ANÁLISE MULTIDIMENSIONAL DA PRODUÇÃO CIENTÍFICA EM CIÊNCIA DA COMPUTAÇÃO

MARCO AURÉLIO MODESTO BARRETO

ANÁLISE MULTIDIMENSIONAL DA PRODUÇÃO CIENTÍFICA EM CIÊNCIA DA COMPUTAÇÃO

Dissertação apresentada ao Programa de Pós-Graduação em Ciência da Computação do Instituto de Ciências Exatas da Universidade Federal de Minas Gerais como requisito parcial para a obtenção do grau de Mestre em Ciência da Computação.

ORIENTADOR: NIVIO ZIVIANI

Belo Horizonte

Julho de 2009

Barreto, Marco Aurélio Modesto
M1234a Análise Multidimensional da Produção Científica em Ciência da Computação / Marco Aurélio Modesto Barreto. — Belo Horizonte, 2009
xviii, 68 f. : il. ; 29cm

Dissertação (mestrado) — Universidade Federal de Minas Gerais
Orientador: Nivio Ziviani

1. Armazéns de Dados. 2. Bancos de Dados. 3. Bibliometria. I. Título.

CDU 519.6*82.10

UNIVERSIDADE FEDERAL DE MINAS GERAIS
INSTITUTO DE CIÊNCIAS EXATAS
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

FOLHA DE APROVAÇÃO

Análise multidimensional da produção científica em Ciência da Computação

MARCO AURÉLIO BARRETO MODESTO

Dissertação defendida e aprovada pela banca examinadora constituída pelos Senhores:

PROF. NIVIO ZIVIANI - Orientador
Departamento de Ciência da Computação - UFMG

PROF. JAIME ARTURO RAMIREZ
Departamento de Engenharia Elétrica - UFMG

PROF. ALBERTO HENRIQUE FRADE LAENDER
Departamento de Ciência da Computação - UFMG

PROF. WAGNER MEIRA JÚNIOR
Departamento de Ciência da Computação - UFMG

Belo Horizonte, 28 de julho de 2009.

À minha mãe (in memorian).

Resumo

O número de programas de pós-graduação em Ciência da Computação no Brasil cresceu mais de 200% nos últimos doze anos, tendo aumentado de 17 para 50 programas. Em face desse crescimento é importante avaliar como a qualidade da produção científica desses programas se compara à de programas congêneres em outros países. O objetivo desta dissertação é realizar uma análise da produção científica em Ciência da Computação, a partir de dados coletados sobre os oito principais programas no país de acordo com dados oficiais da CAPES para o triênio 2004-2006 e 22 dos mais importantes programas norte-americanos e europeus. Para isso, o primeiro passo foi projetar e implementar um armazém de dados de publicações científicas dos 30 programas para permitir análises multidimensionais de dados bibliográficos. O armazém de dados proporciona uma análise dos dados sob diversas dimensões, o que permite obter estatísticas sobre os programas de pós-graduação abordados pelo estudo, tais como: média de publicações por docente por programa, distribuição das publicações entre as subáreas da Ciência da Computação, subáreas mais populares por programa, média de citações por artigo por programa, indicadores de produtividade por docente, por programa ou por país, entre outras. Dessa forma, o arcabouço proposto permitiu a análise do perfil de publicação dos programas brasileiros em relação ao perfil dos principais programas da América do Norte e da Europa. Os resultados mostram que a produção científica dos programas nacionais, representada por artigos publicados em periódicos e conferências internacionais, é comparável em volume, qualidade e impacto a de alguns dos principais programas da América do Norte e da Europa.

Abstract

The amount of Computer Science graduate programs in Brazil has increased over 200% in the past twelve years, from 17 to 50 programs. In light of this growth, it is important to assess how the quality of the scientific production of these programs compares to their equivalent in other countries. The objective of this dissertation is to perform an analysis of the Computer Science scientific production, using data collected from the eight major programs in the country according to official data from CAPES for the triennium 2004-2006, and 22 of the most important programs in North America and Europe. Therefore, the first step was to design and implement a data warehouse of scientific publications from the 30 programs to allow multidimensional analysis of bibliographic data. The data warehouse provides an opportunity for analysis using several dimensions, which allows the acquisition of statistical data regarding the graduate programs related in this study, such as: average of publications per professor per program, publications distribution among subareas of Computer Science, most popular subareas per program, average of citations per article per program, productivity measurements per professor, per program or per country, among others. Thus, the data warehouse proposed allowed the analysis of the Brazilian programs publication profile, compared to major programs in North America and Europe. Results demonstrate that national scientific production, represented by articles published in journals and international conferences, is comparable in volume, quantity and impact to some of the major programs in North America and Europe.

Lista de Figuras

Lista de Tabelas

Sumário

Capítulo 1

Introdução

1.1 Motivação

Segundo dados da CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior), o número de programas de pós-graduação em Ciência da Computação no Brasil cresceu mais de 200% nos últimos doze anos, conforme ilustrado na Figura 1.1. Em face desse crescimento torna-se interessante fazer uma avaliação qualitativa da produção e da inserção internacional dos principais programas da área no Brasil, de modo que se possa estabelecer o papel que a área vem desempenhando no desenvolvimento científico e tecnológico do país. Em vista disso, foi criado na Universidade Federal de Minas Gerais (UFMG) um grupo para estudar o perfil da produção científica dos principais programas brasileiros de modo a compará-los com seus congêneres da América do Norte e da Europa.

