: ./modelos/PR-Programas.config

#Configurações do arquivo curriculos.csv
dataset_curriculos: ./curriculos.csv
separador_csv:

#Criação da hierarquia
Node_Nome_Coluna_1 = Universidade
Node_Valor_Celula_Coluna_1 = UNINOVE
Node_Descricao_1 = Universidade Nove de Julho
Node_Modelo_Config_Scriptlattes_1 = ./modelos/template.config

Node_Nome_Coluna_11 = Áreas
Node_Valor_Celula_Coluna_11 = programasExatas
Node_Descricao_11 = Pós-graduação stricto sensu em Exatas
Node_Modelo_Config_Scriptlattes_11 = ./modelos/template.config

Node_Nome_Coluna_111 = Programas
Node_Valor_Celula_Coluna_111 = PPGI
Node_Descricao_111 = Stricto Sensu em Informática e Gestão do Conhecimento
Node_Modelo_Config_Scriptlattes_111 = ./modelos/template.config

Node_Nome_Coluna_12 = Area
Node_Valor_Celula_Coluna_12 = programasSaude
Node_Descricao_12 = Pós-graduação stricto sensu em Saúde
Node_Modelo_Config_Scriptlattes_12 = ./modelos/template.config

Node_Nome_Coluna_121 = Programas
Node_Valor_Celula_Coluna_121 = PPM
Node_Descricao_121 = Pós-graduação stricto sensu em Medicina
Node_Modelo_Config_Scriptlattes_121 = ./modelos/template.config
```

Figura 4.2: *Exemplo de configuração do arquivo config.txt*
Fonte : Próprio autor

O arquivo *curriculos.csv* deve conter os dados dos currículos Lattes, dispostos na hierarquia através da relação do currículo com o grupo. Primeiramente é necessário fazer uma busca para encontrar o ID Lattes dos currículos (é com esse ID que é possível criar os arquivos com extensão *list*, que serão utilizados pelo *ScriptLattes*). Depois de obter obter estes IDs, deve-se informar os dados que se associam a hierarquia.

De forma padrão, foi criado um cabeçalho para o arquivo *curriculos.csv*, sendo: *ID_Lattes*, *Nome*, *Periodo_Ini_x* e *Periodo_Fim_x*, onde x representa o vínculo dos

períodos com o grupo da hierarquia, ou seja, é possível extrair os relatórios em diferentes períodos. Esta associação de período, grupo e currículo será detalhada a seguir.

Como dito anteriormente, para que seja feita a associação entre um currículo informado no arquivo *curriculos.csv*, com um grupo na hierarquia, e o grupo na hierarquia, é fundamental que o mencionado arquivo, tenha colunas com o mesmo nome, e valor dos parâmetros *Node_Nome_Coluna_* e *Node_Valor_Celula_Coluna_n* informados na hierarquia. Para representar melhor esta associação, vamos supor que, a Figura 4.3, represente um arquivo de configuração *config.txt* pré-configurado, e a Tabela 4.1, também represente um arquivo *curriculos.csv* pré-configurado. Por este exercício hipotético, é possível ver que as colunas representadas pela Tabela 4.1, possuem o mesmo nome e valores dos parâmetros descritos no arquivo, representados pela Figura 4.3. Nesta figura é apresentada a correspondente hierarquia, já com os currículos agrupados.

```
...
Node_Nome_Coluna_11 = Programas_1
Node_Valor_Celula_Coluna_11 = engenharia
...
Node_Nome_Coluna_12 = Programas_2
Node_Valor_Celula_Coluna_12 = agronomia
...
Node_Nome_Coluna_13 = Programas_3
Node_Valor_Celula_Coluna_13 = oceanografia
...
...
```

Figura 4.3: *Exemplo de associação de um currículo com um grupo da hierarquia*
Fonte : Próprio autor

O arquivo representado na Figura 4.3, possui 3 grupos, destacados em cores diferentes. No sufixo de cada parâmetro *Node_ ** contem o valor incremental de 2 algorismo, representando o segundo grupo da hierarquia, ou seja, é possível pressupor que exista um grupo acima definido com apenas um algorismo no sufixo, tal como:**Node_..._1**.

Tabela 4.1: *Estrutura do arquivo* curriculos.csv*, para um exemplo de associação com a hierarquia*

Id Lattes*	...	Programa_1	Programa_2	...	Programa_3 ...
C_1	...	engenharia		...	
C_2	...	engenharia		...	oceanografia...
C_3	...		agronomia	...	
C_4	...			..	oceanografia...

Na Tabela 4.1 tem-se 3 colunas destacadas, que se assemelham com os nomes dos *Node_ ** da Figura 4.3, ou seja, o cabeçalho das colunas respeitaram os valores dos parâmetros descritos no arquivo *config.txt*, sendo: a coluna de cabeçalho *Programa_1* possui o mesmo nome do valor do parâmetro *Node_Nome_Coluna_11*, o *Programa_2* possui

o mesmo nome do valor do parâmetro *Node_ Nome_ Coluna_ 12* e o *Programa_* 3 tem o mesmo nome do *Node_ Nome_ Coluna_ 13*, fazendo assim relações entre os grupos e estas colunas.

Essa relação se estende para os valores das colunas, que, para terem a relação com o grupo da hierarquia, devem dispor dos mesmos valores dos parâmetros *Node_ Valor_ Celula_ Coluna_* de mesmo sufixo, ou seja, a coluna de nome *Programa_* 1 deve conter os valores dos parâmetros *Node_ Valor_ Celula_ Coluna_ 11*. Com isso, é possível atribuir um determinado currículo (linha correspondente a coluna em questão) a um grupo. Na Figura 4.4 podemos ver uma hierarquia com os currículos C_n associados.

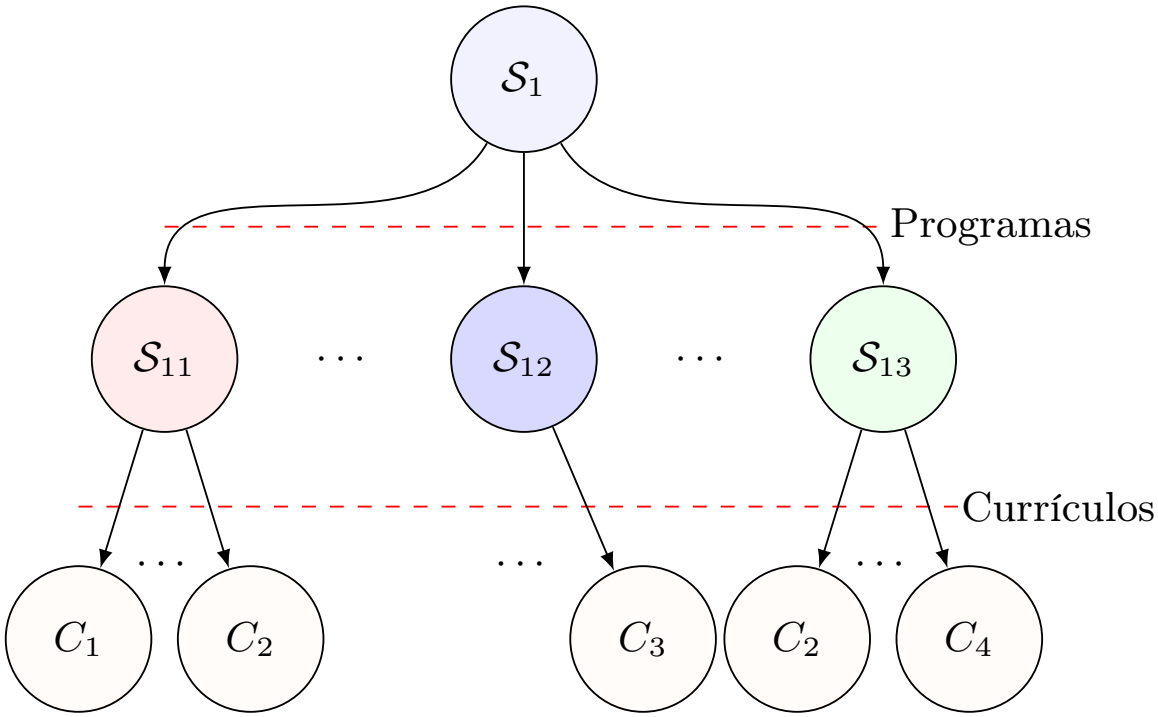


Figura 4.4: *Exemplo de hierarquia com associação dos currículos nos grupos*

Fonte : Próprio autor

Nesta hierarquia Figura 4.4), é possível identificar que a associação entre a hierarquia apresentada no arquivo de exemplo *config.txt* (ver Figura 4.3), com o arquivo *curriculos.csv* (ver Tabela **??**), foi feita com sucesso, isto é, todos os currículos informados no arquivo *curriculos.csv*, foram alocados nos grupos da hierarquia. Um detalhe interessante desta alocação, está no currículo C_2, que, no arquivo *curriculos.csv* dispõem de 2 valores em duas colunas que representam os grupos (*Programa_* 1 e *Programa_* 3). Assim sendo, o algoritmo faz alocação do mesmo registro para 2 grupos distintos. Neste caso, C_2 pertence aos grupos $\mathcal{S}_{11}$ (*Programa_* 1) e $\mathcal{S}_{13}$ (*Programa_* 3).

Caso o nomes das colunas, assim como seus valores, forem diferentes dos valores informados nos parâmetros *Node_ Nome_ Coluna_ n* e *Node_ Valor_ Celula_ Coluna_ n*, o algoritmo não faz as alocações, ou seja, para a realizar as alocações é obrigatório que estes

valores sejam iguais.

Como dito anteriormente, o cabeçalho padrão para o arquivo *curriculos.csv*, contém informações *ID_Lattes* (Id de um currículo na plataforma Lattes), *Nome* (nome do proprietário do currículo), *Periodo_Ini_x* e *Periodo_Fim_x*.

As colunas *Periodo_Ini_x* e *Periodo_Fim_x* descrevem o período, em anos, que se deseja gerar os relatórios do *ScriptLattes* de um determinado grupo, ou seja, para fazer o vínculo do currículo, grupos e os períodos, o nome das colunas de períodos deve conter o sufixo com o nome do parâmetro *Node_Nome_Coluna_n*, como exemplo, Supõe-se que a coluna *Programa_1* represente um grupo (*Node_Nome_Coluna_n*). Para informar os períodos que determinados currículos estiveram neste grupo, basta, tão somente, substituir o *x* das colunas *Periodo_Ini_x* e *Periodo_Fim_x* pelo nome do grupo (*Node_Nome_Coluna_n*), ficando *Periodo_Ini_Programa_1* e *Periodo_Fim_Programa_1*. Sendo indispensável informar o *Periodo_Ini_x* e *Periodo_Fim_x* para todos os grupos (*Node_Nome_Coluna_n*) da hierarquia. Deste modo, o vínculo entre currículos, grupos e os períodos que os currículos estiveram nos grupos da hierarquia, ficará completo. A Tabela 4.2, representa um arquivo *curriculo.csv* preenchido com os períodos.

Tabela 4.2: *Estrutura do arquivo* curriculos.csv

ID_Lattes	Nome	Programa_1	Periodo_Ini_Programa_1	Periodo_Fim_Programa_1	...
8352781922744228	Julio Medeiros	PPGI	2013	2016	...
3254781925474920	Carlos Padro	PPGI	2013	2015	...
4234234222343212	Mario Furlan	PPGI	2013	2016	...
8726383987762839	Gabriel Tavares	PPGI	2013	2016	...
4324242422234324	Juliana Prisco	PPGI	2016	2017	...
3211312311232122	David Hilbert	PPGI	2013	2016	...
5435312345134545	George Pólya	PPGI	2013	2016	...
9854358353985349	Carl Gauss	PPGI	2013	2016	...

4.1.2 Geração de Scripts

Com a hierarquia criada, através do preenchimento dos arquivos *config* e *curriculo.csv*, é possível gerar um *script's* que contém todas as chamadas para o *ScriptLattes*. Conforme já exposto anteriormente o vínculo entre os currículos e os grupos da hierarquia, devem ser feitos pelo cruzamento dos valores dos parâmetros *Node_Nome_Coluna_* e *Node_Valor_Celula_Coluna_n*. Com isso, é possível iterar a lista de currículos e alocar cada um deles em seus devidos grupos.

