Universidade Federal de Juiz de Fora
Instituto de Ciências Exatas
Bacharelado em Ciência da Computação

Uma ferramenta de apoio à análise quantitativa de cursos de pós-graduação

Welson de Avelar Soares Filho

JUIZ DE FORA
MARÇO, 2013

Uma ferramenta de apoio à análise quantitativa de cursos de pós-graduação

Welson de Avelar Soares Filho

Universidade Federal de Juiz de Fora
Instituto de Ciências Exatas
Departamento de Ciência da Computação
Bacharelado em Ciência da Computação

Orientador: Marcelo Lobosco

JUIZ DE FORA
MARÇO, 2013

Uma ferramenta de apoio à análise quantitativa de cursos de pós-graduação

Welson de Avelar Soares Filho

MONOGRAFIA SUBMETIDA AO CORPO DOCENTE DO INSTITUTO DE CIÊNCIAS EXATAS DA UNIVERSIDADE FEDERAL DE JUIZ DE FORA, COMO PARTE INTEGRANTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE BACHAREL EM CIÊNCIA DA COMPUTAÇÃO.

Aprovada por:

Marcelo Lobosco
D.Sc., Engenharia de Sistemas e Computação

Bernardo Martins Rocha
M.Sc., Modelagem Computacional

Luis Paulo da Silva Barra
D.Sc., Engenharia Civil

Rodrigo Weber dos Santos
D.Sc., Matemática

JUIZ DE FORA

27 DE MARÇO, 2013

Aos meus pais, pelo apoio e sustento.

Aos meus irmãos, familiares e amigos.

Obrigado, para sempre, pelo carinho.

Resumo

Este trabalho apresenta uma nova ferramenta de apoio gerencial para coordenadores de cursos de pós-graduação. Foi utilizada, como base para o seu desenvolvimento, o ScriptLattes, um sistema de extração de dados web a partir da plataforma Lattes. A principal modificação realizada no ScriptLattes foi em seu *parser* HTML-Python, de modo que este passasse a cruzar dados da plataforma Lattes com consultas no banco de dados Webqualis. Desta forma, pode-se fazer a análise de alguns dos parâmetros quantitativos adotados pela CAPES para avaliar os programas de pós-graduação *stricto sensu*. Em particular o presente trabalho apresenta exemplos e estudos de caso focados nos critérios adotados pela área Interdisciplinar, apesar dos mecanismos apresentados neste trabalho serem genéricos, podendo portanto ser modificados de forma a adequá-los as particularidades de outras áreas.

Palavras-chave: Sistema de Apoio Gerencial, ScriptLattes, CAPES, Webqualis.

Abstract

This work presents a new management support tool that can be used by coordinators of Brazilian graduate courses. The new tool was based on ScriptLattes, a system used to extract data from the Lattes web platform. The main modification performed in ScriptLattes was in its HTML-Python parser, in order to allow data extracted from Lattes platform to be joined with queries performed on the Webqualis database. Thus, one can make the analysis of some of the quantitative parameters adopted by CAPES to evaluate Brazilian graduate courses. In particular this work presents examples and case studies focused on the criteria adopted by the CAPES Interdisciplinary area, although the mechanisms presented in this work are generic and therefore can be modified to fit the particularities found on distinct areas.

Keywords: Management Support Tool, ScriptLattes, CAPES, Webqualis.

Agradecimentos

Primeiramente aos meus pais, Welson e Regina, por terem me dado toda base necessária para chegar até aqui e ir muito além. Aos meus irmãos, Raphael e Patrícia, por todo incentivo. E à todos meus parentes por serem uma família maravilhosa.

Ao querido professor Marcelo "Dom" Lobosco pelo exemplo de humildade, amizade e principalmente, pela paciência, sem a qual este trabalho não se realizaria.

Aos professores do Departamento de Ciência da Computação/UFJF pelos seus ensinamentos e aos funcionários do curso, que durante esses anos, contribuíram de algum modo para o meu enriquecimento pessoal e profissional.

Por fim, mas não menos importante, aos meus amigos e amigas de faculdade e de fora dela. Vocês são companheiros para uma vida inteira.

"Olha, lá vejo meu pai.
Olha, lá vejo minha mãe, meus irmãos e irmãs.
Olha, lá vejo meus ancestrais desde o início.
Olha, lá eles clamam por mim,
Pedem que assuma meu lugar entre eles nos salões do Valhalla,
Onde os bravos vivem para sempre.".

Oração viking

Sumário

Lista de Figuras

Lista de Abreviações

CAInter	Coordenação da Área Interdisciplinar da CAPES
CAPES	Coordenação de Aperfeiçoamento de Pessoal de Nível Superior
CNPq	Conselho Nacional de Desenvolvimento Científico e Tecnológico
DOI	Digital Object Identifier
FAPEMIG	Fundação de Amparo à Pesquisa do Estado de Minas Gerais
GLP	General Public License
GNU	GNU's Not Unix!
ISSN	International Standard Serial Number
IndFor	Indicador de Formação Docente
IndOri	Indicador de Orientações
IndProd	Indicador de Produção Intelectual
PGMC	Pós-Graduação em Modelagem Computacional
UFJF	Universidade Federal de Juiz de Fora

1 Introdução

1.1 Motivação

O número de programas de pós-graduação *stricto sensu* cresceu nos últimos anos no Brasil. Segundo dados da fundação do Ministério da Educação responsável pela expansão e consolidação da pós-graduação *stricto sensu*, a CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior), o triênio 2007-2009 teve um aumento de 20,8% na quantidade de cursos de mestrado, doutorado e mestrado profissional [5]. E a tendência é que este número aumente no triênio 2010-2012, ainda sem dados divulgados.