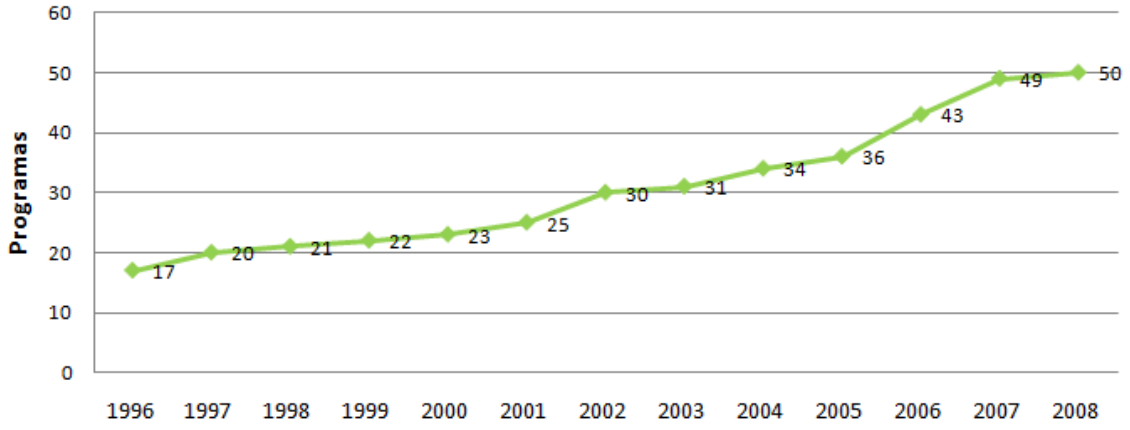


Figura 1.1. Número de programas em Ciência da Computação no Brasil.

Como consequência nasceu o Projeto Perfil-CC[1], cujo objetivo principal é comparar a produção científica dos oito principais programas da área do país de acordo com a avaliação da Capes para o triênio 2004-2006 com a produção de vinte e dois dos mais importantes programas da América do Norte e da Europa. Para realizar o estudo, o Projeto Perfil-CC utiliza a *DBLP Computer Science Bibliography*[2], biblioteca digital que contempla o maior volume de publicações da área de Ciência da Computação em todo o mundo [Petricek et al., 2005; Ley, 2002].

Um armazém de dados (do inglês *data warehouse*) é um repositório utilizado para armazenar dados relativos às atividades de uma organização, de forma consolidada. O desenho desse repositório favorece os relatórios, a análise de grandes volumes de dados e a obtenção de informações estratégicas que podem facilitar a tomada de decisão. O armazém de dados permite uma análise dos dados sob diversas perspectivas ou dimensões, de forma flexível e bastante ágil. A facilidade de analisar os dados sob diversas dimensões leva ao conceito de análise multidimensional. A ferramenta mais popular para exploração de um armazém de dados é a *Online Analytical Processing* (OLAP) ou Processo Analítico em Tempo Real [Kimball et al., 2008].

O objetivo desta dissertação é realizar uma análise multidimensional da produção científica em Ciência da Computação. O primeiro passo foi projetar e implementar um armazém de dados de publicações científicas em Ciência da Computação para permitir análises multidimensionais dos dados gerados pelo Projeto Perfil-CC. A partir do armazém de dados podem ser obtidas diversas estatísticas sobre os 30 programas abordados pelo estudo, tais como: média de publicações por docente por programa, distribuição das publicações entre as subáreas da Ciência Computação, subáreas mais populares por programa, média de citações por artigo por programa, indicadores de produtividade por docente, por programa ou por país, entre outras.

Neste contexto, foram coletados dados bibliográficos da DBLP para cada docente relacionado aos 30 programas considerados no Projeto Perfil-CC. A seguir foi utilizado o repositório de publicações Google Scholar[3] para obter o número de citações recebidas pelos artigos, o qual foi adicionado a um banco de dados relacional. Finalmente, os dados consolidados do banco de dados relacional foram transformados em um armazém de dados para permitir uma análise multidimensional dos dados, de forma *online* e em tempo real.

[1] http://latin.dcc.ufmg.br/perfilcc
[2] http://dblp.uni-trier.de
[3] http://scholar.google.com. Biblioteca digital que contempla publicações de todas as áreas.

1.2 Trabalhos Relacionados

O estudo do perfil de publicação em Ciência da Computação tem recebido uma certa atenção nos últimos dois anos. Em Laender et al. [2008], a produção científica de programas brasileiros de pós-graduação em Ciência da Computação é comparada com programas congêneres na América do Norte e na Europa. Para cada um dos programas, foram coletados da DBLP os dados referentes à produção científica de seus corpos docentes. A pesquisa conclui que a produção científica da área é fortemente focada em artigos, publicados em anais de conferências em uma relação superior a dois artigos em conferência para um em periódico, tanto aos programas nacionais quanto aos internacionais. Além disso, o trabalho conclui que os programas brasileiros têm desempenho semelhante aos programas norte-americanos e europeus em termos de publicações científicas e número de doutores formados.

Menezes et al. [2008] analisam o perfil de publicação científica da área utilizando a rede de colaboração formada entre os seus pesquisadores. O artigo mostra, por exemplo, que a colaboração mútua entre os pesquisadores europeus é menor que a dos brasileiros e a dos norte-americanos. A medição do coeficiente de aglomeração[4] mostrou diferenças na diversidade dos relacionamentos da rede formada pelas subáreas. Numa linha próxima, Menezes et al. [2009] analisam e mesma rede e faz uma análise temporal das publicações das três regiões estudadas (Brasil, América do Norte e Europa). O estudo mostra a evolução do número de pesquisadores por artigo, o crescimento do tamanho da rede e as divergências das redes formadas pelos novos campos e pelos campos tradicionais da área. Um estudo semelhante foi feito para as áreas de Biologia, Física e Matemática [Newman, 2004]. Nascimento et al. [2003] analisa o grafo de colaboração obtido por todos os artigos publicados entre 1975 e 2002 na conferência ACM SIGMOD (*International Conference on Management of Data*). Entre os resultados, são identificados os autores mais "próximos" a todos os demais. Além disto, é mostrado que a rede formada pelos autores da conferência SIGMOD é mais um exemplo de grafo que segue o fenômeno *small world*[5].