Com esta alocação, é gerado, para cada grupo, os arquivos de entrada *list* e *config* do *ScriptLattes*, assim como um *script* que unifica todos os dados de entrada dos arquivos gerados, possibilitando uma única execução para toda a hierarquia. O Algoritmo 1, demostra as funções em alto nível para geração das entradas(*list* e *config*) para as chamadas

do *ScriptLattes*.

Algoritmo 1 Algoritmo para geração dos arquivos de entrada do *ScriptLattes*, para cada currículos dos grupos

função $computarScriptPelaHierarquia(\mathcal{S})$
  $\mathcal{C}^* \Leftarrow buscarCurriculoNoGrupo(\mathcal{S})$
  $gerarArquivosEntradaScriptlattes(\mathcal{C}^*, \mathcal{S})$
  para cada $\mathcal{S}' \in Filhos(\mathcal{S})$ **faça**
    $computarScriptPelaHierarquia(\mathcal{S}')$
  fim para
fim função

função $buscarCurriculoNoGrupo(\mathcal{S})$
  $\mathcal{C}^* \leftarrow \emptyset$
  para cada $\mathcal{C} \in CurriculosCSV$ **faça**
    se $\mathcal{C}[Node\_Nome\_Coluna(\mathcal{S})] = Node\_Valor\_Celula\_Coluna(\mathcal{S})$ **então**
      $\mathcal{C}^* \leftarrow \mathcal{C}^* \cup \{\mathcal{C}\}$
    fim se
  fim para
  retorno $\mathcal{C}^*$
fim função

função $gerarArquivosEntradaScriptlattes(\mathcal{C}^*, \mathcal{S})$
  $gerarArquivoList(\mathcal{C}^*)$ ▷ Gerar arquio de entrada do ScriptLattes *list*
  $gerarArquivoConfig(\mathcal{S})$ ▷ Gerar arquio de entrada do ScriptLattes *config*
fim função

A função $computarScriptPelaHierarquia$ faz uma iteração recursiva em pré-ordem para todos os grupos da hierarquia. Inicialmente essa função faz o processamento do grupo chamando as funções *buscarCurriculoNoGrupo* e $gerarArquivosEntradaScriptlattes$. Caso esse grupo contiver filhos em sua hierarquia, esta função é chamada recursivamente até o vértice filho. O grupo $\mathcal{S}$ é passado como parâmetro para a função *buscarCurriculoNoGrupo*, assim sendo, em alto nível, essa função faz a iteração no conjunto de currículos representados no arquivo *curriculos.csv* ($CurriculosCSV$), em cada iteração é realizada uma comparação do valor do parâmetro do currículo $\mathcal{C}$ $Node\_Nome\_Coluna$, com o valor do parâmetro de $Node\_Valor\_Celula\_Coluna$ de $\mathcal{S}$. Esses valores serão retornados pelo vetor $\mathcal{C}^*$, se, e somente se, forem iguais, ou seja, se estiverem associados ao mesmo grupo. A função $gerarArquivosEntradaScriptlattes$, recebe o conjunto $\mathcal{C}^*$ e $\mathcal{S}$, e gera os arquivos de entrada *list* e *config* do *ScriptLattes* para os dados informados.

4.1.3 Geração de relatórios

Após executar o módulo de geração de *script's*, o próximo passo é gerar os relatórios para cada grupo da hierarquia. É neste módulo que se tem uma interação com a ferramenta *ScriptLattes*, pois, para cada grupo da hierarquia, será feita uma execução, passando as informações contidas no grupo (arquivos *list* e *config*) para o *ScriptLattes*.

Uma grande vantagem do *ScriptLattes* é permitir extração de um conjunto de currículos da plataforma Lattes, e gerar relatórios de períodos específicos. Com isso, ao ser informado os anos para extração no arquivo *curriculos.csv*, estas informações são passadas para o arquivo de entrada *list* do StriptLattes, gerando um relatório específico por período de cada grupo da hierarquia.

Outra vantagem do *ScriptLattes*, está na geração dos relatórios em formato *HTML*. Isso facilita a integração com outros sistemas. Neste módulo toda hierarquia é representada em formato *html*, com *link's* para os relatórios gerados após execução do *ScriptLattes* em cada agrupamento.

4.2 Aplicação do método proposto para cadastro da plataforma Sucupira e E-Mec

Os programas de pós-graduação *stricto sensu*, após serem avaliados pela CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior) ao término de um período de quatro anos, devem novamente ser submetidos à prestação de contas, afim de que se determine se tais programas continuam a atender as exigências indispensáveis para se manterem em atividade. O processo de prestação de contas é iniciado através do cadastro das informações do programa na plataforma Sucupira(plataforma *online*).

As informações necessárias para o cadastro dizem respeito aos dados pessoais, acadêmicos e profissionais de cada discente, docente e egresso, assim como diversas informações sobre as produções e orientações. A plataforma Lattes concentra a maior parte dessas informações exigidas pela CAPES, auxiliando, com isso, os gestores dos programas a fazerem os levantamentos necessários para realizar este cadastro.

Sendo assim, utilizar ferramentas para extrações dos dados, tal como o *ScriptLattes*, se tornou algo essencial para o levantamento de informações de um programa. Pensando nisso, foi criado uma adaptação da hierarquias de currículos, apresentada neste trabalho, a fim de criar grupos de programas de pós-graduação *stricto sensu*.

Esta adaptação foi implementa por meio de um *script* denominado *MultiScriptLattes*. O referido *script* foi desenvolvido para auxiliar os gestores dos programas de pós-graduação *stricto sensu* a levantarem as informações cadastradas na plataforma Lattes, de determinado período. Para executar o *MultiScriptLattes*, o utilizador poderá montar a hierarquia da forma que desejar, sempre respeitando os dados iniciais de entrada desse *script*. A

Figura 4.5, demostra uma possível combinação da hierarquia, com foco em programas de pós-graduação *stricto sensu*.

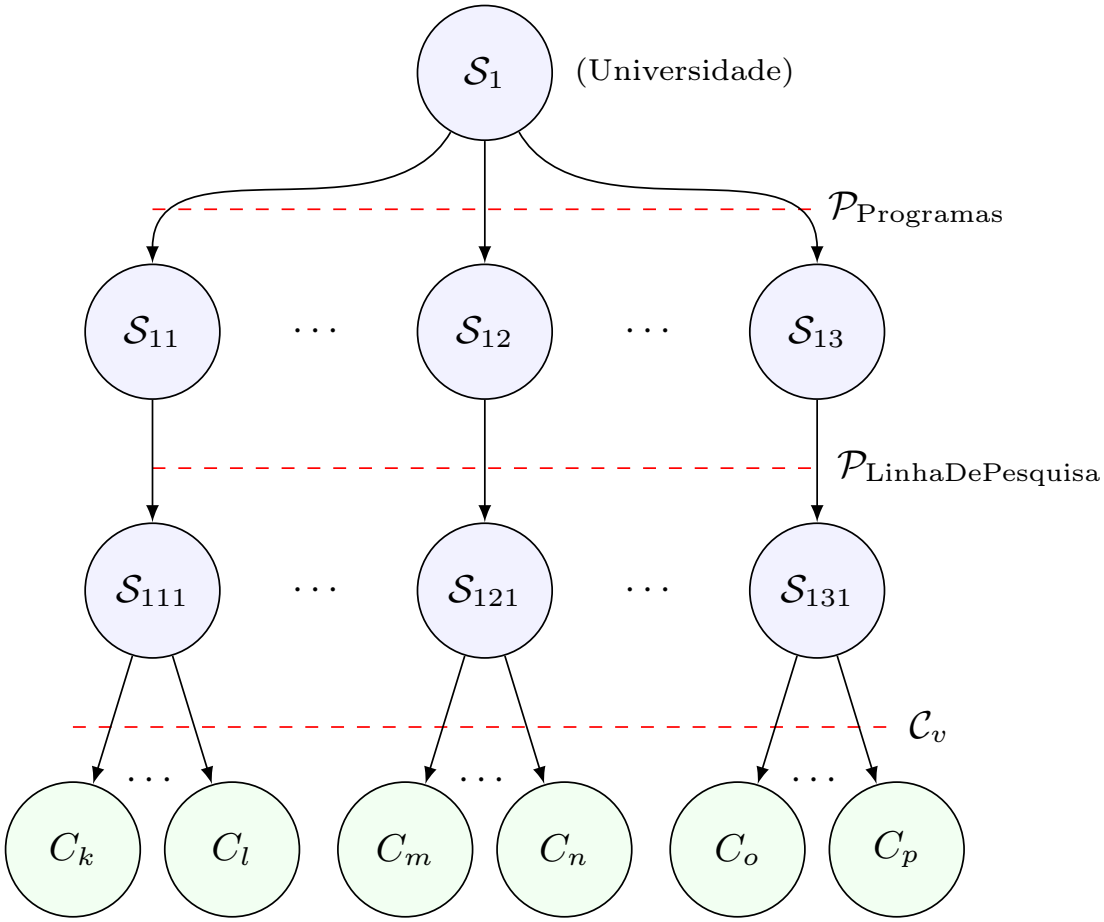


Figura 4.5: *Processo para construir hierarquia de currículos*

Fonte : Próprio autor

A primeira versão da ferramenta *MultiScriptLattes* foi desenvolvida em linguagem de programação *python*, com o conceito de *software* livre. Sua estrutura contém 3 módulos, similares aos apresentados na implementação de referência de hierarquia de currículos, cumprindo mencionar também, as adaptações realizadas em sua etapa de parametrização da hierarquia.

4.2.1 Configuração do *MultiScriptLattes*, para cadastro da plataforma Sucupira

As adaptações feitas pelo *MultiScriptLattes* na etapa de parametrização da hierarquia, foram significativas para implementação das extrações dos dados de programas de pós-graduação *stricto sensu*. Primeiramente, o arquivo *curriculos.csv*, foi transformado em outros dois arquivos, sendo: *docentes.csv* e *discentes.csv*. Para o arquivo *config.txt*, foram

acrescentado parâmetros dos diretórios dos novos arquivos, e incluído na hierarquia um parâmetro para o arquivo Qualis da área. Na Figura 4.6 é apresentado o processo de execução do *MultiScriptLattes*.

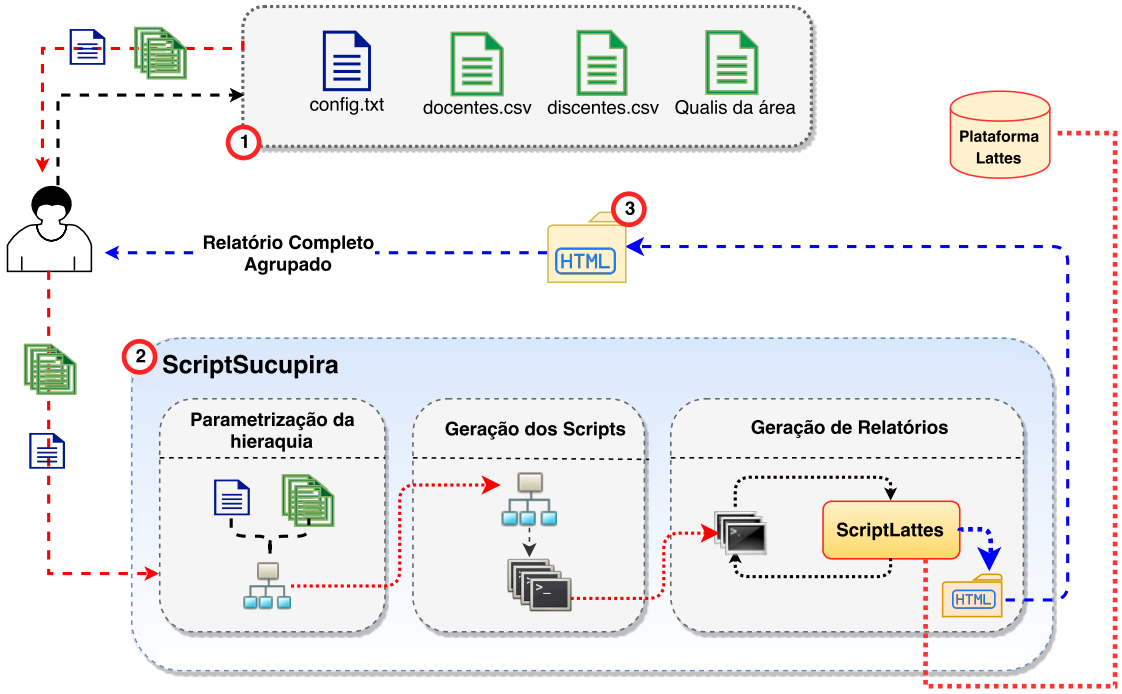


Figura 4.6: *Processo de execução* MultiScriptLattes.