Um curso, para se manter funcionando, precisa obter no mínimo avaliação igual a 3 pela CAPES. Esta avaliação segue uma escala: notas 1 e 2 reprovam o programa; nota 3 corresponde a um desempenho regular, atendendo ao padrão mínimo de qualidade; nota 4 equivale a um bom desempenho; nota 5 é considerado um nível muito bom; notas 6 e 7, um desempenho equivalente ao alto padrão internacional. E são considerados, na avaliação, 5 eixos: a proposta do programa, o corpo docente, o corpo discente (teses e dissertações produzidas pelos alunos), a produção intelectual e a inserção social (impacto e integração com outros centros de pesquisa) [6]. Numa comparação entre o triênio 2007-2009 com 2004-2006, 10% dos cursos avaliados diminuíram o conceito e 85 dos 4.099 cursos avaliados não alcançaram o conceito mínimo 3 [6].

O acompanhamento constante, por parte do coordenador, dos 5 eixos que compõem a avaliação do curso é primordial para manter ou aumentar a nota do curso. Contudo, tal tarefa revela-se árdua. Isto ocorre em parte pelo fato da informação a ser consultada estar distribuída em diferentes bases de dados. Por exemplo, para acompanhar a qualidade das publicações em periódicos de seu programa, o coordenador deveria primeiro consultar a base de dados que lista as publicações, a plataforma Lattes, extraindo, para cada docente, informações contidas na seção "Produções". Em seguida, para avaliar a qualidade de cada produção, outra base de dados deveria ser consultada, o banco de dados Webqualis. Do cruzamento e consolidação de tais informações o coordenador calcula o índice de pro-

dutividade do programa de pós-graduação. E este é apenas um dos parâmetros a serem avaliados. Tal acompanhamento se soma às atividades administrativas de gestão do curso, sobrecarregando ainda mais o coordenador.

1.2 Objetivo

O intuito deste trabalho é exatamente o de propor e implementar uma ferramenta para facilitar a tarefa de consulta e cruzamento das informações nos programas de pós-graduação *stricto sensu*, de modo a automatizar parte do processo de avaliação dos cursos. Como destacado anteriormente, o resultado esperado é facilitar o acompanhamento contínuo de alguns dos indicadores do programa por parte de seu coordenador, tornando-se assim uma valiosa ferramenta administrativa.

1.3 Organização

Este trabalho está organizado da seguinte forma. No capítulo 2 será apresentado o ScriptLattes, programa que serviu de ponto de partida para este trabalho. Originalmente, o ScriptLattes extrai dados da plataforma Lattes segundo informações contidas em um arquivo de configuração. Dados como publicações, participações em banca examinadora, orientações, entre outros, são consultados para cada pesquisador, sendo ao final do processo gerada uma página HTML com estas informações concentradas, para facilitar a consulta. O ScriptLattes também gera gráficos quantitativos que apresentam os números de publicações do pesquisador ao longo do tempo, sendo esta uma opção que também pode ser configurada por meio de um arquivo de entrada.

O capítulo 3 apresenta a ferramenta proposta, detalhando a sua implementação. São tratadas neste capítulo as alterações no *parser* HTML-Python necessárias para consultar a base de dados do Webqualis, bem como os cálculos dos índices de avaliação. Este capítulo mostra também as principais dificuldades de implementação da ferramenta.

O quarto capítulo apresenta um caso de uso da ferramenta, aplicado ao curso de Pós-Graduação em Modelagem Computacional (PGMC) da Universidade Federal de Juiz de Fora (UFJF).

Por fim, no último capítulo serão apresentadas as conclusões e possibilidades de melhorias futuras na ferramenta, como a sua integração com outro sistema utilizado pela CAPES para coletar indicadores dos cursos de pós-graduação *stricto sensu*, chamado "COLETA CAPES", que também são usados para avaliação.

2 ScriptLattes

O Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) faz um importante trabalho para manter unificado os dados dos currículos acadêmicos de professores, alunos e pesquisadores, de instituições públicas e particulares, em uma plataforma denominada Lattes. Os chamados "Currículos Lattes" (ou CV Lattes) são considerados um padrão nacional de avaliação representando um histórico das atividades científicas, acadêmicas e profissionais de pesquisadores cadastrados [4].

Diversas instituições de ensino e pesquisa usam dados do Lattes para formular relatórios de pesquisa, de produção científica e de orientações. De forma convencional, esta reunião de informações seria feita manualmente, levando muito tempo em caso de grupos grandes de pesquisadores e acadêmicos. Além do mais, este processo manual estaria sujeito à falha, mesmo que a Plataforma Lattes traga as informações de forma bem estruturada e bem dividida.

Em início de 2005, Jesús P. Mena-Chalco e Roberto M. Cesar Jr. disponibilizaram publicamente a primeira versão do ScriptLattes. Inicialmente o programa foi desenvolvido com a finalidade de auxiliar a Secretaria de Pós-graduação do IME-USP na elaboração de relatórios de produção bibliográfica de professores alocados no Departamento de Ciência da Computação do IME. Estes relatórios seriam feitos, exclusivamente, com dados cadastrados e disponibilizados publicamente na Plataforma Lattes do CNPq [9].

É importante destacar que o ScriptLattes é um software livre desenvolvido em Python, de código aberto e sob a licença GNU-GPL. Em termos gerais, a GPL baseia-se em quatro liberdades [10]:

- A liberdade de executar o programa, para qualquer propósito;
- A liberdade de estudar como o programa funciona e adaptá-lo para as suas necessidades. O acesso ao código-fonte é um pré-requisito para esta liberdade;
- A liberdade de redistribuir cópias de modo que você possa ajudar ao seu próximo.

- A liberdade de aperfeiçoar o programa, e liberar os seus aperfeiçoamentos, de modo que toda a comunidade se beneficie deles. O acesso ao código-fonte é um pré-requisito para esta liberdade.

E por ser distribuído desta forma é que foi possível o desenvolvimento deste trabalho.