Martins et al. [2009] apresentam um método para classificação automática da qualidade de conferências utilizando técnicas de aprendizado de máquina. Tal estudo utilizou técnicas de aprendizado de máquina que testaram o uso de várias variáveis para a solução como o número de citações, o número de submissões das conferências,

[4]Métrica que mede a conectividade da rede. Determinado pela razão entre o número de conexões entre vizinhos comuns a um nó de referência, dividido pelo número de possíveis conexões entre os vizinhos comuns ao nó.

[5]Características de redes com alto grau de agrupamento (conectividade) e baixa distância média entre os vértices, independente do tamanho e da densidade da rede.

as taxas de aceitação, a tradição da conferência e a reputação dos membros do comitê de programa. Concluiu-se que as mais importantes são as citações e a tradição da conferência.

Arruda et al. [2009] fazem um estudo sobre o perfil dos pesquisadores brasileiros em Ciência da Computação. Basicamente analisam a produção dos pesquisadores em termos da localização geográfica dos seus programas e do sexo. Foram identificadas subáreas com uma presença feminina expressiva, como por exemplo, Interação Humano-Computador e Inteligência Artificial e subáreas predominantemente masculinas como Hardware e Redes de Computadores. Nas subáreas com presença feminina, concluiu-se que as pesquisadoras têm mais publicações que os pesquisadores. Nas outras áreas, a diferença não foi estatisticamente significativa. Em relação à região, concluiu-se que os pesquisadores estão concentrados no Sudeste e Sul do Brasil, regiões que também obtêm as maiores taxas de publicação por pesquisador. Wainer et al. [2009], em um trabalho complementar do mesmo grupo de pesquisa, fazem um estudo comparativo sobre publicações em Ciência da Computação no Brasil e em outros países, medido pelo número de artigos em periódicos e conferências indexados pelo ISI (*Institute for Scientific Information*). Entretanto, como veremos mais adiante, esta base não se mostra adequada para estudar as publicações em Ciência da Computação [Mattern, 2006].

1.3 Objetivos

O objetivo deste trabalho é analisar o perfil de publicação científica da área de Ciência da Computação, comparando-o com o perfil de programas de grande prestígio internacional da América do Norte e da Europa. São analisados o número de publicações por docente, a qualidade das publicações e diversas métricas de citações. Também são analisadas a distribuição das subáreas da Ciência da Computação entre os programas analisados. O objetivo deste estudo é analisar o comportamento dos programas brasileiros em relação ao comportamento dos principais programas da América do Norte e da Europa, ressaltando as principais diferenças entre as subáreas.

Para realizar o estudo dos dados são utilizadas técnicas de análise multidimensional por meio da implementação de um armazém de dados de publicações científicas. Um armazém de dados fornece um método para se acessar, visualizar e analisar uma grande quantidade de dados com alta flexibilidade e desempenho, além de disponibilizar recursos de criação automática de gráficos e exportação dos dados para uma planilha eletrônica. A principal vantagem do ambiente proposto é permitir a obten-

ção de respostas rápidas a consultas de natureza tipicamente dimensional, permitindo uma análise flexível, ágil e mais rica dos dados gerados. Algumas das vantagens da utilização do ambiente proposto são:

- Maior flexibilidade para analisar os dados, permitindo diferentes visões;
- Obtenção de informação mais rica;
- Facilidade para criação de consultas sem depender de um especialista que conheça a linguagem SQL[6];
- Geração automática de gráficos a partir dos filtros aplicados e das visões selecionadas;
- Flexibilidade na granularidade dos dados, permitindo tanto a análise macro (alto nível) dos dados, como por exemplo "o número de publicações por professor", quanto a análise micro (baixo nível), como por exemplo, "a listagem de todas as X publicações do professor Y";
- Exportação dos dados por meio de planilhas eletrônicas, permitindo que os usuários possam manipulá-los para novas análises.

Em relação às dimensões disponibilizadas para análise, podemos citar:

- Programa de pós-graduação (Ex.: UFMG, Princeton, Paris VI).
- Localização geográfica do programa (Ex.: Brasil, América do Norte ou Europa).
- Docente do programa (Ex.: Virgílio Almeida).
- Autor de um artigo (pode ou não ser um docente dos programas analisados)(Ex.: Cristiano Cazita).
- Artigo publicado (Ex.: "Characterizing a spam traffic").
- Veículo onde um artigo foi publicado (Ex.: ACM SIGCOMM).
- Classificação Qualis do veículo de publicação (Ex.: AI[7]).
- Ano da publicação (Ex.: 2004).

[6] *Structured Query Language*) - Linguagem de consulta declarativa para bancos de dados relacional.

[7] Avaliação da qualidade do veículo de publicação fornecida pela CAPES. AI corresponde a um veículo de alta (A) qualidade e de circulação internacional (I)

O ambiente gerado foi disponibilizado na Web[8], sendo facilmente adaptável para análise de outras bases de dados. Este trabalho utilizou ferramentas gratuitas e de código aberto, por se adequarem melhor aos recursos existentes no laboratório onde o ambiente foi disponibilizado e por possuírem, em geral, custos mais baixos.

1.4 Contribuições

Conforme já mencionado, a principal contribuição deste trabalho é uma análise multidimensional do perfil de publicação dos programas brasileiros em relação ao perfil dos principais programas da América do Norte e da Europa, ressaltando as principais diferenças entre as subáreas destes programas. Podemos citar as seguintes contribuições específicas:

1. Construção de um armazém de dados de publicações científicas disponibilizado na Web (vide Capítulo 4). A construção gerou um ambiente com uma base com grande qualidade dos dados, além de permitir a obtenção de diversas estatísticas sobre os 30 programas abordados com flexibilidade de análise.