Fonte : Próprio autor

O arquivo *docentes.csv*, foi criado com base no arquivo *curriculos.csv*, contando com alguns cabeçalhos a mais, tais como: caraga horária (*CH*), matricula (*MATRICULA*), *email* (*MAIL*) e categoria (*CATEG_n*). A coluna categoria, contém a informação que atesta se um docente é permanente (P) ou colaborador (C). Também foram adicionado informações relacionadas ao programa, como linha de pesquisa e responsáveis pelo programa.

Assim como o arquivo *docentes.csv*, foi criado o arquivo *discentes.csv*, que contempla os dados dos discentes de um determinado programa. Este arquivo deve conter o cabeçalho *ID_Lattes*, *Nome*, pois representa, a descrição das colunas no arquivo. Assim como é feito quanto aos docentes, é necessário fazer a busca do ID Lattes de cada discentes informado no arquivo, assim como o programa e o ano que estes discentes cursaram ou estão cursando.

Para facilitar a geração dos arquivos *discentes.csv* e *discentes.csv*, foi criado uma planilha padrão (*template*) e disponibilizada de forma *online* e gratuita através do *link*:

https://goo.gl/uFfzFt

.

Com essa planilha é possível cadastrar as informações docentes e discentes de um programa de pós-graduação *stricto sensu* e fazer o *download* dos arquivos *discentes.csv* e *discentes.csv*, já preenchidos. Esta planilha foi construída para facilitar o utilizador a gerir as informações que serão utilizadas para extração. Um ponto positivo da planilha em questão é a validação do ID Lattes, que é feita no momento da geração dos arquivos. Além da validação do ID Lattes, também são validados os nomes corretos dos programas informados na coluna *Programa_n*, assim como o *Ano*, tanto dos docentes quanto dos discentes.

Tanto o arquivo *docentes.csv* quanto o *discentes.csv*, utilizam os valores dos parâmetros*Node_Nome_Coluna_n* e *Node_Valor_Celula_Coluna_n* para cruzar os dados com a hierarquia construída no arquivo *config.txt*, seguindo o mesmo procedimento de associação já informado. Os valores devem estar sincronizados para que as informações dos docentes e discentes estejam relacionados ao mesmo grupo na hierarquia.

No arquivo *config.txt* foram adicionados 3 parâmetros pertinentes à construção da hierarquia, sendo: diretório do Qualis da área em formato *csv* (*Node_Arq_Qualis_Periodido_csv_n*) e em *pdf* (*Node_Arq_Qualis_Periodido_pdf_n*) e área CAPES do programa (*Node_desc_Qualis_Periodido_n*). A forma como a hierarquia é estruturada não teve alteração, mantendo-se a criação incremental através do sufixo. Outra alteração neste arquivo, foi a inclusão dos parâmetros referentes ao apontamento para os diretórios dos arquivos *discentes.csv* (*dataset_doscentes*) e *discentes.csv* (*dataset_discentes*). Na Figura 4.7 é apresentado o arquivo *config.txt* com as alterações citadas e com a hierarquia construída.

```
# Configurações do ScriptLattes
diretorio_de_cache_dos_cvs: ./cache/uninove/ppgi/cache
diretorio_de_cache_dos_doi: ./cache/uninove/ppgi/doi
modelo_programa: ./modelos/PR-Programas.config

#Configurações do arquivo curriculos.csv
dataset_doscentes: ./docentes.csv
dataset_discentes: ./discentes.csv
separador_csv: |

#Criação da hierarquia
Node_Nome_Coluna_1 = Universidade
Node_Valor_Celula_Coluna_1 = UNINOVE
Node_Descricao_1 = Universidade Nove de Julho
Node_Modelo_Config_Scriptlattes_1 = ./modelos/template.config
Node_Nome_Coluna_11 = Exatas
Node_Valor_Celula_Coluna_11 = programasExatas
Node_Descricao_11 = Pós-graduação stricto sensu em Exatas
Node_Modelo_Config_Scriptlattes_11 = ./modelos/template.config

Node_Nome_Coluna_111 = ProgramasPPGI
Node_Valor_Celula_Coluna_111 = PPGI
Node_Descricao_111 = Stricto Sensu em Informática e Gestão do Conhecimento
Node_Modelo_Config_Scriptlattes_111 = ./modelos/template.config
Node_Arq_Qualis_Periodido_csv_111 = ./qualis/inter_2016-2017.csv
Node_Arq_Qualis_Periodido_pdf_111 = ./qualis/inter_2016-2017.pdf
Node_desc_Qualis_Periodido_111 = Interdisciplinas

Node_Nome_Coluna_12 = Saude
Node_Valor_Celula_Coluna_12 = programasSaude
Node_Descricao_12 = Pós-graduação stricto sensu em Saúde
Node_Modelo_Config_Scriptlattes_12 = ./modelos/template.config

Node_Nome_Coluna_121 = ProgramasPPM
Node_Valor_Celula_Coluna_121 = PPM
Node_Descricao_121 = Pós-graduação stricto sensu em Medicina
Node_Modelo_Config_Scriptlattes_121 = ./modelos/template.config
Node_Arq_Qualis_Periodido_csv_111 = ./qualis/saude_2016-2017.csv
Node_Arq_Qualis_Periodido_pdf_111 = ./qualis/saude_2016-2017.pdf
Node_desc_Qualis_Periodido_111 = Medicina II
```

Figura 4.7: *Configurações dos parâmetros do arquivo* config.txt *para adaptação do* MultiScriptLattes

Fonte : Próprio autor

Com essa adaptação é possível gerar as hierarquias e executar os próximos módulos de geração de *script's* e geração de relatórios para os grupos da hierarquia. Esta adaptação fora utilizada em trabalhos recentes, na próxima sessão é apresentado um resumos dos trabalhos que utilizaram o *MultiScriptLattes*, para criação de hierarquia e extração dos dados de programas de pós-graduação *stricto sensu*.

4.2.2 Resultados utilizando a configuração do *MultiScriptLattes* para cadastro da plataforma Sucupira

Foi utilizado por NIGRO et al. (2015) esta implementação para extração, organização e apresentação dos dados existentes na plataforma Lattes, visando tornar pública a produção científica de um programa de pós-graduação *stricto sensu* em Direito, com base no requerido pelos avaliadores da CAPES. O resultado desta pesquisa facilitou o processo de prestação de contas, contribuindo com a transparência da produtividade do programa, possibilitando aos gestores do curso citado acompanhar em tempo quase real, o desempenho do programa e dos professores sob sua gestão. Esta pesquisa, possibilitou estabelecer estratégias para alavancar a produção científica ainda no quadriênio vigente, afastando a possibilidade da ocorrência de uma avaliação insatisfatória.

De forma a extrair a produção científica e acadêmica de um grupo de professores integrantes de um programa de pós-graduação *stricto sensu* em Engenharia de Produção, Nigro et al. (2017) utilizaram este *script*, tendo como resultado a disponibilização pública dos dados do programa em questão.

No trabalho, objeto deste tópico, Ferraz, Quoniam e Maccari (2014) utilizaram este método para extrair as produções científicas, técnicas e tecnológicas, atuação acadêmica e relações entre professores pesquisadores pertencentes ao departamento de pós-graduação em Administração da UNINOVE (Universidade Nove de Julho). De acordo com o resultado desta pesquisa, utilizar esta implementação, trouxe um enriquecimento satisfatório no que tange, a concentração dos dados de produção acadêmica.

4.3 Configuração do *MultiScriptLattes*, para cadastro da plataforma e-Mec

O Conselho Nacional de Educação e Instituição (CNE) fora criado através da Lei n 9.131/95, assim como a instituição do Exame Nacional de Curso (ENC), após a Lei 9394/96, que formaliza as diretrizes da educação nacional, o Ministério da Educação e Cultura (MEC), passou a priorizar a avaliação institucional. Com isso, em parceria com o Estado e Instituições de Ensino, assumiu um papel de regulamentador, com foco em excelência e qualidade do ensino (CATANI; OLIVEIRA, 2002).

A plataforma e-Mec é um sistema online de processos que regulamenta a educação

superior no Brasil. Os pedidos de credenciamento e recredenciamento de instituições, autorização, reconhecimento e renovação de cursos são administrados pelo e-Mec. Esta plataforma é a base oficial e única de IES (Instituições de Educação Superior). Os dados são armazenados em conformidade com os atos autorizados, e editados com base na Portaria Normativa MEC n 40/2007.

As IES por sua vez, devem prestar contas em um período trienal, de seus cursos de graduação, através do cadastro na plataforma e-Mec. O *MultiScriptLattes*, é uma adaptação para facilitar o levantamento de dados, se tornando um opção para as instituição a reunir informações de seus cursos.

Esta implementação do *script MultiScriptLattes*, não contem informações de discentes e de *Qualis* de artigos. Para facilitar a criação dos arquivos de entrada, de maneira similar o *script*, foi uma planilha padrão utilizando o *Google Sheeats*, que faz a validação do ID Lattes de um docente e da relação dos cursos de graduação informado, a mesma poderá ser acessada pelo *link*:

https://goo.gl/pSTPxo

Na Figura 4.8 podemos ver uma possível hierarquia utilizando o *MultiScriptLattes*, configurado para levantamento dos dados para cadastro da plataforma e-Mec, com grupos de universidades e cursos.

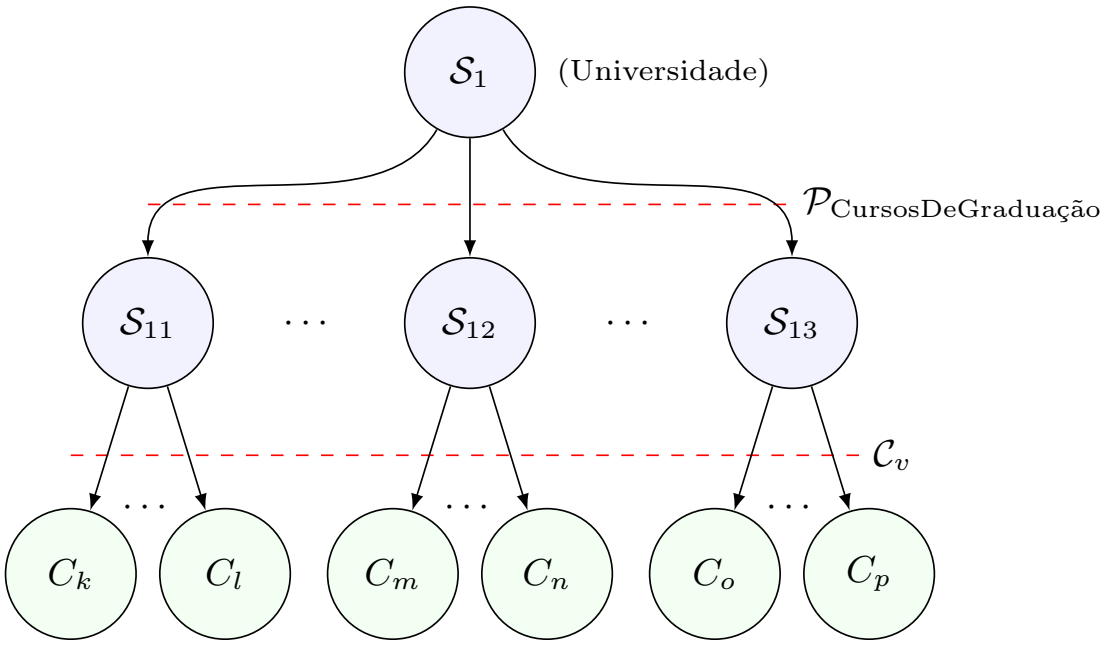


Figura 4.8: *Processo para construir hierarquia de currículos*

Fonte : Próprio autor

Esta configuração do *MultiScriptLattes* recentemente fora utilizado para criações de hierarquias, na próximo sessão segue resumos dos trabalhos que utilizam esse *script*.