2.1 Módulos do ScriptLattes

Para iniciar a prospecção dos dados, o sistema lê um arquivo de entrada de dados, em formato texto puro, que contêm o(s) identificador(es) (ID) do(s) CV Lattes, nome do(s) pesquisador(es) e o período de tempo que se deseja realizar a busca (Figura 2.1). O ID é um número de 16 algarismos atribuído a todo pesquisador cadastrado na plataforma Lattes.

Figura 2.1: Dados de entrada do ScriptLattes. Extraído de [4].

As saídas de dados do sistema são relatórios sumarizados, configuráveis, que podem conter a produção bibliográfica, produção técnica, orientações, etc. Tais relatórios são gerados como uma página HTML com *links* para outras páginas contendo as informações mais detalhadas.

2.1.1 Seleção dos Dados

Este módulo do ScriptLattes consulta as informações, em formato HTML, do *site* da Plataforma Lattes. Este módulo trabalha, essencialmente, apenas com *strings* e com-

parações entre *strings* para preencher os atributos dos objetos internos da aplicação. Os objetos internos são uma estrura de dados interna do programa que representa seções do currículo Lattes do pesquisador. Para que estes atributos funcionem como esperado, é feita a normalização da codificação de caracteres, ou seja, transforma os caracteres do formato ASCII para UTF-8. Isso mantém um mesmo padrão durante toda aplicação e evita perda de formatação ou informação [4].

2.1.2 Pré-processamento dos Dados

Neste módulo, o *parser* HTML extrai informações sobre produção técnica, publicação em periódicos, entre outros, de acordo com o período temporal estabelecido pelo arquivo de entrada. Nenhuma informação é criada pelo ScriptLattes; todos os dados são derivados dos campos extraídos do CV Lattes. A Figura 2.2 apresenta as informações que são extraídas.

2.1.3 Tratamento de Redundâncias

É comum que haja colaboração nas produções dentro de um mesmo grupo. Desta forma, uma mesma produção pode aparecer duplicada porque é declarada nas páginas de distintos autor(es)/co-autor(es). O título da produção é usado como ponto de início da busca por duplicatas. Este parâmetro é utilizado porque é padrão usar o título da produção nos currículos. Mas também são usados o nome do autor ou, caso exista, o código DOI, um padrão para a identificação de documentos na internet. Entretanto, vale ressaltar que por não haver um padrão na Plataforma Lattes quanto a estes valores, ou seja, padrão para os títulos (o pesquisador pode abreviar da forma que achar melhor) e nem obrigatoriedade do uso do código DOI (algumas publicações podem nem ter) tal processamento é complexo e passível de erros [4].

Todas as produções redundantes são armazenadas em uma matriz bidimensional, *matrizDeColaboracoes*, a qual armazena o número de colaborações bibliográficas, técnicas e artísticas a respeito dos membros do grupo de pesquisa que se deseja fazer os relatórios, sendo posteriormente utilizada para determinar a teia de colaboradores dos pesquisadores.

Table 1. Information extracted from the Lattes curriculum.

Personal information
Name
Professional address
Bibliographical production
Articles in scientific journals
Book published/organized
Book chapter published
Articles in newspapers/magazines
Complete works published in proceedings of conferences
Expanded summary published in proceedings of conferences
Summary published in proceedings of conferences
Articles accepted for publication
Presentations of work
Other kinds of bibliographical production
Technical production
Patented or registered software
Not patented or registered software
Technological products
Techniques or process
Technical works
Other kinds of technical production
Artistic productions
Artistic/cultural production
Ongoing / concluded supervisions
Postdoctorate supervision
Ph.D. thesis
Master's thesis
Monograph of completion for improvement/specialization
Works of completion for graduation
Scientific initiation
Other academic advisory

Figura 2.2: Informações extraídas da base de dados Lattes. Extraído de [4].

2.1.4 Geração do Grafo de Colaborações

Geralmente, um grafo de colaborações científicas descreve atividades de pesquisa que têm sido produzidas por um grupo. Neste módulo, o ScriptLattes cria uma saída gráfica para representar o grafo de colaboração entre os membros de um grupo com base em suas produções científicas. Neste grafo, cada membro é representado por um nó. Uma aresta é criada entre um par de nós sempre que uma produção comum dos pesquisadores correspondente é detectada pelo módulo de tratamento de redundância. Em outras palavras, se dois pesquisadores são co-autores de uma produção em comum, seus respectivos nós no grafo de colaborações são ligados por uma aresta.

O processo neste módulo é simplificado por utilizar a *matrizDeColaboracoes*, uma matriz simétrica que contém a quantidade total de produções em co-autoria entre os

membros, a qual foi gerada e calculada pelo módulo de tratamento de redundâncias. A *matrizDeColaboracoes* é tomada como uma matriz de adjacência (*matrizDeAdjacencia* dentro do código do *script*) que representa o grafo de colaborações. No grafo gerado, é possível observar a colaboração entre os membros e grupos de cooperação científica. Este grafo é um instrumento que ajuda a descobrir os pesquisadores com maior atividade de co-autoria dentro do grupo e pode ser usado na análise detalhada de co-autoria [4].

2.1.5 Mapa de Geolocalização

Muitas vezes deseja-se saber a localização geográfica dos membros de um grupo em particular, bem como de seus ex-alunos. Neste contexto, o ScriptLattes gera um "Mapa de Pesquisa", que representa a localização geográfica dos membros do grupo no mundo. Este módulo permite a criação de um Mapa de Pesquisa utilizando uma plataforma externa de endereços geográficos.

O nome do país, o nome da cidade, e o número de código postal de cada membro, disponível no Currículo Lattes, é usado para consultar o local. Além disso, a localização geográfica de cada doutor formado pelo grupo está também representada no mapa. Assim, o mapa mostra a pesquisa, onde os alunos formados estão trabalhando, passando uma idéia da influência mantida pelo grupo de pesquisa. É importante notar que, para realizar a recuperação dos valores de posicionamento geográfico de cada membro é usada uma interface com o Google Maps [4].