2. Análise do perfil de publicação sob a ótica do número de publicações por programa, comparando os programas brasileiros, os norte-americanos e os europeus. Discutimos esta análise na Seção 5.1. Concluímos que a produção científica dos programas nacionais, representada por artigos publicados em periódicos e conferências internacionais, é comparável em volume e qualidade a de alguns dos principais programas da América do Norte e da Europa.

3. Análise das métricas de citações das publicações dos programas. Esta discussão é apresentada na Seção 5.2. O número de citações por artigo para os dois programas brasileiros mais bem colocados é comparável ao impacto de alguns dos principais programas da América do Norte e da Europa. Considerando o índice h, os oito programas brasileiros possuem um índice h compável a de vários programas da América do Norte e Europa.

4. Análise da distribuição das subáreas dos programas analisados. Este tópico é detalhado na Seção 5.3. Foi possível notar que cada programa "especializa-se" numa determinada subárea e que a distribuição das subáreas é heterogênea. Além disto, as áreas mais tradicionais da Ciência da Computação são as que possuem mais artigos: Arquitetura de Computadores, Redes de Computadores, Bancos de Dados, Algoritmos e Inteligência Artificial.

[8] http://www.latin.dcc.ufmg.br:8080/perfilccDW/

1.5 Indicadores Bibliométricos

Indicadores bibliométricos vêm sendo empregados para medir a atividade científica com base na análise estatística dos dados quantitativos obtidos da literatura científica e técnica. Esses indicadores têm sido usados para quantificar características como a qualidade, a importância e o impacto de artigos, periódicos, autores e instituições na pesquisa.

Por meio de indicadores bibliométricos é possível determinar, entre outros aspectos:

- O crescimento de qualquer campo da ciência, segundo a variação cronológica do número de trabalhos publicados nesse campo;
- O envelhecimento dos campos científicos [Alvarado, 2009], segundo a vida média das referências de suas publicações;
- A evolução cronológica da produção científica, segundo o ano de publicação dos documentos;
- A produtividade dos autores ou instituições, medida pelo número de seus trabalhos publicados;
- A colaboração entre os pesquisadores ou instituições, medida pelo número de autores por trabalho ou centros de investigação que colaboram;
- O impacto ou visibilidade das publicações dentro da comunidade científica internacional, medido pelo número de citações que recebem em trabalhos posteriores;
- A análise e avaliação das fontes difusoras dos trabalhos, por meio de indicadores de impacto das fontes;
- A dispersão das publicações científicas entre as diversas fontes e outros.

Um dos indicadores bibliométricos mais utilizados é o fator de impacto [Garfield & Merton, 1979], também chamado de SCI (*Science Citation Index*), criado pelo ISI (*Institute for Scientific Information*). Esse indicador tem sido utilizado durante décadas pela comunidade acadêmica para avaliar periódicos em diversos campos científicos, incluindo a área de Ciência da Computação. O fator de impacto oferece uma forma de avaliar ou comparar a importância relativa dos artigos de um periódico em relação a artigos de outros periódicos do mesmo campo, sendo aplicado sobre o conjunto de artigos de um periódico para avaliar indiretamente esse periódico. O cálculo do fator

de impacto é feito dividindo-se o número de citações de artigos publicados em um determinado ano em artigos dos dois anos anteriores, pelo número de artigos publicados nesses dois anos. Por exemplo, o fator de impacto de um periódico X em 2004 é determinado pela soma das citações recebidas em 2004 dos artigos publicados em 2002 e 2003. Assim, se esse periódico publicou 542 artigos em 2002 e 543 em 2003, e em 2004 esses artigos receberam 4.122 citações, o fator de impacto deste periódico é 3,799 (4.122 / (542 + 543)).

Uma das desvantagens do fator de impacto é sua cobertura limitada para determinadas áreas como a Ciência da Computação. De acordo com Mattern [Mattern, 2006], os dados do ISI não se mostram adequados para estudos bibliométricos na área de Ciência da Computação já que focam principalmente as áreas de Ciências Naturais e Ciências da Vida, e abrangem um número reduzido de conferências. Ainda segundo Mattern [Mattern, 2006], um levantamento feito com base na produção científica de 2003 do ETH de Zurique mostrou que o banco de dados do ISI cobria apenas 14% das publicações em Ciência da Computação daquela instituição enquanto a cobertura nas áreas de Física, Química e Biologia era bem maior, alcançando patamares em torno de 60%.

O índice h (do inglês h-index) é outro indicador proposto para quantificar a produtividade e o impacto de cientistas baseando-se nos seus artigos mais citados. Foi proposto em 2005 por Jorge E. Hirsch como uma ferramenta para determinar a qualidade relativa dos trabalhos de físicos teóricos [Hirsch, 2005]. A vantagem do índice h em relação a outras métricas de citações é que ele não é influenciado por poucos artigos de grande visibilidade. O índice é determinado pelo número de artigos com citações maiores ou iguais a esse número. Por exemplo: um pesquisador com h = 5 tem 5 artigos que receberam 5 ou mais citações; um programa de pós-graduação com h = 20 tem 20 artigos com 20 ou mais citações; e assim por diante.

O índice h também pode ser aplicado para estimar a produtividade e o impacto de um grupo de cientistas, um programa de pós-graduação, um país, e assim por diante. Apesar de ainda ter que provar seu valor e suplantar outras métricas tradicionais, como a enumeração do número de artigos, enumeração do número de citações e fator de impacto dos periódicos nos quais se publica, o índice h está ganhando cada vez mais adeptos.