4.3.1 Resultados utilizando a configuração do *MultiScriptLattes* para cadastro da plataforma e-Mec

Neste trabalho, Bavaresco (2017) utilizou esta configuração do *script* para extração dos dados de cursos de uma Instituição de Ensino Superior (IES) privada. Esta pesquisa foi a percursora em aplicar o conceito disposto por esta ferramenta na esfera de cursos de graduação. Com objetivo de concentrar as informações para o gerenciamento e acompanhamento, foi possível analisar as produções do corpo docente, chegando a conclusão que, determinados docentes deverão ser incentivados a apresentar produtividade para o próximo triênio, de 2017 a 2019. O ponto central desta pesquisa, está na favorabilidade de utilização deste *script*, que por ser uma ferramenta de uso livre, poderá ser utilizada em qualquer IES, privada ou não.

Tendo em vista, auxiliar uma IES do setor privado, na gestão de desempenho dos cursos de graduação, Silva (2017) propôs no presente trabalho, aplicação de uma campanha para conscientização do corpo docente, em manter os dados atualizados na plataforma Lattes. Utilizando esta configuração da ferramenta *MultiScriptLattes*, antes e depois de uma campanha de conscientização, demonstrou-se uma evolução nos indicadores avaliativos, concluindo-se que, utilizar esta ferramenta em conjunto de companhas de conscientização acarretaram em resultados congruente, no que tange, gestão de indicadores de produtividade acadêmica.

4.4 Método de Agrupamentos Hierárquicos aplicado em programas de pós-graduação *stricto sensu*

No Brasil, o sistema de avaliação da pós-graduação *stricto sensu* é considerado referência no que tange modernidade e eficiência, sendo de grande vália para o desenvolvimento da ciência e tecnologia no país (MACCARI et al., 2009). Este processo de avaliação é realizado em um período quadrienal, e são avaliados quesitos como: corpo docente, teses e dissertações, produção intelectual, produções técnicas, inserção social e proposta do programa.

Para um programa pós-graduação *stricto sensu* se manter em atividade, deve-se obter o conceito igual ou superior a 3, tendo 7 como avaliação máxima. Se o conceito obtido na avaliação quadrienal pelo o programa, for inferior a 3, a CAPES não mais o recomenda, prejudicando sua imagem e reputação.

Criada em 1999, a grande área multidisciplinar foi reformulada em 2008, se tornando área interdisciplinar. Desde 2001, foram feitas cerca de 6 avaliações trienais, sendo: 2001, 2004, 2007, 2010, 2012 e 2016. Após a avaliação realizada em 2012, o processo avaliativo passou a ser aplicado em um período quadrienal.

De acordo com a última avaliação feita em 2016, a câmera temática Engenharia/Tec-

nologia/Gestão da área interdisciplinar, passou a ter 54 programas em funcionamento, sendo:

a) 4 programas de mestrado e 1 doutorado com nota 2, ver Tabela 5.1.

b) 19 programas de mestrado com nota 3, ver Tabela 5.2.

c) 16 programas de mestrado e 21 doutorado com nota 4, ver Tabela 5.3.

d) 5 programas de mestrado e doutorado com nota 5, ver Tabela 5.4.

e) 4 programas de mestrado e doutorado com nota 6, ver Tabela 5.5.

f) Não temos nenhum programa com nota 7 para câmera temática Engenharia/Tecnologia/Gestão.

A Figura 4.9, demonstra a soma dos programas de mestrado e doutorado para cada conceito desta câmera temática.

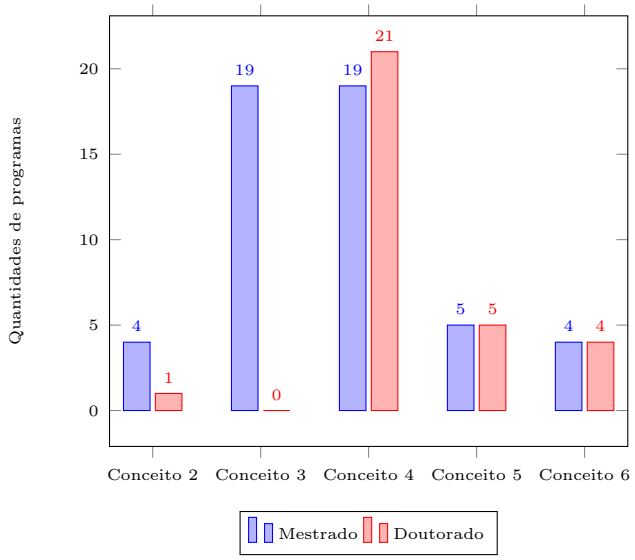


Figura 4.9: *Quantidades de Programas por nota Capes*
Fonte : Próprio autor

Sendo assim, identificar o quão distante um determinado programa está de outro da mesma área de avaliação, é uma opção positiva para gestão de um programa durante um período. Nesta dissertação iremos apresentar medições por similaridade de programas da mesma área de avaliação, através dos indicadores CAPES. Primeiramente será apresentado uma seção sobre criação dos dados, e logo após, falaremos sobre a geração de agrupamentos hierárquicos sobre os dados destes 49 programas.

4.4.1 Coleta dos dados

Tendo em vista aplicar o conceito de agrupamento hierárquico nos dados câmera temática Engenharia/Tecnologia/Gestão da área interdisciplinar, foi realizado um levantamento das informações de cada programas de pós-graduação *stricto sensu*, assim como o conceito obtido por cada programa no quadriênio de 2013 à 2016. Os dados dos programas estão disponíveis, no plataforma Sucupira, de forma online e gratuita, e podem ser extraídos em formato *xlsx* através do *link*:

https://sucupira.capes.gov.br/sucupira/public/consultas/coleta/programa/listaPrograma.jsf

.

Para esta pesquisa, foram utilizados os filtros de busca *Área Avaliação* como Interdisciplinar e *Área Básica* como Engenharia/Tecnologia/Gestão. Com isso foi possível encontrar os programas em atividade da câmera temática Engenharia/Tecnologia/Gestão da interdisciplinar. No entanto, tais informações não eram o bastante para conseguirmos extrair os dados dos programas, foi preciso obter os dados dos docentes, tanto permanentes como colaboradores.

Para buscar estas informações, na plataforma Sucupira, existe uma área específica contendo dados estatísticos dos programas cadastrados. Para ter acesso a esses dados, o interessado deverá se cadastrar na plataforma, informando seus dados pessoais, não sendo necessário entrar em contato com CAPES para solicitar uma permissão de acesso. Dentro da área restrita podemos consultar diversos relatórios, tais como: relatórios de pagamentos de bolsas e auxílios, dados de discentes, programas, bem como os dados de docentes, dados estes de maior interesse para os fins do presente trabalho. No *link* abaixo segue o endereço da área restrita.

http://dadosabertos.capes.gov.br/dadosabertos/login.do

O nome do relatório de docentes contido na área restrita é Coleta de Dados, *Docentes de Pós-Graduação stricto sensu no Brasil 2015*. Este relatório teve sua última modificação em 13 de setembro de 2017, e possui o ID CAPES de referencia *BR-CAPES-COLSUCUP-DOCENTE2015-2016-03-01*. Neste relatório contém a relação de docentes por programa de pós-graduação *stricto sensu*, além de informações específicas dos docentes, como: formação, área de formação, região, entre outras. Um ponto negativo desse relatório, está em não fornecer os dados por câmera temática, ou seja, não é possível relacionar os docentes, de determinados programas, da câmera temática específica de uma área, sem utilizar outro relatório, tal como o citado acima, extraído através do plataforma Sucupira.

Sendo assim, foi necessário unificar os dados dos relatórios de programas com os de docentes, e gerar um novo relatório de programas por câmera temática de uma área de

avaliação. Após a criação desse relatório unificado, o próximo passo, foi fazer a busca dos ID Lattes de cada currículo dos docentes contidos nesse relatório. Para isso criamos um *script*, em linguagem de programação *python* para fazer buscas automáticas, direto na plataforma Lattes. Ao todo, foram extraídos mais de 6.000 IDs. Este *script* auxiliou a buscar em torno de 90% dos IDs, os outros 10%, fora feita de forma manual, devido as inconsistências dos nomes dos docentes cadastrados na plataforma Lattes, com os cadastrados na plataforma Sucupira.

Com os dados necessários para construção dos arquivos *config.txt* e *curriculos.csv*, fora feita uma configuração simples da ferramenta *MultiScriptLattes*, apresentado no começo deste capítulo. Esta configuração permitiu extrair os dados de 49 dos programas desta câmera temática. Na Figura 4.10 são apresentados alguns dos programas dispostos na hierarquia criada.

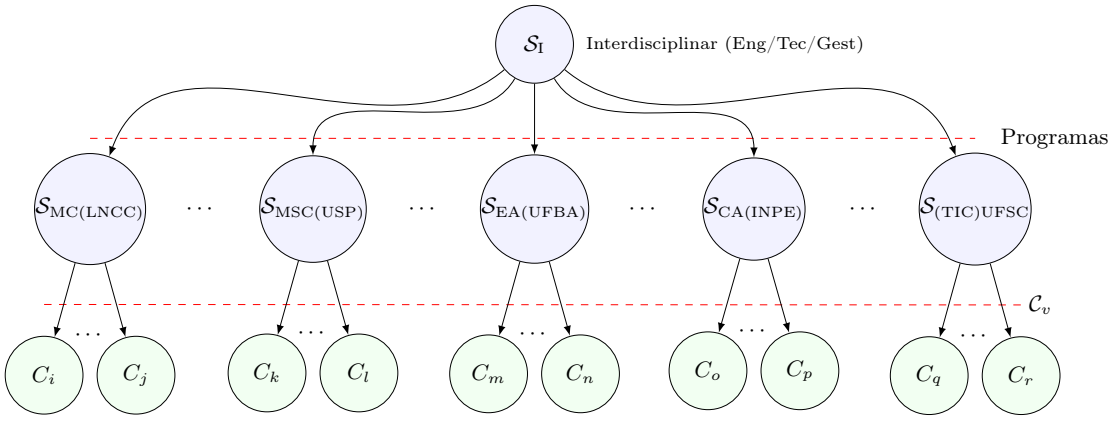


Figura 4.10: *Hierarquia de currículos dos programas câmera temática de Engenharia/Tecnologia/Gestão*

Conforme já explicitado, ao todo foram extraídos os 49 programas de pós-graduação *stricto sensu*, no período de 2013, 2014, 2015 e 2016, assim como um consolidado de todo o quadriênio 2013 à 2016. Não foi possível extrair os 54 programas dessa câmera temática. Na Tabela 4.3 descreve detalhadamente os motivos pelos quais restaram inviabilizadas as extrações de desses 5 programas. faltantes.

Todos os relatórios resultantes dessa execução, foram disponibilizados de forma pública, a fim de contribuir com uma base de dados para futuras pesquisas. Abaixo, segue o link de acesso aos relatórios disponibilizados:

https://goo.gl/PSA4aL

Tabela 4.3: *Programas não extraídos*

Programa	Instituição	Motivo
Propriedade Intelectual e Inovação	INPI	Não foi encontrado registro de docentes na plataforma Sucupira
Engenharia e Gestão da Inovação	UFABC	Não foi encontrado o programa na plataforma Sucupira
Modelagem e Métodos Quantitativos	UFC	Não foi encontrado registro de docentes na plataforma Sucupira
Nanociência e Nanotecnologia	UNB	Não foi encontrado registro de docentes na plataforma Sucupira
Ciências Físicas Aplicadas	UECE	Não foi encontrado registro de docentes na plataforma Sucupira

4.4.2 Construções dos agrupamentos hierárquicos

Com intuito de criar agrupamentos hierárquicos dos 49 programas extraídos, foram utilizados, como parâmetros, os indicadores CAPES utilizados na avaliação dos programas do quadriênio, para calcular as distâncias entre os programas, de 2013 a 2016. Isto possibilitou a utilização do conceitos de agrupamento hierárquico a fim de medir a similaridade entre os programas.

Embora hajam questões a serem discutidas sobre a real efetividade do sistema de avaliação da CAPES (SGUISSARDI, 2006), os cadernos de indicadores de área da CAPES, são utilizados para descrever a forma como serão avaliados os programas de determinadas áreas. Com isso, cruzar as informações de um programa com o caderno de indicadores auxiliam os gestores a se anteciparem a possíveis avaliações negativas no quadriênio.