2.1.6 Geração de relatórios

Neste módulo são criados os relatórios das produções, bem como das supervisões em andamento e concluídas. Estes relatórios são separados por tipo e mostram informação quantitativa classificada por ano.

Os gráficos de barras estão associados com os relatórios, onde os comprimentos das barras são proporcionais à quantidade de produções científicas do pesquisador/grupo. Com esta informação é possivel ver o volume de publicações de um pesquisador/grupo e a partir disso, perceber a relevância de um pesquisador em questão ou do andamento de produção do grupo de pesquisa [8].

3 Uma Ferramenta de Apoio à Análise Qualitativa de Cursos de Pós-Graduação

3.1 Introdução

Este capítulo apresenta a proposta de uma nova ferramenta de apoio à analise qualitativa de cursos de pós-graduação. A ferramenta proposta baseia-se no sistema ScriptLattes, descrito no capítulo anterior. O sistema ScriptLattes foi modificado de modo a introduzir consultas no banco de dados de periódicos Webqualis, permitindo assim que sejam calculados automaticamente alguns dos índices usados na avaliação do desempenho de programas de pós-graduação.

Dentre os módulos do ScriptLattes apresentados no capítulo anterior, os que foram usados como ponto de partida para este trabalho foram os módulos "Entrada de Dados" e "Pré-processamento dos dados". Também foi necessária a modificação do módulo de geração de relatórios, visando à escrita das novas informações obtidas pelo cruzamento de informações com a base de dados Webqualis nas páginas HTML geradas.

De maneira geral, os demais módulos também sofreram pequenas modificações, de modo a permitir o encaminhamento das novas informações criadas entre os módulos e a gravação dos dados nos atributos dos objetos da aplicação.

O restante do capítulo apresenta as principais modificações realizadas no ScriptLattes, de modo a viabilizar o desenvolvimento da nova ferramenta.

3.2 Base Teórica

O principal alvo das modificações no ScriptLattes de modo a viabilizar o desenvolvimento da nova ferramenta foi o *parser* HTML-Python. Mas antes de descrever as modificações propriamente ditas, vamos primeiro explicar o que é e para que serve um *parser*.

Um *parser* [1, 2] é um programa, baseado em análise textual, que permite iden-

tificar e extrair regiões ou trechos específicos de uma entrada de dados. O *parser* é um importante elemento constituinte de compiladores. Nestes, cabe ao *parser* fazer a transformação do código fonte de entrada, que está numa linguagem de programação específica, em uma árvore de *tokens* (palavras) com igual poder de expressão da entrada dos dados.

A Figura 3.1 mostra esquematicamente como um *parser* funcionaria para esta linguagem. O *parser* para a linguagem separa a expressão em palavras (*tokens*), cada qual com seu valor, e monta a ordem de prioridade na forma de uma árvore. Perceba que o poder de expressão da linguagem de entrada se mantém, ou seja, o *parser* transforma uma linguagem em outra, sem alterar sua expressividade.

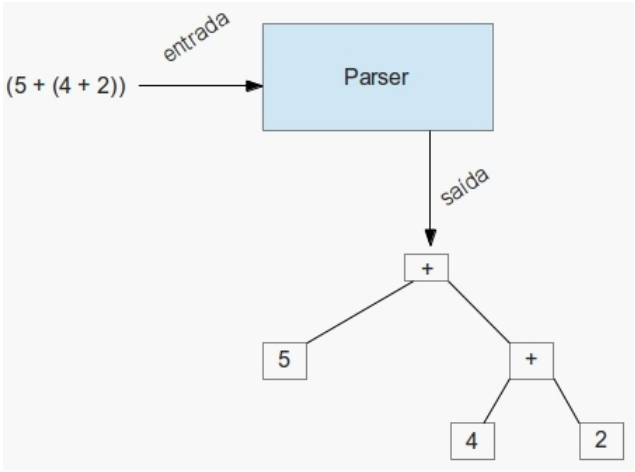

Figura 3.1: Representação esquemática do funcionamento do *parser*. Para cada entrada, é montada uma árvore baseada na gramática da linguagem.

Os métodos mais usados para a implementação de um *parser* são o *bottom-up* e o *top-down*[1]:

- O método *bottom-up* constrói uma árvore de sintaxe abstrata de baixo para cima, ou seja, das folhas para a raiz;
- O método *top-down* constrói a árvore de sintaxe abstrata de cima para baixo, partindo do nó raiz para as folhas.

3.3 O *Parser* HTML-Python Original

No ScriptLattes foi desenvolvido um *parser* HTML-Python deteminístico *top-down*, que extrai os seguintes dados do CV Lattes: nome completo, nome completo do membro, nome

em citações bibliográficas, endereço profissional, tipo de bolsa de produtividade, foto, sexo e data de atualização do currículo. Adicionalmente, são extraídas as listas completas de produções acadêmicas pertencentes ao período a ser pesquisado. É importante destacar que um desafio computacional para o programa é o tratamento dos dados em formato HTML, onde as partes constituintes das produções acadêmicas (por exemplo, nomes dos autores, título da publicação, título do projeto, nome do meio da publicação, número de páginas, volume, páginas, ano) são apresentadas sem alguma indicação de separação. Assim, o *parser* desenvolvido, bem como o modificado para este trabalho, identifica, na grande maioria dos casos, todas as partes constituintes das produções acadêmicas [8].

O *parser* originalmente desenvolvido no ScriptLattes recebe como entrada uma página HTML (linguagem A), com todas suas *tags* e valores, e alimenta os objetos Python (linguagem B). A Figura 3.2 ilustra este processo. Caso isso não fosse feito, o que se teria em mãos o tempo todo seria um grande volume de *strings*, mais difíceis de se manipular do que os dados estruturados.