Hirsch comparou o índice h com outros índices comumente usados para analisar a produção cientifica de um pesquisador e fez as seguintes observações:

1. Número total de artigos. Vantagem: mede a produtividade. Desvantagem: não mede a importância e o impacto de cada artigo.

2. Número total de citações. Vantagem: mede o impacto total do pesquisador. Desvantagem: pode ser insuflado por um pequeno número de artigos de grande visibilidade, os quais podem não ser representativos do indivíduo se ele é um co-autor com vários outros autores nos artigos.

3. Citações por artigo. Vantagem: permite a comparação de cientistas de diferentes idades. Desvantagem: privilegia a baixa produtividade e penaliza a alta produtividade.

Alguns autores ressaltam que o índice h, quando tomado de modo absoluto, não pode ser usado para comparar pesquisadores de diferentes áreas. Um índice h considerado bom em determinada área, em outras pode não ser tão bom assim ou mesmo ser considerado ruim. Os maiores valores de índice h são encontrados entres pesquisadores ligados às ciências da vida.

Além do fator de impacto e do índice h, vários outros indicadores aparecem propostos na literatura, como por exemplo, o *Weighted PageRank* e o *G-index*. Entretanto, uma descrição e discussão desses indicadores foge ao escopo desta dissertação e podem ser encontradas em Martins [2009].

1.6 Estrutura da Dissertação

Esta dissertação está estruturada da seguinte forma. O Capítulo 2 apresenta uma visão geral sobre armazéns de dados. O Capítulo 3 detalha os dados sobre publicações utilizados neste trabalho. O Capítulo 4 descreve a construção do armazém de dados de publicações científicas. O Capítulo 5 analisa os dados por meio do ambiente construído. Finalmente, as conclusões e os trabalhos futuros são apresentados no Capítulo 6.

Capítulo 2

Uma Visão Geral Sobre Armazéns de Dados

Uma importante questão estratégica para o sucesso de uma organização está relacionada com a sua capacidade de analisar e reagir rapidamente a mudanças nas condições de seus empreendimentos. Para que isso ocorra, torna-se necessário que a organização disponha de uma quantidade maior de informação qualificada. Os avanços na área de tecnologia da informação estão possibilitando que essas organizações possam manipular um grande volume de dados.

As informações encontram-se geralmente espalhadas por diferentes sistemas e exigem um esforço considerável para serem integradas e poderem dar apoio efetivo à tomada de decisão. Embora tenham ocorrido avanços tecnológicos nas áreas de armazenamento e manipulação de dados, ainda se observa uma enorme deficiência na obtenção de informações estratégicas que possam auxiliar o processo decisório [Kimball et al., 2008].

Em vista disso, tecnologias que suportam a análise de informações vêm ganhando destaque na atualidade. Uma delas é o processo de armazenagem de dados (do inglês *data warehousing*), que oferece às organizações uma maneira flexível e eficiente de obter informações a partir dos dados que apoiem seus processos de tomada de decisão.

Com o objetivo de apresentar uma visão geral desses conceitos, este capítulo está estruturado como se segue. A Seção 2.1 enfatiza os principais conceitos relacionados a armazém de dados. A Seção 2.2 mostra os esquemas de dados utilizados em armazém de dados. A Seção 2.3 detalha as consultas OLAP - Processo Analítico em Tempo Real.

2.1 Análise Multidimensional

Os sistemas de gerenciamento de bancos de dados evoluíram bastante nas últimas décadas, agregando funções ao seu objetivo principal de manter e disponibilizar dados para as aplicações computacionais. Ao mesmo tempo, ocorreu a especialização dos sistemas de gerenciamento de bancos de dados para o tratamento diferenciado do armazenamento e acesso às informações, levando em conta não somente a natureza e relacionamento dos dados, mas também as necessidades das aplicações. Neste contexto, surgiram os sistemas de apoio à decisão com seus sistemas de gerenciamento de bancos de dados e ferramentas específicas para a manipulação de informações analíticas. Entre estas ferramentas destaca-se a tecnologia de armazenagem de dados que é considerada a evolução natural dos ambientes de apoio à decisão [Kimball et al., 2008].

Um armazém de dados é um sistema de computação utilizado para armazenar informações relativas às atividades de uma organização em bancos de dados. O desenho do banco de dados favorece os relatórios, a análise de grandes volumes de dados e a obtenção de informações estratégicas que podem facilitar a tomada de decisão. Esse banco de dados é preparado em vários níveis de granularidade e obtido a partir de outros sistemas computacionais da organização (sistemas legados). A ideia é extrair dados analíticos dos sistemas de produção, transformá-los e armazená-los em vários graus de relacionamento e sumarização, de forma a facilitar e agilizar os processos de tomada de decisão. Os dados armazenados em um armazém de dados são analisados por meio de ferramenta específica para *Online Analytical Processing* (OLAP) [Barbieri, 2001].

2.1.1 Esquema Dimensional

Nesta seção apresentamos uma estrutura de dados que é eficiente para análise de dados. Ela permite o cruzamento de dados de forma intuitiva e garante flexibilidade nas consultas dos usuários.

Nos bancos de dados relacionais convencionais, a redundância dos dados é evitada, sendo aceita somente de forma controlada nos casos em que é realmente necessária. Esta redundância é eliminada por meio de processos de normalização, onde cada tabela do banco de dados que possua dados redundantes é dividida em tabelas distintas, originando, deste modo, apenas tabelas contendo dados não redundantes.

A normalização das tabelas traz benefícios nos casos em que muitas transações são efetuadas, pois estas se tornam mais simples e rápidas. Já no caso das aplicações que utilizam a análise OLAP ocorre o contrário: as transações operam sobre um grande

volume de dados e não são simples nem frequentes, não sendo conveniente a normalização das tabelas, pois no ambiente de análise OLAP ocorrem poucas transações concorrentes e cada transação acessa um grande número de registros.