Pensando nisso, utilizamos 4 indicadores para comparar por similaridade os 49 programas, sendo que 2 destes indicadores, foram retirados do caderno de avaliação do ano de 2016, e outros 2, que foram, inspirados neste caderno. Esta comparação é útil para verificar o quão distante um programa está de outro, que contém um conceito CAPES inferior ou superior. Para facilitar essa comparação, criamos um *script*, determinado por *ScriptComp* que implementa o conceito de agrupamento hierárquico, proposto nesse trabalho, sobre os relatórios gerados pelo *ScriptLattes* de cada programa.

Na seção anterior geramos um grande massa de dados utilizando uma adaptação da hierarquia de currículos. Com isso foram gerados diversos relatórios para cada programa. Como já consignado acima, estes relatórios são gerados através da execução do *ScriptLattes* para cada grupo da hierarquia. Sendo assim, o *ScriptComp*, utiliza dos relatórios gerados pelo *ScriptLattes*, a fim de comparar os programas por similaridade utilizando os dados extraídos da plataforma Lattes, para um período específico.

Com os relatórios dos programas gerados, o *ScriptComp* utiliza os arquivos de sufixo **database.xml* e **publicacoesPorMembro.csv*, gerados pelo ScritpLattes, para unificar as

informações de publicações, orientações, publicações qualificadas, produções técnicas e artísticas, participações em eventos e projetos de pesquisas dos currículos dos docentes. Esta operação se tornou viável, pois os arquivo **database.xml* e **publicacoesPorMembro.csv*, estão estruturas em formatos **xml* e *csv*. Isso facilitou a transformação destes dados em dados orientados a objetos pelo *ScriptComp*.

Com todos os programas transformados em objetos, é possível ,através dos dados, calcular os valores dos indicadores para cada programa, independente do período. O *ScriptComp* está adaptado para fazer o cálculo dos indicadores por ano, ou seja, é possível analisar a distância de um programa de conceito 6, comparado com outro de conceito 4 num determinado ano.

O *ScriptComp* se tornou uma opção *opensource* para a gestão de um programa, ter a opção de comparar um determinado programa com outro da mesma área de avaliação, auxilia os gestores a tomarem decisões sobre o andamento e rumo do programa. Este *script* está disponível no *link*:

link: https://goo.gl/PSA4aL

A fim de medir a similaridade entre os programas, como dito, foram eleitos 4 indicadores para execução do *ScriptComp* sobre estes relatórios, sendo: produção de artigos qualificados (indProdArt), produção de artigos qualificados até $B1$ (IndProdSup), quantidade de orientações de mestrado e doutorado (IndProdOrient) e produtividade do programa (indProd). Como descrito na metodologia proposta, o resultado do agrupamento hierárquico é apresentado através de um dendrograma.

No dendrograma a descrição do eixo horizontal (x), separado por dois-pontos, contém o conceito CAPES obtida no quadrimestre de $2013 - 2016$, e o nome das universidades que pertencem aos 49 programas. Se a universidade tiver mais de um programa para área analisada, foi inserido o caractere para separar as as siglas dos programas, tal como: 4:UNINOVE, 6:LNCC, 6:UERJ_MC e 6:UERJ_CC. No eixo vertical (y), é definido como distância real, a distância obtida na execução do algoritmo de agrupamento hierárquico. No eixo central de cada agrupamento, é descrito o média dos valores obtidos no cálculo do indicador em questão, de modo que é possível conseguir identificar em uma escala do menor para o maior, o desempenho de um grupo em comparação com outro.

Para para medir a similaridade no algoritmo hierárquico, fora utilizado como distância no *ScriptComp* a ligação por média. Ao utilizar a distância por média, é possível descobrir com mais facilidade a média entre os grupos gerados. Na próxima seção será apresentado o resultado do *ScritpComp* para cada um dos 4 indicadores citados.

4.4.2.1 Agrupamento Hierárquico pelo indicador IndProdArt

Este indicador mede toda produção intelectual de um programa em formato de artigos científicos, publicados e qualificados de um programa da área interdisciplinar, através da

seguinte equação:

$$indProdArt = \sum_{j=1}^{qtdDocentes} (1 \times A1_j + 0,85 \times A2_j + 0,7 \times B1_j + 0,55 \times B2_j + 0,4 \times B3_j + 0,25 \times B4_j + 0,1 \times B5_j)/DP \quad (4.1)$$

onde $A1$, representa um artigo classificado com Qualis $A1$;
$A2$ representa um artigo classificado com Qualis $A2$;
$B1$ representa um artigo classificado com Qualis $B1$;
$B2$ representa um artigo classificado com Qualis $B2$;
$B3$ representa um artigo classificado com Qualis $B3$;
$B4$ representa um artigo classificado com Qualis $B4$;
$B5$ representa um artigo classificado com Qualis $B5$;
DP é a quantidade de docentes permanentes no programa.

A Figura 4.11 demonstra similaridade entre os programas para o indicador IndProdArt do ano 2016, através de um dendrograma.

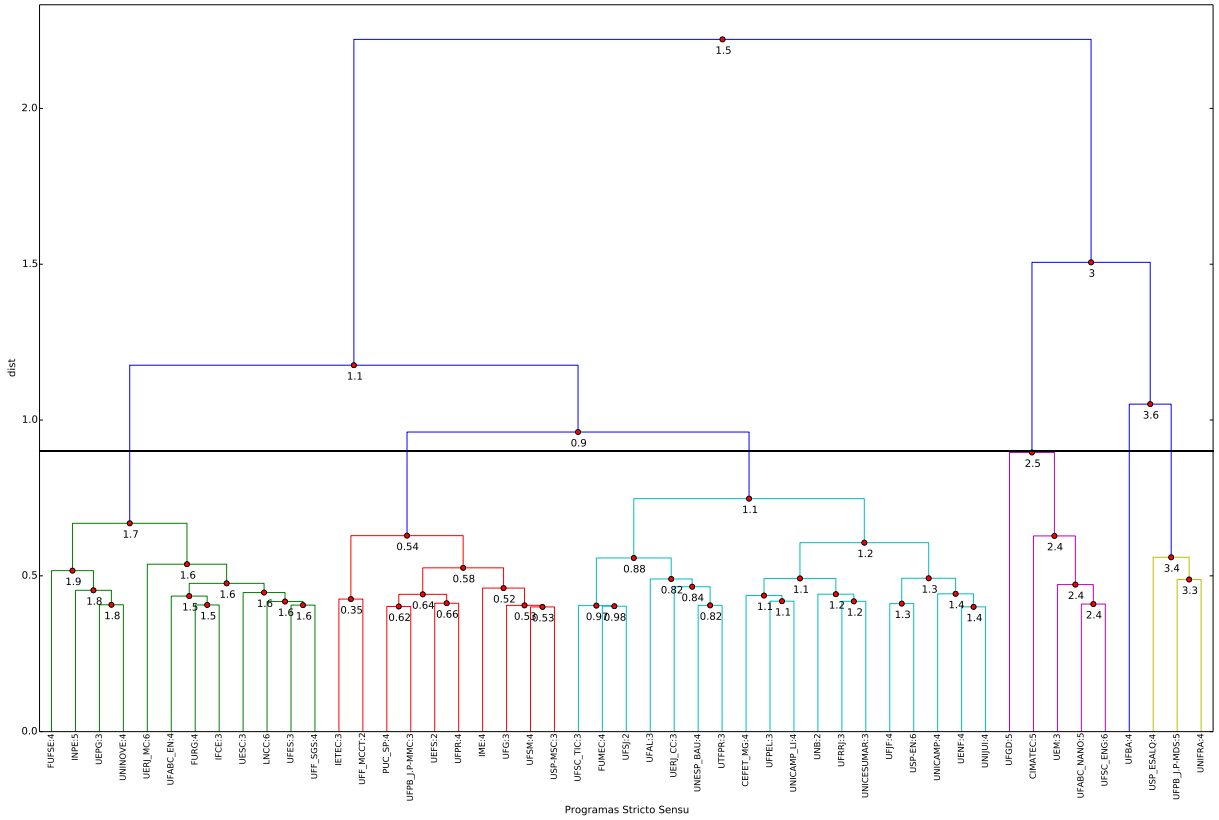


Figura 4.11: *Resultado da execução do* ScriptComp *com corte na hierarquia, para o indicador IndProdArt*

Fonte : Próprio autor

De acordo com a Figura 4.11, a extração da média aritmética do conceito CAPES, de cada programa nos grupos, é uma possibilidade válida de análise, sendo possível, com isso, verificar a produtividade desse grupo em comparação com a média do conceito CAPES. Para isso foi definido um *corte*, na região do eixo x na distância 0.9, separando os programas em 6 grupos.

No eixo central de cada grupo, temos a média do indicador IndProdArt calculada. Portanto, se analisarmos os 6 grupos gerados após o *corte*, temos uma crescente do grupo com maior produtividade para o menor. Na Tabela 4.4 é demonstrada essa análise.

Tabela 4.4: *Análise dos programas após o agrupamento hierárquico - IndProdArt*

indProdArt por grupo	Programas por grupo	Média dos conceitos
3.8	1	4
3.4	3	5
2.5	5	3.6
1.7	12	3.5
1.1	18	3.5
0.54	10	3.2

Com isso podemos fazer as seguintes afirmações:

a) O grupo com valor de indProdArt 1.1, contém maior numero de programas, porém obteve o cálculo do indicador inferior aos grupos 1.7, 2.5, 3.4 e 3.6, sendo penúltimo grupo em comparação com os demais. Com isso podemos concluir que, embora este grupo possua mais programas, a média de produtividade obtida no cálculos do indicador indProdArt foi inferior a 4 grupos, ou seja, estes programas não produziram o suficiente comparado com os demais já informados.

b) O índice de maior produtividade para o indicador de artigos qualificados, pertence ao programas do grupo 3.6. Embora este programa não tenha o conceito CAPES 6, podemos concluir que obteve o melhor desempenho neste indicador, em comparação com os demais no ano de 2016.

c) O programa de pós-graduação *stricto sensu* em informática e gestão do conhecimento da universidade Nove de Julho, no ano de 2016, obteve um índice superior de produtividade de artigos qualificados, em comparação com os demais programas do agrupamento 1.1 e 0.54.

4.4.2.2 Agrupamento Hierárquico pelo indicador IndProdArtSUP

O indicador IndProdArtSUP avalia toda produção intelectual do programa em formato de artigos científicos, similar ao IndProdArt. Contudo, leva em consideração apenas os

artigos $A1, A2, B1$ através da seguinte equação:

$$indProdArtSUP = \sum_{j=1}^{qtdDocentes} (1 \times A1_j + 0,85 \times A2_j + 0,7 \times B1_j)/DP \quad (4.2)$$

onde $A1$, representa um artigo classificado com Qualis $A1$;
$A2$ representa um artigo classificado com Qualis $A2$;
$B1$ representa um artigo classificado com Qualis $B1$;
DP é a quantidade de docentes permanentes no programa.

O resultado do *ScriptComp* para este indicador no ano de 2016, pode ser verificado através da Figura4.12.