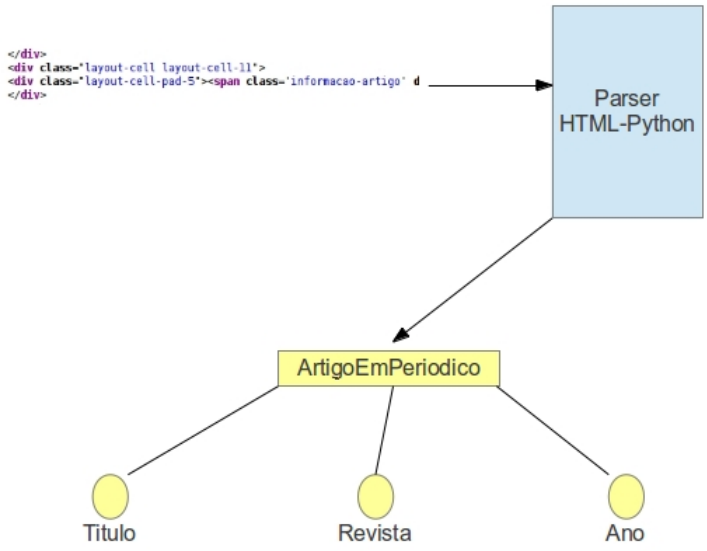

Figura 3.2: Funcionamento do *parser* original do ScriptLattes.

3.4 O *Parser* HTML-Python Modificado

O novo parser inclui no processamento uma consulta ao banco de dados de periódicos do Webqualis, conforme ilustrado na figura 3.3. Este banco de dados é extraído em formato de texto puro, com três campos: ISSN do periódico, título do periódico e o seu estrato, todos a partir do endereço *http* : *//qualis.capes.gov.br/webqualis/*.

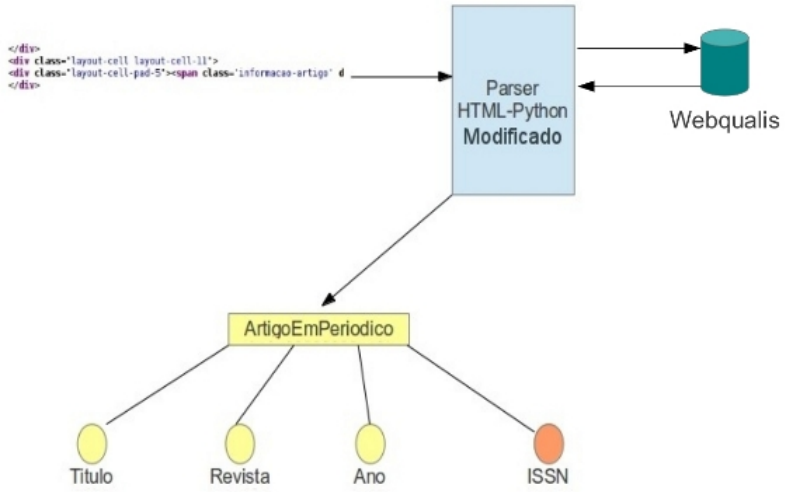

Figura 3.3: Funcionamento do *parser* modificado. O acesso a base de dados Webqualis foi adicionado ao processo.

Identificou-se no *parser* HTML-Python original o trecho de código correspondente à extração dos dados de periódicos. Para consultar a base Webqualis, foi usado como índice o campo ISSN, um identificador mundial único para periódicos, e, portanto, uma chave primária. Deste modo garante-se que não haverá problemas de se procurar por um periódico e pegar o estrato de outro, algo que poderia acontecer, por exemplo, caso o nome do periódico fosse usado na busca.

Foi necessário incluir um atributo novo *issn* na classe *ArtigoEmPeriodico* (do módulo *artigoEmPeriodico.py*) para armazenar esta nova informação. Um método também foi criado para consultar o arquivo com a base de periódicos Webqualis, conforme ilustrado pela Figura 3.4.

Esta consulta será usada para apresentar nos relatórios, junto aos dados do artigo, o estrato do periódico, conforme pode ser observado na Figura 3.5. Para tal, o módulo

```
# ------------------------------------------------------------------- #
    def getEstrato(self, strISSN):
            webQualisDB = csv.reader(open(sys.argv[2], 'rb'), delimiter = ';')
            lista = list(list(linha) for linha in webQualisDB)

            for i in range(0, len(lista)):
                if lista[i][0].find(strISSN[0:4] + "-" + strISSN[4:]) >= 0:
                    if lista[i][2] == 'A1':
                        scriptLattes.numA1 += 1
                    elif lista[i][2] == 'A2' :
                        scriptLattes.numA2 += 1
                    elif lista[i][2] == 'B1' :
                        scriptLattes.numB1 += 1
                    elif lista[i][2] == 'B2' :
                        scriptLattes.numB2 += 1
                    elif lista[i][2] == 'B3' :
                        scriptLattes.numB3 += 1
                    elif lista[i][2] == 'B4' :
                        scriptLattes.numB4 += 1
                    elif lista[i][2] == 'B5' :
                        scriptLattes.numB5 += 1

                    return lista[i][2]

            # Nao encontrou o periodico.
                return 'N/A'
# ------------------------------------------------------------------- #
```

Figura 3.4: Código da operação de consulta ao estrato.

de geração de relatórios também foi modificado. O estrato serve de apoio gerencial ao coordenador do programa de pós-graduação porque um dos critérios levados em consideração pela CAPES na avaliação de um programa considera a qualidade das publicações, de maneira que publicações em periódicos mais bem pontuados tendem a melhorar o desempenho do programa na avaliação da CAPES.

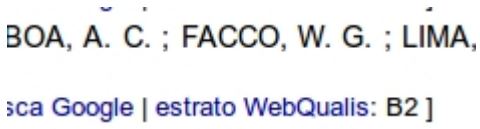

Figura 3.5: Destaque de parte de uma saída gerada pela ferramenta proposta. Pode-se perceber no destaque o estrato do periódico listado(B2).