Outro ponto que distingue os bancos de dados relacionais das aplicações OLAP está relacionado com a modelagem dos dados. As aplicações OLAP não utilizam o esquema relacional tradicional, como ocorre com os bancos de dados, pois este esquema é utilizado no projeto de bancos de dados com dados não redundantes. Elas utilizam o esquema dimensional.

Diferente do esquema relacional, o esquema dimensional é muito assimétrico. Nele existe uma grande tabela "dominante" no centro do esquema, a qual se conecta com as demais por meio de múltiplas junções, enquanto o restante das tabelas se liga à tabela central por meio de uma única junção. A tabela central é chamada de tabela fato e as demais tabelas são chamadas de tabelas de dimensões. As junções só ocorrem entre as tabelas Fato e Dimensões, melhorando o desempenho para a leitura. O esquema dimensional possui uma estrutura mais intuitiva, permitindo uma consulta mais fácil para usuários não especializados [Kimball & Ross, 2002].

Um exemplo de esquema dimensional é apresentado na Figura 2.1. Trata-se de uma modelagem que representa um banco de dados com dados bibliográficos de pesquisadores da área de Ciência da Computação. As informações extraídas são simplesmente contagens das ocorrências das dimensões, como por exemplo, o número de artigos, número de autores e o número de veículos de publicações. Na tabela fato, cada linha representa um artigo escrito por um professor (autor) de um dos programas avaliados. Uma possível medida para o fato seria o número de citações contabilizadas para cada artigo.

Figura 2.1. Exemplo de um esquema dimensional de uma base bibliográfica.

Normalmente, a tabela de dimensões contém uma única chave primária e vários atributos que descrevem essa dimensão com detalhes. Na tabela fato, a chave primária

é a combinação das demais chaves primárias das tabelas de dimensão, constituindo-se, assim, de várias chaves estrangeiras, de acordo com o número de dimensões. Os dados pertencentes à tabela fato são normalmente numéricos. O esquema dimensional é também chamado de esquema estrela (do inglês *star schema*), devido ao formato com que são dispostas as tabelas do diagrama com a tabela fato no centro e um conjunto de tabelas de dimensão nas extremidades (o que pode ser percebido na Figura 2.1).

Segundo Pedersen [Pedersen et al., 2000], o esquema multidimensional permite a realização de consultas visuais e suporta a semântica do esquema, podendo automaticamente escolher as funções mais adequadas para agregar em um nível mais alto os dados que manipula. Para tanto, os dados são organizados em cubos de diversas dimensões.

Cada dimensão consiste em um conjunto de descritores categóricos organizados em estruturas hierárquicas [Messaoud et al., 2004]. O usuário pode realizar operações no cubo, agregando dados em dimensões superiores (*roll-up*), desagregando-os, descendo nas inferiores (*drill-down*), ou selecionando e projetando dados (*slice-and-dice*). A abordagem dimensional permite o uso automático de funções de agregação e de consulta visual, além de bom desempenho e do fato de ser mais natural para a análise de dados [Pedersen et al., 2002].

As duas tecnologias principais para a construção de cubos multidimensionais são a ROLAP (OLAP Relacional) e a MOLAP (OLAP Multidimensional) [Shoshani, 1997]. A primeira usa bancos de dados relacionais tradicionais, nos quais os dados são armazenados em tabelas esquematizadas na forma de uma estrela (do inglês *star schema*) ou de flocos de neve (do inglês *snow flake schema*). A segunda normalmente utiliza estruturas de dados proprietárias para armazenar o cubo de dados, tornando o processamento mais rápido. Esta última, no entanto, não se integra naturalmente à tecnologia existente, exigindo uma importação dos dados para o sistema multidimensional proprietário.

2.1.2 ETL - Extração, Transformação e Carga dos Dados

No ambiente de um armazém de dados os dados são inicialmente extraídos de sistemas operacionais e de fontes externas. A seguir são integrados e transformados (limpos, eliminados, combinados, validados, consolidados, agregados e sumariados) antes de serem carregados no armazém de dados. Esta é uma etapa crítica da construção de um armazém de dados, pois envolve toda a movimentação dos dados. A mesma se dá basicamente em três passos, conhecidos como ETL: Extração (*Extraction*), Transformação (*Transformation*) e Carga (*Loading*). A Figura 2.2 exemplifica esse processo.

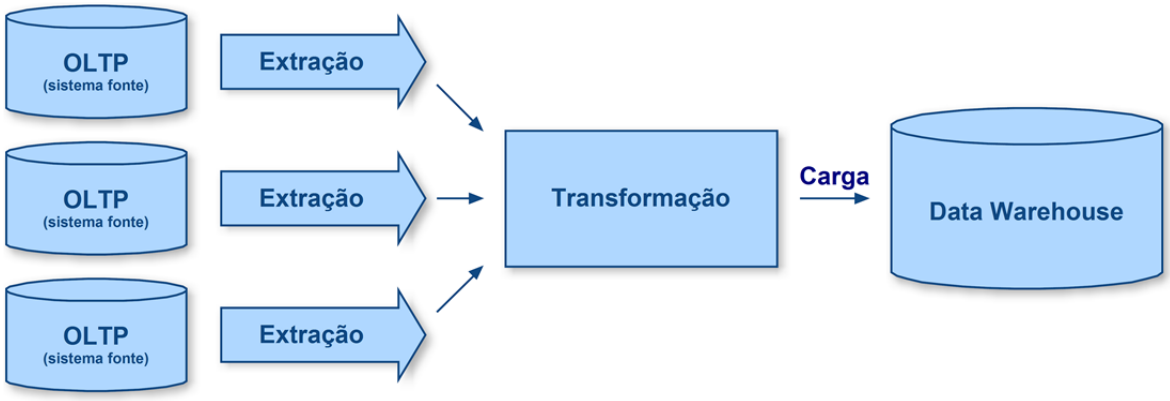


Figura 2.2. Extração, transformação e carga dos dados.