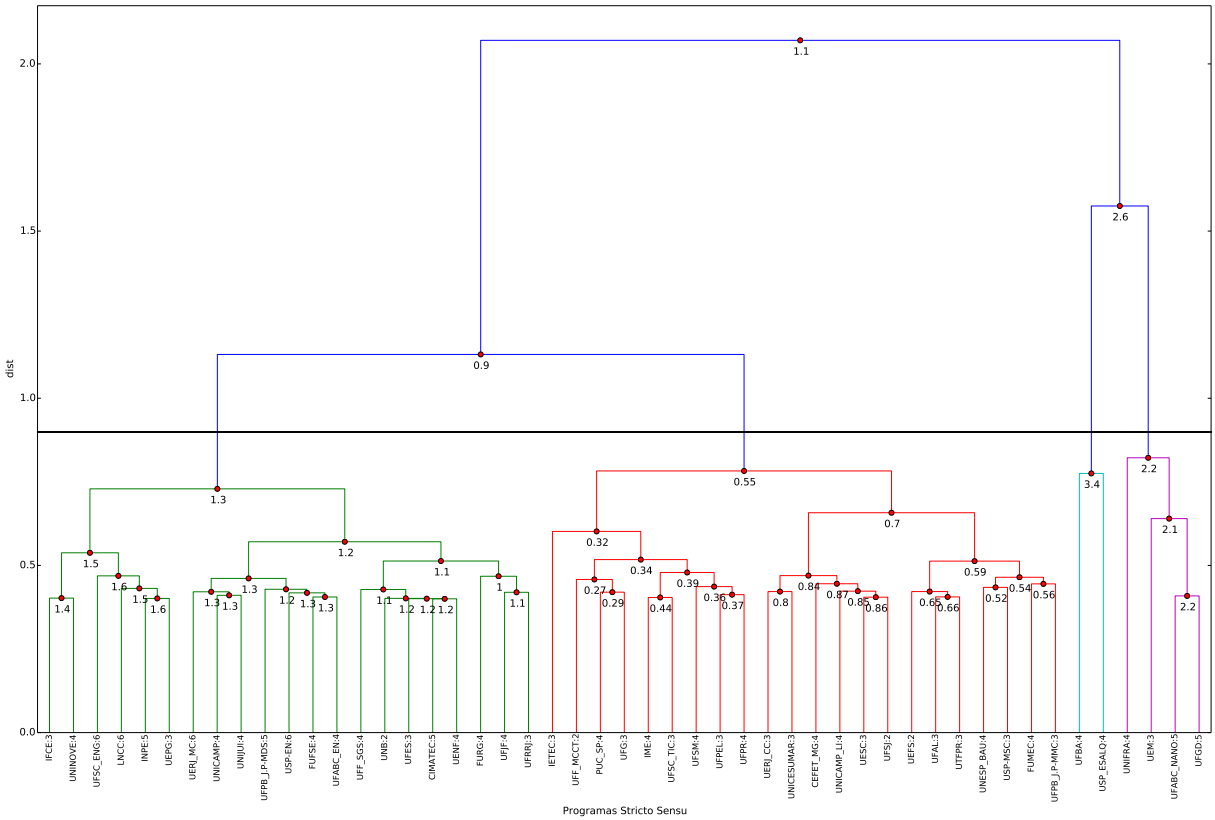


Figura 4.12: *Resultado da execução do* ScriptComp *com corte na hierarquia para o indicador IndProdArtSUP*

Fonte : Próprio autor

Na Figura 4.12 foi definido um *corte*, na região do eixo x na distância 0.8, separando os programas em 4 grupos. Com isso é possível analisar através do eixo central de cada grupo, seu desempenho para este indicador. Na Tabela 4.5 é apresentada a média do conceito CAPES, assim como a quantidade de programas em cada grupo.

Diante disso, podemos fazer as seguintes afirmações:

Tabela 4.5: *Análise dos programas após o agrupamento hierárquico - IndProdArtSUP*

indProdArtSUP por grupo	Programas por grupo	Média dos conceitos
3.4	3	4
2.2	4	4.25
1.3	21	4.2
0.55	23	3.4

a) O grupo com maior índice de produtividade de artigos qualificados - IndProdArtSUP, possui 3 programas, sendo a média dos conceitos CAPES igual a 4. Na Figura4.12 o grupo 3.4, não possui nenhum programa com o conceito CAPES superior a 5. Com isso podemos inferir que os programas pertencentes ao grupo 3.4, tiveram o índice de produtividade maior, do que os programas de conceito CAPES superiores a 5, para o ano de 2016.

b) Os programas relacionados ao grupo 0.55, tiveram o pior índice entre os programas analisados. Este grupo não possui nenhum programa com nota superior a 5.

c) O programa de pós-graduação *stricto sensu* em informática e gestão do conhecimento da universidade Nove de Julho, está no grupo 1.3, o qual possui, a nível de conceito CAPES, uma média superior a 4, ou seja, os programas desse grupo, pelo indicador indProdArtSUP estão se projetando acima do conceito 4 da CAPES.

4.4.2.3 Agrupamento Hierárquico pelo indicador IndProdOrient

O Indicador que avalia a quantidade de orientações de mestrado e doutorado, é expresso a partir da seguinte equação:

$$\begin{aligned} IndProdOrient \quad = \sum_{j=1}^{qtdDocentes} \quad & (qtdOriAndamentoMestrado + \\ & qtdOriAndamentoDoutorado + \\ & qtdOriConcluidoMestrado + \\ & qtdOriConcluidoDoutorado)/DP \end{aligned} \tag{4.3}$$

onde $qtdOriAndamentoMestrado$, representa a quantidade de orientações de mestrado em andamento;
$qtdOriAndamentoDoutorado$ representa a quantidade de orientações de doutorado em andamento;
$qtdOriConcluidoMestrado$ representa a quantidade de orientações de mestrado concluídas;
$qtdOriConcluidoDoutorado$ representa a quantidade de orientações de doutorado concluídas;
DP é a quantidade de docentes permanentes no programa.

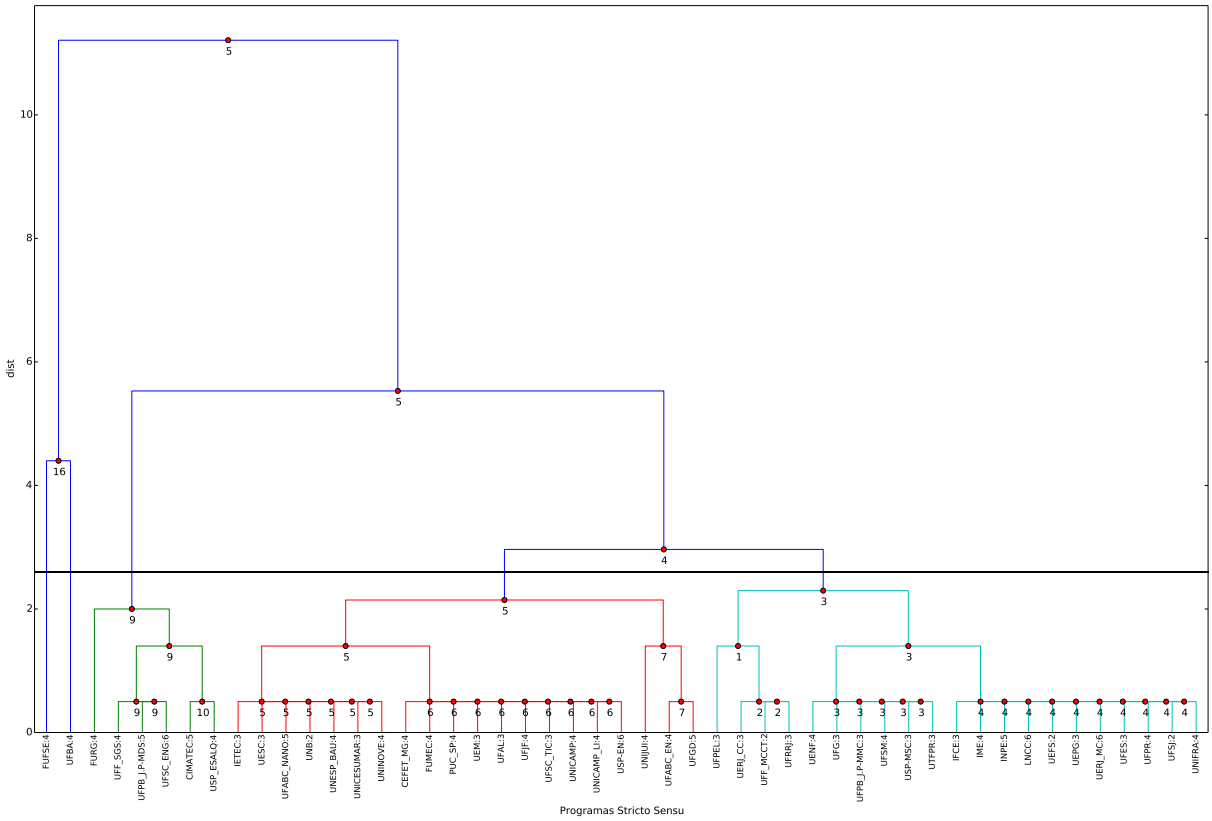


Figura 4.13: *Resultado da execução do* ScriptComp *com corte na hierarquia, para o indicador IndProdOrient*

Fonte : Próprio autor

Ao analisar os resultados desse indicador através da Figura fig:IndProdOrient podemos afirmar que:

a) Os programas pertencentes às universidades FFUSE, UFBA, FURG, UNIJUI e UFPEL tiveram um melhor desempenho no indicador de índice de orientações, do que os demais programas.

b) Dentre os programas que se destacaram, nenhum deles possuem conceito CAPES superior a 5, com isso podemos afirmar que, para o ano de 2016, estes programas obtiveram uma produtividade maior, neste indicador, do que os demais com conceitos superiores.

4.4.2.4 Agrupamento Hierárquico pelo indicador IndProd

O indicador IndProd é utilizado para avaliar toda produção intelectual de um programa. Para isso, o valor deste indicador é a soma dos indicadores de avaliação de livros, capítulos, produção técnica/tecnológica e produção intelectual, em formato de artigos científicos. Abaixo seguem as equações que compõem esse indicador.

a) IndProdLiv: Indicador que avalia toda produção intelectual do programa no formato de livros, sendo participação obrigatória de um docente permanente como autor, através da seguinte equação:

$$IndProdLiv = \sum_{j=1}^{qtdDocentes} liv_j/DP \tag{4.4}$$

onde liv representa os livros do programa;
DP é a quantidade de docentes permanentes no programa.

b) IndProdCap: Indicador que avalia toda produção intelectual do programa no formato de capítulos de livros, sendo participação obrigatória de um docente permanente com autor, através da seguinte equação:

$$IndProdOrient = \sum_{j=1}^{qtdDocentes} cap_j/DP \tag{4.5}$$

onde cap representa os capítulos de livros do programa;
DP é a quantidade de docentes permanentes no programa.

c) IndProdTec: Indicador que avalia toda produção intelectual do programa classificada como produção técnica e tecnológica, através da seguinte equação:

$$IndProdTec = \sum_{j=1}^{qtdDocentes} tec_j/DP \tag{4.6}$$

onde te representa todo produção técnica e tecnológica de programa;
DP é a quantidade de docentes permanentes no programa.

d) IndProd: Indicador que avalia toda produção intelectual do programa, através da soma dos indicadores IndProdArt, IndProdLiv, IndProdCap e IndProdTec, através da seguinte equação:

$$IndProd = indProd + indProdLiv + IndProdCap + IndProdTec \tag{4.7}$$

A Figura 4.14 demostra o dendrograma gerado após execução do ScriptComp para o indicador IndProd.

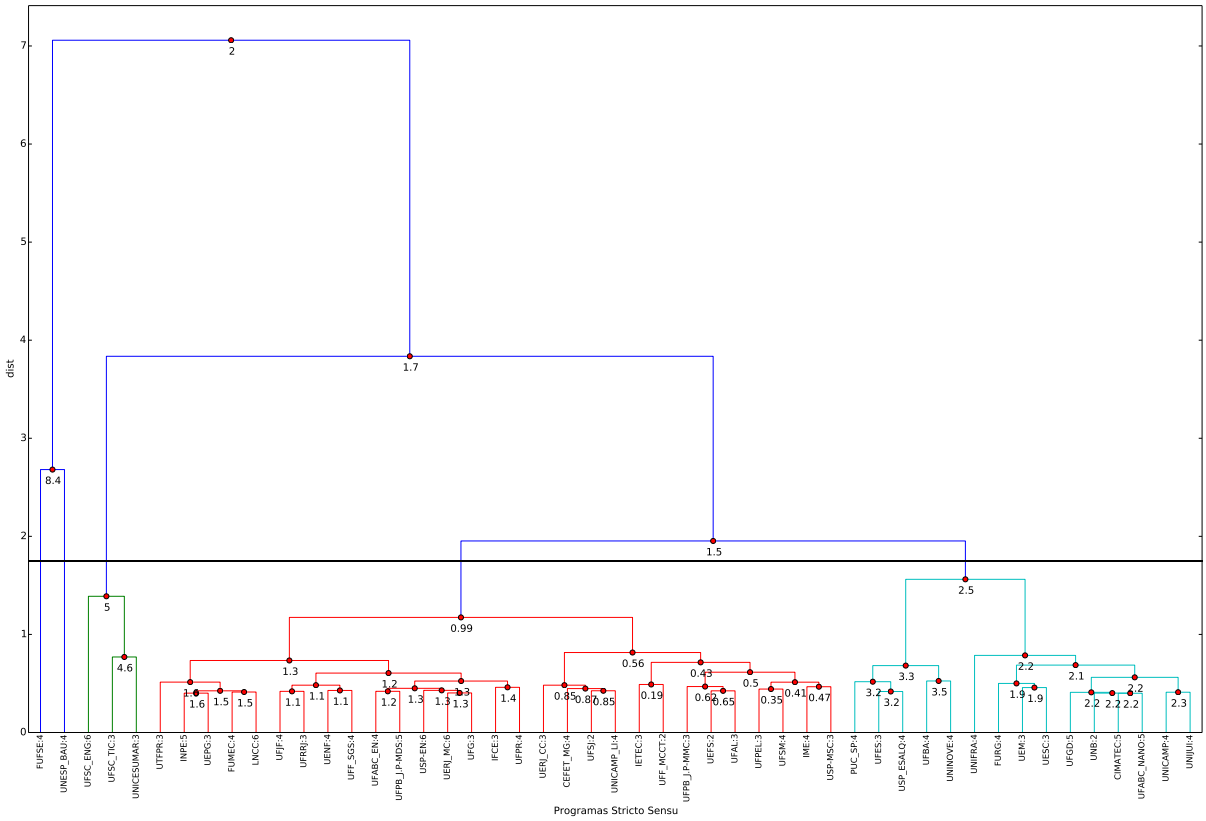


Figura 4.14: *Resultado da execução do ScriptComp com corte na hierarquia, para o indicador IndProd*

Fonte : Próprio autor

O *corte* definido na Figura 4.14, na região do eixo x na distância 1.8, separou os programas em 5 grupos. Ao analisar o eixo central de cada grupo, é possível verificar o desempenho de cada programa para este indicador, Desta forma, na Tabela 4.6 é apresentado a média do conceito CAPES, assim como a quantidade de programas em cada grupo.