3.5 Critérios para Avaliação dos Programas de Pós-Graduação

Cada programa de pós-graduação está ligado a uma área de conhecimento. No caso da CAPES, são listadas 76 áreas do conhecimento [3]. Cada área possui suas particulari-

dades, e por isso, possuem os seus próprios critérios de avaliação. Os documentos de área apresentam conceitos, critérios e diretrizes que norteiam o processo de avaliação adotados por cada área. Neste trabalho apresentamos uma implementação baseada no documento da área Interdisciplinar, preparado pela Coordenação de Área Interdisciplinar da CAPES (CAInter). O documento apresenta índices e critérios objetivos e subjetivos de avaliação, que juntos são utilizados para calcular a nota do programa de pós-graduação.

Assim, dado o caráter subjetivo da avaliação, diversos destes critérios não são possíveis de serem calculados automaticamente. Mesmo muitos dos critérios objetivos não podem ser calculados por não estarem disponíveis na Plataforma Lattes, e nem na Plataforma Webqualis, dependendo de informações que estão armazenadas em outras bases de dados não-públicas. Vamos, assim, nos ater aos indicadores que foram usados neste trabalho.

O Indicador de Formação Docente ($IndFor$) observa a participação de docentes permanentes bolsistas do CNPq; a diversidade de instituições onde os docentes permanentes concluíram o doutorado; o apoio a projetos por órgãos de fomento; a distribuição do corpo docente pelas áreas disciplinares que abrangem a proposta; o grau de inter-multidisciplinaridade, compatibilidade e integração do corpo docente com a Proposta do Programa [7]. É calculado da seguinte maneira:

$IndFor(\%) = \frac{(Form1+Form2+Form3)}{3}$, onde:

$Form1 = \frac{A}{B} \times 100\%$

A = Número de docentes permanentes que são bolsistas do CNPq

B = Número total de docentes permanentes

$Form2 = \frac{C}{D} \times 100\%$

C = Número de instituições onde os docentes permanentes concluíram o doutorado

D = Número total de docentes permanentes

$Form3 = \frac{E}{F} \times 100\%$

E = Número de docentes permanentes com projetos apoiados por órgãos de fomento

F = Número total de docentes permanentes

O Indicador de Orientações (*IndOri*) verifica o número de dissertações e teses defendidas e aprovadas no período e sua proporção em relação ao corpo docente permanente e ao corpo discente[7]. É calculado da seguinte forma:

$IndOri = \frac{(A+2\times B)}{C}$

A = Número de dissertações defendidas e aprovadas

B = Número de teses defendidas e aprovadas

C = Número total de docentes permanentes

Por fim, o Indicador de Produção Intelectual (*IndProd*) calcula a produção anual, por docente, de artigos em periódicos, de livros e de capítulos de livros. Por não estarem disponíveis as classificações nestes dois últimos, este trabalho considerou, para os cálculos de *IndProd*, apenas a produção de artigos em periódicos. O cálculo deste índice é apresentado abaixo:

$IndProd = \frac{1\times A1+0{,}85\times A2+0{,}7\times B1+0{,}55\times B2+0{,}4\times B3+0{,}25\times B4+0{,}1\times B5)}{D}$

A1, A2, B1, B2, B3, B4 e B5 = Número de artigos em periódicos classificados em cada um dos estratos. São desconsiderados periódicos classificados no estrato C.

D = Número total de docentes permanentes

No ScriptLattes, tudo é executado a partir do arquivo principal *scriptLattes.py*. Para facilitar o cálculo destes indicadores, neste trabalho preferiu-se usar o conceito de Variáveis Globais. Isso evitou a criação de novos módulos com novas classes apenas para armazenar esses indicadores. Contudo, à medida que o projeto crescer, essa opção deve ser reavaliada, até porque evitar usar variáveis globais é uma boa prática de programação em Python [11, 12].

Os módulos que precisaram de modificações para os cálculos descritos acima foram: *membro.py*(A), *formacaoAcademica.py*(B), *projetoDePesquisa.py*(C) e *geradorDePaginasWeb.py*(D). No caso A, foi incluída, no método *carregarDadosCVLattes()*, uma verificação condicional que atesta se o pesquisador é bolsista do CNPq ou não. Em caso

afirmativo, a variável global que controla a soma dos professores permanentes que são bolsistas do CNPq é incrementada (*numDocentesPermanentesBolsistas*). Esta informação já era buscada pelo *parser* original. Este é um dado importante no cálculo do *IndFor*. No caso B foi adicionado à classe *FormacaoAcademica* um método para calcular em quantas instituições diferentes os docentes permanentes concluíram o doutorado. O método *verificarDoutorado(strTipo)* recebe o texto extraído da Plataforma Lattes e faz uma busca textual pela string 'Doutorado em', que é a única forma de saber se um professor fez doutorado. Foi usada a variável global *numInstituicoesDocentesPermanentesDoutorado* para guardar a soma. Este é um dado importante no cálculo do *IndFor*. Adiante, no caso C, houve outra criação de método. A classe *ProjetoDePesquisa* recebeu um método, *contarProjetosApoioados(strDescricao)*, que verifica na *string* de entrada se um projeto é financiado por um órgão de fomento. Novamente, uma busca textual é feita procurando por algumas palavras-chave como, por exemplo, 'Fundação de Amparo à Pesquisa do Estado de Minas Gerais' (FAPEMIG), 'Conselho Nacional de Desenvolvimento Científico e Tecnológico'. Quando encontrado um projeto, a variável global *numDocentesPermanentesProjetosOrgaosFomento* é incrementada. Por fim, o módulo *python* do caso D é o responsável por toda geração de páginas de relatório. Nele foi preciso fazer as modificações para escrever os novos dados gerados para o usuário. O método *menuHTMLdeBuscaPB()* foi modificado para aceitar a escrita do estrato do periódico. O método *gerarPaginasDeOrientacoes* foi modificado para apresentar o cálculo do total de dissertações e teses concluídas e aprovadas no periódo de extração (*totalDissertacoesTesesConcluidasAprovadas*; usado para o $IndOri$). E o método *gerarPaginaPrincipal()* foi alterado para escrever os valores dos indicadores *IndFor* e $IndOri$.