Extração

A extração é o primeiro passo na obtenção de dados para o ambiente de um armazém de dados. Significa basicamente ler e entender as fontes de dados e copiar as partes necessárias para a área de transformação de dados, a fim de serem trabalhadas posteriormente [Kimball & Ross, 2002]. Os programas de extração devem dar suporte à captura incremental dos dados que equivale a uma replicação baseada em dados modificados para posterior distribuição ao armazém de dados.

Transformação

Uma vez que os dados tenham sido extraídos dos sistemas-fonte, um conjunto de transformações deve ser processado sobre esses dados, convertendo-os em formato válido para o negócio e adequado para carga. A transformação dos dados pode envolver um ou vários processos, dependendo da necessidade e situação. Alguns dos processos mais comumente utilizados são:

Limpeza - constitui no conjunto de atividades realizadas sobre os dados extraídos, de modo a corrigir o uso incorreto ou inconsistente de códigos e caracteres especiais, resolver problemas de conflito de domínios, tratar dados perdidos, corrigir os valores duplicados ou errados. A finalidade é deixar os elementos de dados de acordo com formatos padrões (uniformizados), não duplicados, corretos, consistentes e que reflitam a realidade.

Combinação - é realizada quando fontes de dados possuem exatamente os mesmos valores de chaves representando registros iguais ou complementares.

Desnormalização - o padrão no processo de transformação é reunir as hierarquias de dados separadas em várias tabelas devido à normalização, dentro de uma única dimensão, de forma desnormalizada.

Cálculos, derivação e alocação - são transformações a serem aplicadas às regras de negócio identificadas durante o processo de levantamento de requisitos. É conveniente que as ferramentas a serem empregadas possuam um conjunto de funções, tais como manipulação de textos, aritmética de data e hora, entre outras.

Carga

Após os dados serem transformados, eles são carregados no armazém de dados. A carga dos dados também possui uma enorme complexidade, sendo que os seguintes fatores devem ser levados em conta:

Integridade dos dados - no momento da carga, é necessário verificar os campos que são chaves estrangeiras com suas respectivas tabelas para certificar-se de que os dados existentes na tabela da chave estrangeira estão de acordo com a tabela da chave primária;

Tipo de carga a ser realizada, incremental ou total - a carga incremental normalmente é feita para tabelas de fatos e a carga total é feita em tabelas de dimensão onde o analista tem que excluir os dados existentes e incluí-los novamente. Mas isso depende da necessidade do negócio em questão.

Otimização do processo de carga - todo banco de dados possui um conjunto de técnicas para otimizar o processo de carga, tais como evitar a geração de log durante o processo, criar índices e agregar dados. Muitas dessas características podem ser executadas nos bancos de dados ou registradas em *scripts* por meio da utilização de ferramentas sobre a área de organização de dados.

Suporte completo ao processo de carga - o serviço de carga também precisa suportar as exigências antes e depois da carga atual, como eliminar e recriar índices e particionamento físico de tabelas e índices.

2.2 Esquema de Dados para o Armazém de Dados

Um armazém de dados lida com basicamente dois tipos de dados: numéricos e descritivos. Os dados numéricos são chamados de fatos e os descritivos de dimensões. Os fatos guardam um histórico de valores relacionando-os com as dimensões participantes do fato. Eles são considerados dinâmicos, pois crescem ao longo do tempo, uma vez que novas medidas são adicionadas a cada nova carga do armazém de dados. As dimensões não costumam ter modificações significativas na constituição de seus membros e nos seus dados descritivos, tal como ocorre com os fatos.

O esquema dimensional também é chamado de esquema estrela (do inglês *star schema*), pois sua representação gráfica lembra o formato de uma estrela, com a tabela fato no centro e as dimensões nas pontas (Figura 2.3). O esquema estrela deu origem a diversas variantes, principalmente nos esquemas lógicos e físicos, adaptando-se às diferentes necessidades e aos produtos comerciais. Entre estes, os mais utilizados em bancos de dados relacionais são o esquema estrela tradicional, que possui os níveis de dimensões desnormalizados e o esquema floco de neve (do inglês *Snow Flake*), que possui os níveis de dimensões normalizados, formando uma hierarquia de níveis explícita e não redundante [Kimball et al., 2008].

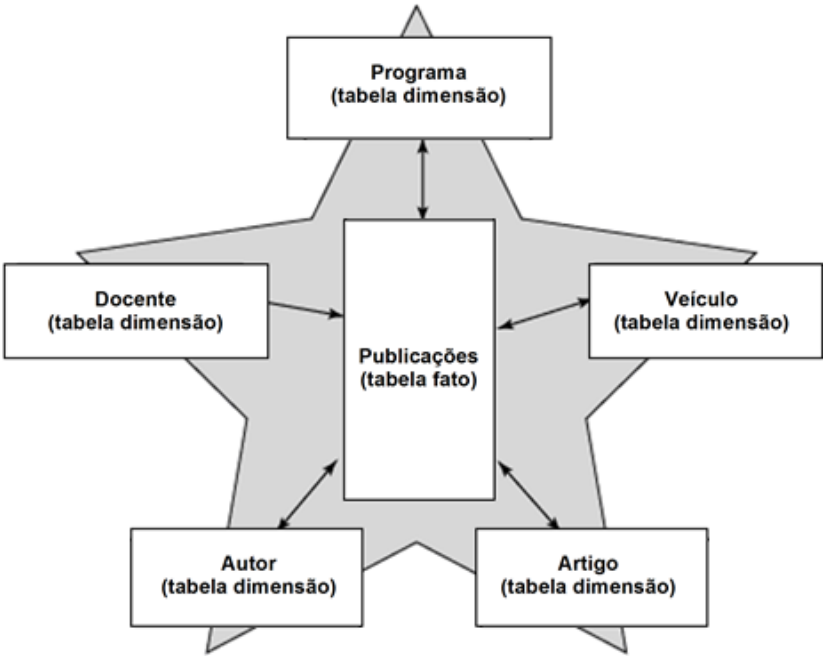


Figura 2.3. Esquema estrela.