Tabela 4.6: *Análise dos programas após o agrupamento hierárquico - IndProd*

indProdArtSUP por grupo	Programas por grupo	Média dos conceitos
5	3	4
4.2	1	4
4.2	1	4
2.5	15	3.8
0.99	107	3.6

Com isso podemos fazer as seguintes afirmações:

- Analisando a Figura 4.14 em conjunto com a Tabela 4.6, é possível afirmar que os programas das universidades FUFSE, UNESP_BAU, UFSC_ENG, USP_TIC

e UNICESUMAR, no ano de 2016, tiveram produtividade maior do que os demais programas dos grupos.

- A média do conceito CAPES para os programas das universidades USP_TIC e UNICESUMAR, estão acima de seu conceito atual.

- O grupo de valor 0.99, possui média CAPES de 3.8. Entretanto, este grupo dispõe de 2 programas com conceito 5 e 3 com conceito 6, com isso é possível afirmar que os programas com conceito acima de 4 que estão alocados no grupo 3.8, tiveram um desempenho similar a um programa com conceito 4 para indicador indProd.

Capítulo 5

Conclusão

5.1 Conclusão

Com o método produzido neste trabalho, foi possível criar grupos em hierarquia, dos dados extraídos da plataforma Lattes.

A técnica de aprendizado de máquina, denominada agrupamento hierárquico, proporcionou, de modo satisfatório a criação de agrupamentos em diversos níveis dos dados. Um dos benefícios deste técnica, é possibilitar, de maneira prática, a customização das funções de distância dos agrupamentos, permitindo assim, adequar nosso objetivo de categorizar os relatórios gerados pela ferramenta *ScriptLattes*.

A utilização do *Web-Crawler ScriptLattes*, foi fundamental para extração do dados pertencentes aos grupos da hierarquia. Explorar a limitação do ScriptLattes e utilizá-lo como ferramenta de extração foi fundamental para os resultados obtidos e conclusão desta pesquisa.

O *script MultScriptLattes* implementou o método proposto de hierarquia de currículos, com intuito de agrupar os dados e contribuir positivamente na transparência e gestão de programas *stricto-sensu* e de cursos de graduação. Como anteriormente citado, tanto a configuração para cadastro da plataforma Sucupira, quanto a da plataforma e-Mec, contribuíram com trabalhos recentemente publicados, demonstrando sua eficácia nos agrupamentos de dados.

Em especial a implementação do método, visando medir a similaridade de programas, através da ferramenta *ScriptComp*, apresentou uma nova possibilidade na comparação de dados com base nos indicadores da área, abrindo um grande leque para novos trabalhos.

Esta pesquisa não tem como intuito, esgotar ou limitar o campo explorado, no que se refere ao agrupamento dos dados existentes na plataforma Lattes, Pretende-se sim, contribuir com o avanço de pesquisas que utilizam técnicas de aprendizagem de máquina sobre os dados existentes nesta plataforma.

5.2 Trabalhos Futuros

Como oportunidades de trabalhos futuros, propomos que:

a) Seja aplicada a método proposto, a fim de criar novos agrupamentos sobre os dados da plataforma Lattes.

b) Analisar os relatórios categorizados de outros programas de graduação e pós-graduação *stricto sensu*.

c) Desenvolver outras visualizações dos agrupamentos hierárquicos, que não seja apresenta por dendrograma.

d) Utilizar este método, para medir a similaridade de outras áreas do conhecimento com base em seus indicadores.

e) Utilizar outras técnicas de aprendizagem de máquina, não supervisionada, para realizar os agrupamentos dos dados.

f) Predizer o conceito CAPES de um programa de pós-graduação *stricto sensu*, utilizando técnicas de aprendizagem de máquina.

Apêndice

5.3 Produções durante o Mestrado

a) Alves, Wonder A.L, Santos, Saulo D e Schimit, Pedro HT . Hierarchical clustering based on reports generated by scriptlattes. IFIP International Conference on Advances in Production Management Systems. Springer, Foz do Iguaçu-PR, 2016.

b) Santos, Saulo D e Alves, Wonder AL. Extrações de informações automática para prestação de contas na plataforma Sucupira. SETII Seminário em tecnologia da informação inteligente. UNINOVE, São Paulo-SP, 2016.

Referências Bibliográficas

ALVES, A. D.; YANASSE, H. H.; SOMA, N. Y. **LattesMiner: a multilingual DSL for information extraction from lattes platform**. In: ACM. *Proceedings of the compilation of the co-located workshops on DSM'11, TMC'11, AGERE! 2011, AOOPES'11, NEAT'11, & VMIL'11*. [S.l.], 2011. p. 85–92. Citado na pág. 22.

ALVES, W. A.; SANTOS, S. D.; SCHIMIT, P. H. **Hierarchical Clustering Based on Reports Generated by Scriptlattes**. In: SPRINGER. *IFIP International Conference on Advances in Production Management Systems*. [S.l.], 2016. p. 28–35. Citado na pág. 20.

ANDRETTA, P. I. S. et al. **Uma análise sobre a produção, produtividade e colaboração na Ciência da Informação no Brasil entre os anos 2007 a 2009**. *Palabra clave*, SciELO Argentina, v. 1, n. 2, p. 48–52, 2012. Citado na pág. 29.

BAVARESCO, J. ***Utilização e validação da ferramenta computacional SCRIPTEMEC como estratégia inovadora no gerenciamento da produtividade acadêmica de uma instituição privada de ensino superior***. Dissertação (Mestrado), 2017. Citado na pág. 52.

BECHHOFER, S. **OWL: Web ontology language**. In: *Encyclopedia of Database Systems*. [S.l.]: Springer, 2009. p. 2008–2009. Citado na pág. 21.

BONIFACIO, A. S. ***Ontologias e consulta semântica: Uma aplicaçao ao caso Lattes***. Tese (Doutorado) — UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL, 2002. Citado na pág. 16, 20.

BRAS, P. O. **Organização do currículo–plataforma Lattes Curriculum vitae organization–the Lattes software platform**. *Pesqui Odontol Bras*, SciELO Brasil, v. 17, n. Supl 1, p. 18–22, 2003. Citado na pág. 15.

BRUALDI, R. ***Introductory Combinatorics***. Pearson Education International, 2012. (Pearson Education). ISBN 9780132791717. Disponível em: <https://books.google.ca/books?id=7aeWuQAACAAJ>. Citado na pág. 31.

BURNS, C. S.; LANA, A.; BUDD, J. **Institutional repositories: exploration of costs and value**. *D-Lib Magazine*, Corporation for National Research Initiatives, v. 19, n. 1, p. 1, 2013. Citado na pág. 15.

CATANI, A. M.; OLIVEIRA, J. d. **A educação superior**. *A organização do ensino no Brasil: níveis e modalidades na Constituição federal e na LDB. São Paulo: Editora Xamã*, 2002. Citado na pág. 50.

COSTA, A. P. da; YAMATE, F. S. **Semantic Lattes: uma ferramenta de consulta de informaçoes acadêmicas da base Lattes baseada em ontologias**. *Trabalho de Conclusao de Curso, Escola Politécnica da Universidade de Sao Paulo, Sao Paulo, SP*, 2009. Citado na pág. 16, 21.

DUDA, R. O. et al. Pattern classification. 2nd. *Edition. New York*, p. 55, 2001. Citado na pág. 32.

EDGAR, B. D.; WILLINSKY, J. **A survey of scholarly journals using Open Journal Systems**. *Scholarly and Research Communication*, v. 1, n. 2, 2010. Citado na pág. 15.

ELISHAR, A. et al. **Organizational intrusion: Organization mining using socialbots**. In: IEEE. *Social Informatics (SocialInformatics), 2012 International Conference on*. [S.l.], 2012. p. 7–12. Citado na pág. 20.

FERNÁNDEZ-BREIS, J. T. et al. **A semantic platform for the management of the educative curriculum**. *Expert Systems with Applications*, Elsevier, v. 39, n. 5, p. 6011–6019, 2012. Citado na pág. 15, 16, 28.

FERRAZ, R. R. N.; QUONIAM, L. M. **A utilização da ferramenta computacional Scriptlattes para avaliação das competências em pesquisa no Brasil**. *Prisma. com*, n. 21, 2017. Citado na pág. 27.

FERRAZ, R. R. N.; QUONIAM, L. M.; MACCARI, E. A. **The use of Scriptlattes tool for extraction and on line availability of academic production from a department of Stricto Sensu in Management**. In: *11th International Conference on Information Systems and Technology Management–CONTECSI*. [S.l.: s.n.], 2014. v. 17. Citado na pág. 16, 50.

FRIEDMAN, J.; HASTIE, T.; TIBSHIRANI, R. ***The elements of statistical learning***. [S.l.]: Springer series in statistics Springer, Berlin, 2001. v. 1. Citado na pág. 32.

GIORDANO, D. M.; BRUNING, E.; BORDIN, A. S. **Uso do scriptlattes e gephi na análise da colaboração científica**. *Anais do Computer on the Beach*, p. 239–248, 2015. Citado na pág. 16, 28.

HORROCKS, I. et al. **DAML+OIL: A Description Logic for the Semantic Web**. *IEEE Data Eng. Bull.*, v. 25, n. 1, p. 4–9, 2002. Citado na pág. 20.

KADRIU, A. **Discovering value in academic social networks: A case study in ResearchGate**. In: IEEE. *Information Technology Interfaces (ITI), Proceedings of the ITI 2013 35th International Conference on*. [S.l.], 2013. p. 57–62. Citado na pág. 15.

KAUSAR, M. A.; DHAKA, V.; SINGH, S. K. **Web crawler: a review**. *International Journal of Computer Applications*, Foundation of Computer Science, v. 63, n. 2, 2013. Citado na pág. 19.

LANE, J. **Let's make science metrics more scientific**. *Nature*, Nature Publishing Group, v. 464, n. 7288, p. 488–489, 2010. Citado na pág. 15.

LYNCH, C. A. **Institutional repositories: essential infrastructure for scholarship in the digital age**. *portal: Libraries and the Academy*, The Johns Hopkins University Press, v. 3, n. 2, p. 327–336, 2003. Citado na pág. 15.

MACCARI, E. A. et al. **A gestão dos programas de pós-graduação em administração com base no sistema de avaliação da Capes**. *REGE. Revista de Gestão*, REGE, Revista de Gestão, v. 16, n. 4, p. 1, 2009. Citado na pág. 52.

MACHADO, C. C. et al. **Um Web Crawler para Projeções e Análise de Vulnerabilidades de Segurança e Consistência Estrutural de Páginas Web**. *Revista de Empreendedorismo, Inovação e Tecnologia*, v. 2, n. 2, p. 3–12, 2016. Citado na pág. 20.