4 Caso de Uso

4.1 Introdução

Este capítulo apresenta um caso de uso da nova ferramenta proposta neste trabalho envolvendo o grupo de pesquisa da Pós-Graduação em Modelagem Computacional (PGMC) da Universidade Federal de Juiz de Fora (UFJF).

Vale destacar que a base de dados do Webqualis usada corresponde apenas ao subconjunto de periódicos da área Interdisciplinar, visto que o PGMC é um programa ligado a esta área na CAPES. Assim, na seção correspondente aos artigos publicados poderá constar como estrato "N/A" para aqueles artigos que não estejam listados na base Interdisciplinar. Isso quer dizer apenas que o periódico não consta na base extraída e não que ele não exista ou tenha estrato registrado em outra área.

O período, a pesquisar, escolhido no arquivo de entrada *MC_UFJF.config*, está compreendido entre 2010 e 2012, ou seja, busca-se saber o que o grupo de pesquisa da PGMC produziu neste último triênio da CAPES.

Os IDs dos 14 professores permanentes, pertencentes aos quadros do PGMC, foram consultados a partir da página do programa, em $http://www.ufjf.br/mmc/corpo-docente/$, acessada no dia 23 de março de 2013.

Para executar o programa, foi adicionado mais um parâmetro ao *scriptLattes.py*, de modo que agora o nome do arquivo que contem a base de dados Webqualis também deve ser informada:

$scriptLattes.py \quad < nome\_arquivo\_de\_configuracao > \quad < nome\_base\_webqualis >$

Sendo assim, para esse caso de uso, o comando para execução da ferramenta fica:

$python \quad scriptLattes.py \quad ./exemplo/MC\_UFJF.config$
$./webqualis/lista\_completa\_periodicos\_qualis.txt$

O *download* dos fontes do trabalho podem ser baixados no seguinte endereço: $< https://code.google.com/p/script-lattes-webqualis/ >$

4.2 Resultados

A Figura 4.1 apresenta a página principal gerada pela ferramenta. Poucas modificações foram feitas nesta página em relação a versão original do ScriptLattes. A principal foi a inclusão dos resultados dos índices $IndFor$, $IndOri$ e $IndProd$, conforme ilustrado na figura 4.4.

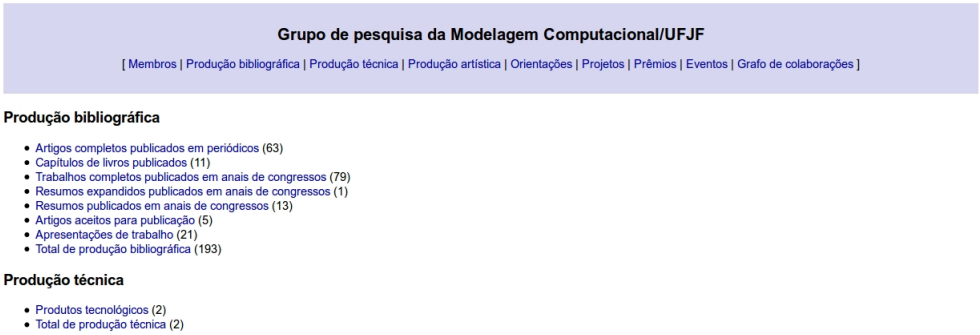

Figura 4.1: Página principal gerada pela ferramenta.

Outra importante modificação foi a inclusão dos estratos nos periódicos. A figura 4.2 ilustra um pedaço da página gerada, e a Figura 4.3 destaca a inclusão dos estratos após os dados de cada periódico.

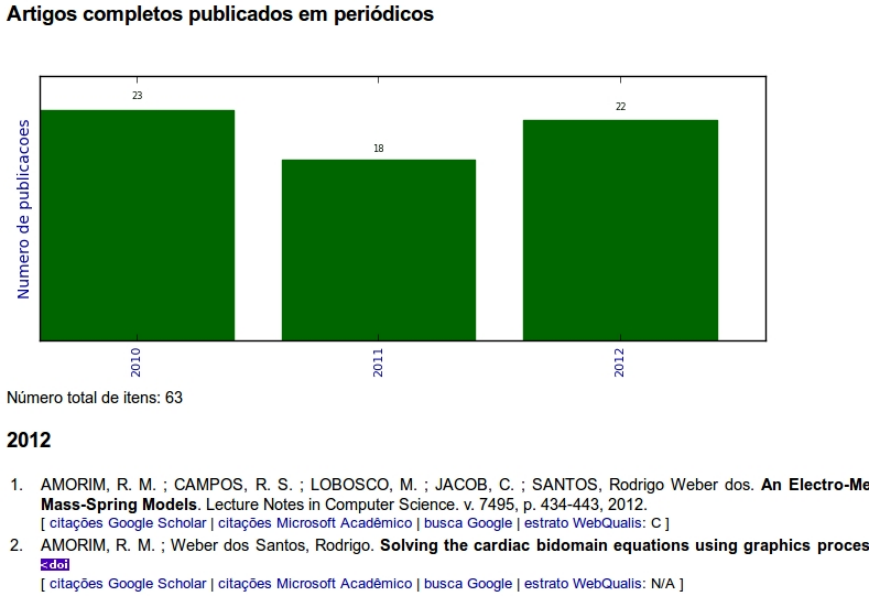

Figura 4.2: Quantidade de períodicos no triênio 2010-2012

Figura 4.3: Extratos dos periódicos em destaque.

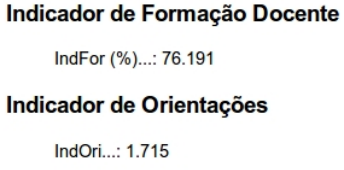

Figura 4.4: Alguns dos indicadores calculados pela ferramenta.

5 Conclusão

Este trabalho tinha por objetivo demonstrar a possibilidade de integração entre dados extraídos da Plataforma Lattes e o banco de dados de periódicos Webqualis, bem como a automação do processo de cálculo de alguns dos indicadores de qualidade dos programas de pós-graduação, segundo critérios estabelecidos por cada um dos comitês de área da CAPES.

A integração com o Webqualis foi concluída com grande êxito, conforme pôde ser visto no capítulo do caso de uso. O fato desta base ser pública, e portanto de livre consulta, foi um agente facilitador da integração. Além do mais, o que havia de informação pública na Plataforma Lattes que pudesse ser extraída para o cálculo de indicadores de avaliação foi também usada.

Uma melhoria futura, sem dúvida, de suma importância, é integração da nova ferramenta computacional proposta e implementada neste trabalho com a base de dados gerada pelo aplicativo "Coleta CAPES" (base não pública). Com os dados que estão contidos nesta base de dados seria possível calcular a maior parte dos indicadores propostos para avaliação dos programas de pós-graduação propostos pelo CAInter da CAPES.

Outras melhorias futuras seriam o aprimoramento das novas estruturas de dados implementadas, bem como a melhor modularização do código em função de suas novas funcionalidades, em especial para se evitar o uso de variáveis globais.

Deve-se ainda destacar uma limitação atual da ferramenta, que realiza uma pesquisa textual para descobrir se o projeto de pesquisa de determinado pesquisador é financiado por órgão de fomento. Na Plataforma Lattes não existe nenhuma representação única, por exemplo, usando-se identificadores únicos (IDs) para diferenciar os órgãos de fomento. Assim, a única forma de se implementar a consulta é buscando-se pelo nome do órgão de fomento como, por exemplo, 'Fundação de Amparo à Pesquisa do Estado de Minas Gerais'. Mas isso é um problema, porque no futuro podem surgir diversos novos órgãos de fomento, inclusive internacionais, de modo que eles não seriam identificados pela versão atual do programa. Para o estudo de caso deste trabalho, dois foram usa-

dos como exemplo: FAPEMIG e CNPq. Mas vale frisar que isto é uma limitação muito grande. Mesmo que todos os órgãos existentes hoje fossem cadastrados, novos órgãos que surgissem deveriam ter a oportunidade de serem consultados. Isso poderia ser feito através de um arquivo externo ao programa (que seria a melhor opção) ou a inclusão de cada um deles no módulo *projetoDePesquisa.py* (uma opção não recomendada).

Por fim, deve-se destacar que mesmo partindo-se de uma implementação já existente, a dificuldade em realizar este trabalho foi grande. Isto porque a aplicação tem dezenas de módulos e milhares de linhas de código sem documentação. A compreensão de seu funcionamento, do seu fluxo de dados (do HTML para a linguagem *Python*) por si só, foi uma tarefa árdua. Isso sem contar que o código original não era perfeitamente funcional: foram descobertos alguns *bugs* na aplicação.

Contudo, o resultado final foi compensador: a ferramenta proposta com certeza será importante para a gestão dos programas de pós-graduação *stricto sensu*. E com sua integração a base gerada pelo "Coleta CAPES", a tendência é que ela se torne ainda mais importante.

Referências Bibliográficas

[1] Grune, D.; Jacobs, C.; Langendoen, K. ; Bal, H. **Modern Compiler Design**. 1st. ed., New York, NY, USA: John Wiley & Sons, Inc., 2000.

[2] Meduna, A. **Elements of Compiler Design**. 1st. ed., Boston, MA, USA: Auerbach Publications, 2007.

[3] Assessoria de Comunicação Social, C. **Tabela de Áreas de conhecimento**, julho de 2007, http://www.capes.gov.br/avaliacao/tabela-de-areas-de-conhecimento, Acessado em 22/03/2013.

[4] 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.

[5] Assessoria de Comunicação Social, C. **Cresce 20,8% o número de cursos de mestrados e doutorados no brasil**. Página da Internet, setembro de 2010, http://www.capes.gov.br/servicos/sala-de-imprensa/36-noticias/4073-cresce-208-o-numero-de-cursos-de-mestrados-e-doutorados-no-brasil, Acessado em 22/03/2013.

[6] Braziliense, C. **Número cursos de mestrado e doutorado com baixo desempenho aumenta**. Página da Internet, setembro de 2010, http://www.correiobraziliense.com.br/app/noticia/brasil/2010/09/15/interna_brasil,213042/inc Acessado em 22/03/2013.

[7] et alli, A. P. J. **Documento de Área 2009**, agosto de 2010, http://www.capes.gov.br/images/stories/download/avaliacao/INTER03ago10.pdf, Acessado em 22/03/2013.

[8] Jesús Pascual Mena-Chalco, R. M. C. J. **Prospecção de dados acadêmicos de currículos lattes através de scriptlattes**. Página da Internet, 2011, http://professor.ufabc.edu.br/ jesus.mena/publications/pdf/scriptLattes-2011-bibliometria.pdf, Acessado em 22/03/2013.

[9] Mena-Chalco, J. P.; Cesar-Jr., R. M. **Breve descrição do scriptlattes**. Página da Internet, 2013, http://scriptlattes.sourceforge.net/description.html, Acessado em 22/03/2013.

[10] Mena-Chalco, J. P.; Cesar-Jr., R. M. **scriptlattes: Uma ferramenta para extração e visualização de conhecimento a partir de currículos lattes**. Página da Internet, 2013, http://scriptlattes.sourceforge.net/faq.html, Acessado em 22/03/2013.

[11] Amit Patel, Antoine Picard, E. J. J. H. M. S. M. S. **Google python style guide**. Página da Internet, 2013, http://google-styleguide.googlecode.com/svn/trunk/pyguide.html, Acessado em 24/03/2013.

[12] Guido van Rossum, B. W. **Style guide for python code**. Página da Internet, 2013, http://www.python.org/dev/peps/pep-0008/#global-variable-names, Acessado em 24/03/2013.