MENA-CHALCO, J. P.; CESAR-JR, R. M. **scriptLattes: an open-source knowledge extraction system from the Lattes platform**. *Journal of the Brazilian Computer Society*, v. 15, n. 4, p. 31–39, 2009. ISSN 0104-6500. Disponível em: <http://link.springer.com/10.1007/BF03194511>. Citado na pág. 16, 20, 23.

MENA-CHALCO, J. P.; CESAR-JR, R. M. **Prospecção de dados acadêmicos de currículos Lattes através de Scriptlattes**. *Bibliometria e Cientometria: reflexões teóricas e interfaces. São Carlos: Pedro & João*, 2013. Citado na pág. 16, 28.

MENA-CHALCO, J. P.; DIGIAMPIETRI, L. A.; CESAR-JR, R. M. **Caracterizando as redes de coautoria de currículos Lattes**. In: *Brazilian Workshop on Social Network Analysis and Mining (BraSNAM)*. [S.l.: s.n.], 2012. p. 1–12. Citado na pág. 25.

MIRTAHERI, S. M. et al. **A brief history of web crawlers**. In: IBM CORP. *Proceedings of the 2013 Conference of the Center for Advanced Studies on Collaborative Research*. [S.l.], 2013. p. 40–54. Citado na pág. 19.

NEWMAN, M. H. A. ***Elements of the topology of plane sets of points***. [S.l.]: Dover Publications, 1992. Citado na pág. 31.

NIGRO, C. A. et al. **Uso da ferramenta computacional scriptlattes-scriptsucupira na prestação de contas anual e quadrienal a CAPES por um programa de pós-graduação stricto sensu em direito**. 2015. Citado na pág. 16, 50.

NIGRO, C. A. et al. **Prestação de contas anual e quadrienal à Capes por um programa de Pós-Graduação stricto sensu em Engenharia de Produção: utilização da ferramenta computacional Scriptlattes-Scriptsucupira**. *PRISMA. COM*, n. 29, 2017. Citado na pág. 50.

NIGRO, C. A. et al. **Uso das ferramentas computacionais Scriptlattes, ScriptGP e Patent2net para análise da produção bibliográfica e tecnológica sobre a dengue**. Universidade Nove de Julho, 2016. Citado na pág. 16, 29.

O.A, M. **Genvl and wwww:Toolsfortam- ing the web**. *In Proceedings of the First Interna- tional World Wide Web Conference*, p. pp. 79– 90, 1994. Citado na pág. 19.

OLIVEIRA, E.; BERMEJO, P. d. S.; KERN, V. M. **GeraLattes: extração de informação gerencial de currículos de pesquisadores usando XML**. In: *Workshop de Computação da Região Sul (WorkCompSul 2004)*. [S.l.: s.n.], 2004. v. 1. Citado na pág. 21.

OLIVEIRA, W. A. de; SILVA, F. F. da; HAYASHI, C. R. M. **Redes de coautoria em educação: uma análise a partir de artigos científicos produzidos nos programas de pós-graduação**. In: *XVIII Seminário Nacional de Bibliotecas Universitárias*. [S.l.: s.n.], 2014. Citado na pág. 29.

PAGE, L. et al. The pagerank citation ranking: Bringing order to the web. 1999. Citado na pág. 20.

ROCHA, D.; KREUTZ, D.; TURCHETTI, R. **Uma Ferramenta Livre e Extensível Para Detecção de Vulnerabilidades em Sistemas Web**. *Computer Science and Engineering July*, v. 14, 2012. Citado na pág. 20.

RODRIGUES, T. G. et al. **Uma Ferramenta de Suporte a Recuperação de Informação na Web focada em Vulnerabilidades e Anomalias Internet**. *X Simpósio Brasileiro em Segurança da Informação e de Sistemas Computacionais (SBSeg)*, p. 227–240, 2010. Citado na pág. 20.

SANTOS, F. L. d. **Mineração de opinião em textos opinativos utilizando algoritmos de classificação**. 2014. Citado na pág. 20.

SANTOS, L. M. **Protótipo para mineração de opinião em redes sociais: estudo de casos selecionados usando o twitter**. 2010. Citado na pág. 20.

SGUISSARDI, V. **A avaliação defensiva no "modelo CAPES de avaliação": É possível conciliar avaliação educativa com processos de regulação e controle do Estado?** *Perspectiva*, v. 24, n. 1, p. 49–88, 2006. Citado na pág. 56.

SILVA, F. M. ***Organização da Informação em sistemas eletrônicos abertos de Informação Científica & Tecnológica: Análise da Plataforma Lattes.*** Tese (Doutorado) — Tese (Doutorado em Ciência da Informação)-Departamento de Biblioteconomia e Documentação, Universidade de São Paulo, São Paulo, 2007. Citado na pág. 16.

SILVA, M. V. C. ***Avaliação contínua e automatizada da produtividade academîca dos cursos de graduação de uma instituição privada de ensino superior***. Dissertação (Mestrado), 2017. Citado na pág. 52.

STELA, G. ***Lattes extrator***. [S.l.]: Florianópolis: Universidade Federal de Santa Catarina, 2002. Citado na pág. 23.

Anexos

Tabela 5.1: *Programas área interdisciplinar com conceito CAPES* 2

Programa	Instituição	M	D
Computação Aplicada	UNIVERSIDADE ESTADUAL DE FEIRA DE SANTANA	2	
MODELAGEM COMPUTACIONAL EM CIÊNCIA E TECNOLOGIA	UNIVERSIDADE FEDERAL FLUMINENSE	2	
TECNOLOGIAS PARA O DESENVOLVIMENTO SUSTENTÁVEL	Universidade Federal de São João del-Rei	2	
TECNOLOGIAS QUÍMICA E BIOLÓGICA	UNIVERSIDADE DE BRASÍLIA	2	

Tabela 5.2: *Programas área interdisciplinar com conceito CAPES* 3

Programa	Instituição	M	D
Engenharia e Gestão de Processos e Sistemas	Instituto de Educação Tecnológica	3	
Energias Renováveis	INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DO CEARÁ	3	
CIÊNCIAS FÍSICAS APLICADAS	UNIVERSIDADE ESTADUAL DO CEARÁ	3	
Bioenergia - UEL - UEM - UEPG - UNICENTRO - UNIOESTE - UFPR	UNIVERSIDADE ESTADUAL DE MARINGÁ	3	
Computação Aplicada	UNIVERSIDADE ESTADUAL DE PONTA GROSSA	3	
CIÊNCIAS COMPUTACIONAIS	UNIVERSIDADE DO ESTADO DO RIO DE JANEIRO	3	
Modelagem Computacional em Ciência e Tecnologia	UNIVERSIDADE ESTADUAL DE SANTA CRUZ	3	
ENGENHARIA E GESTÃO DA INOVAÇÃO	UNIVERSIDADE FEDERAL DO ABC	3	
MODELAGEM COMPUTACIONAL DE CONHECIMENTO	UNIVERSIDADE FEDERAL DE ALAGOAS	3	
MODELAGEM E MÉTODOS QUANTITATIVOS	UNIVERSIDADE FEDERAL DO CEARÁ	3	
ENERGIA	UNIVERSIDADE FEDERAL DO ESPÍRITO SANTO	3	
Modelagem e Otimização	UNIVERSIDADE FEDERAL DE GOIÁS	3	
MODELAGEM MATEMÁTICA E COMPUTACIONAL	UNIVERSIDADE FEDERAL DA PARAÍBA/JOÃO PESSOA	3	
Modelagem Matemática	UNIVERSIDADE FEDERAL DE PELOTAS	3	
Modelagem Matemática e Computacional	UNIVERSIDADE FEDERAL RURAL DO RIO DE JANEIRO	3	
Tecnologias da Informação e Comunicação	UNIVERSIDADE FEDERAL DE SANTA CATARINA	3	
Gestão do Conhecimento nas Organizações	Centro Universitário de Maringá	3	
Modelagem de Sistemas Complexos	UNIVERSIDADE DE SÃO PAULO	3	
TECNOLOGIAS COMPUTACIONAIS PARA O AGRONEGÓCIO	UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ	3	

Tabela 5.3: *Programas área interdisciplinar com conceito CAPES* 4

Programa	Instituição	M	D
MODELAGEM MATEMÁTICA E COMPUTACIONAL	CENTRO FEDERAL DE EDUCAÇÃO TECN. DE MINAS GERAIS	4	
Ciência da Propriedade Intelectual	FUNDAÇÃO UNIVERSIDADE FEDERAL DE SERGIPE	4	
SISTEMAS DE INFORMAÇÃO E GESTÃO DO CONHECIMENTO	UNIVERSIDADE FUMEC	4	
MODELAGEM COMPUTACIONAL	UNIVERSIDADE FEDERAL DO RIO GRANDE	4	
ENGENHARIA DE DEFESA	INSTITUTO MILITAR DE ENGENHARIA	4	
PROPRIEDADE INTELECTUAL E INOVAÇÃO	INSTITUTO NACIONAL DA PROPRIEDADE INDUSTRIAL	4	
TECNOLOGIAS DA INTELIGÊNCIA E DESIGN DIGITAL	PONTIFÍCIA UNIVERSIDADE CATÓLICA DE SÃO PAULO	4	
CIÊNCIAS NATURAIS	UNIVERSIDADE ESTADUAL DO NORTE FLUMINENSE DARCY RIBEIRO	4	
ENERGIA	UNIVERSIDADE FEDERAL DO ABC	4	
ENERGIA E AMBIENTE	UNIVERSIDADE FEDERAL DA BAHIA	4	
Sistemas de Gestão Sustentáveis	UNIVERSIDADE FEDERAL FLUMINENSE	4	
MODELAGEM COMPUTACIONAL	UNIVERSIDADE FEDERAL DE JUIZ DE FORA	4	
CIÊNCIA, GESTÃO E TECNOLOGIA DA INFORMAÇÃO	UNIVERSIDADE FEDERAL DO PARANÁ	4	
EDUCAÇÃO PROFISSIONAL E TECNOLÓGICA	UNIVERSIDADE FEDERAL DE SANTA MARIA	4	
Nanociência e Nanobiotecnologia	UNIVERSIDADE DE BRASÍLIA	4	
MÍDIA E TECNOLOGIA	UNIVERSIDADE EST.PAULISTA JÚLIO DE MESQUITA FILHO/BAURU	4	
PLANEJAMENTO DE SISTEMAS ENERGÉTICOS	UNIVERSIDADE ESTADUAL DE CAMPINAS	4	
Tecnologia	UNIVERSIDADE ESTADUAL DE CAMPINAS/LIMEIRA	4	
NANOCIÊNCIAS	CENTRO UNIVERSITÁRIO FRANCISCANO	4	
MODELAGEM MATEMÁTICA	UNIV. REGIONAL DO NOROESTE DO ESTADO DO RIO GRANDE DO SUL	4	
INFORMÁTICA E GESTÃO DO CONHECIMENTO	UNIVERSIDADE NOVE DE JULHO	4	
Bioenergia USP, UNICAMP E UNESP	UNIV.DE SÃO PAULO/ESCOLA SUP. DE AGRICULTURA LUIZ DE QUEIROZ	4	

Tabela 5.4: *Programa área interdisciplinar com conceito CAPES* 5

Programa	Instituição	M	D
MODELAGEM COMPUTACIONAL E TECNOLOGIA INDUSTRIAL	FACULDADE DE TECNOLOGIA SENAI CIMATEC	5	
COMPUTAÇÃO APLICADA	INSTITUTO NACIONAL DE PESQUISAS ESPACIAIS	5	
NANOCIÊNCIAS E MATERIAIS AVANÇADOS	UNIVERSIDADE FEDERAL DO ABC	5	
Ciência e Tecnologia Ambiental	UNIVERSIDADE FEDERAL DA GRANDE DOURADOS	5	
MODELOS DE DECISÃO E SAÚDE	UNIVERSIDADE FEDERAL DA PARAÍBA/JOÃO PESSOA	5	

Tabela 5.5: *Programa área interdisciplinar com conceito CAPES* 6

Programa	Instituição	M	D
ODELAGEM COMPUTACIONAL	LABORATÓRIO NACIONAL DE COMPUTAÇÃO CIÊNTÍFICA	6	
MODELAGEM COMPUTACIONAL	UNIVERSIDADE DO ESTADO DO RIO DE JANEIRO	6	
ENGENHARIA E GESTÃO DO CONHECIMENTO	UNIVERSIDADE FEDERAL DE SANTA CATARINA	6	
ENERGIA	UNIVERSIDADE DE SÃO PAULO	6	