2.2.1 Propriedades das Dimensões

As dimensões são ortogonais (independentes entre si) e são compostas por membros organizados em níveis hierárquicos para que seja possível executar as funções OLAP de *Drill-Down* e *Roll-Up*, as quais aumentam ou diminuem o nível de detalhamento dos dados de uma consulta [Barbieri, 2001]. Para a dimensão Tempo, por exemplo, podemos ter os seguintes membros: Datas (01/12/2007, 05/01/2008, 10/02/2008, 20/02/2008), Meses (11/2007, 01/2008, 02/2008) e Anos (2007, 2008), como é ilustrado da Figura 2.4. Estes membros são organizados em três níveis hierárquicos: data, mês e ano, sendo que os valores (métricas) associados a 10/02/2008 e 20/02/2008 irão compor a métrica

do mês 02/2008, que por sua vez irá compor a métrica do ano de 2008 juntamente com os valores dos meses 12/2007 e 01/2008. Na Figura 2.4, o membro de nome 12/2008, pertence ao nível Mês e possui como membro superior o membro denominado 2008.

A divisão hierárquica de uma dimensão é chamada de hierarquia de classificação e pode possuir diferentes caminhos hierárquicos [Abelló et al., 2001]. A maioria dos autores costuma representar a hierarquia de classificação por meio de um grafo acíclico dirigido, onde os nós representam os níveis e os arcos apontam os caminhos possíveis de sumarização [Trujillo et al., 2000].

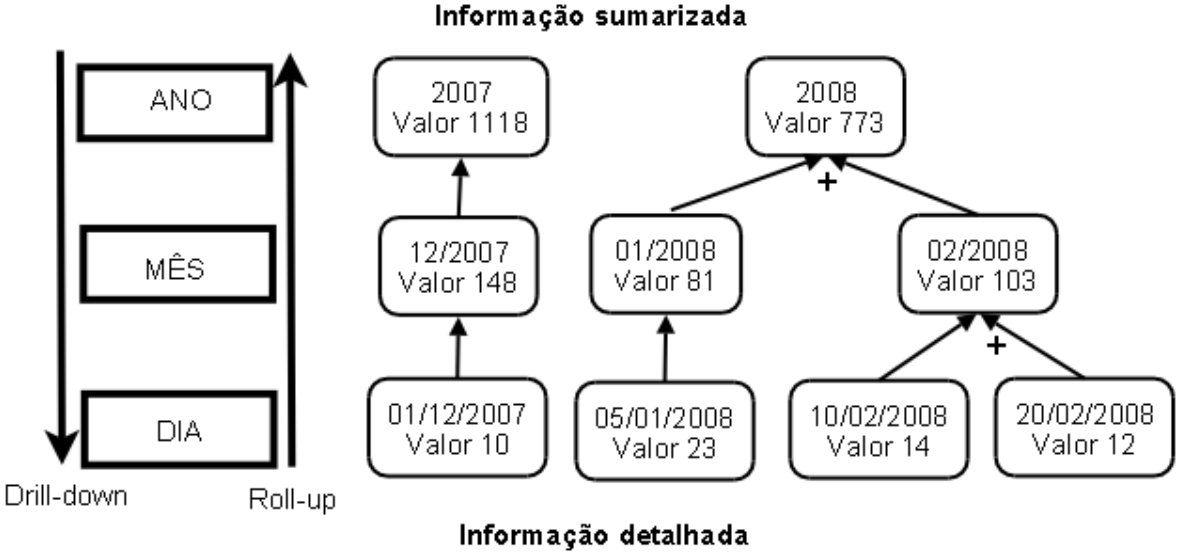


Figura 2.4. Exemplo de hierarquia para a dimensão tempo.

Cada membro de dimensão possui quatro informações indispensáveis:

- um identificador único utilizado nos relacionamentos internos do armazém de dados (geralmente um número binário ou hexadecimal);
- um nome para identificação pelo usuário;
- a indicação do nível hierárquico ao qual pertencem;
- uma lista de identificadores de seus membros superiores (ou ancestrais), isto é, os membros pais que utilizam ele (filho ou descendente) para sumarizar dados.

Os membros de dimensão podem possuir atributos descritivos para caracterizá-los melhor [Hüsemann et al., 2000]. Esses atributos podem servir de parâmetros em operações de seleção de membros ou como informação adicional nos resultados das consultas. Os atributos descritivos são definidos nos níveis de dimensão. Para o nível

Data, por exemplo, poderíamos ter os atributos: *Flag* de Feriado, *Flag* de Fim de Semana, Dia da Semana, etc.

2.2.2 Propriedades dos Fatos

Os fatos são constituídos por um conjunto de elementos formados por todas as combinações possíveis de membros de dimensões distintas. Cada um desses elementos possui uma lista de métricas associadas. Em outras palavras, um fato relaciona um conjunto de dimensões e fornece um conjunto de medidas (métricas) válidas para o ponto de encontro (interseção) dessas dimensões [Abelló et al., 2001]. Um fato não precisa ter um identificador único, pois ele pode ser representado pela lista de identificadores dos membros das dimensões com os quais se relaciona.

A Tabela 2.1 apresenta um exemplo da composição do fato publicação. Um artigo é publicado numa data num veículo por um docente que pertence a um programa de pós-graduação de uma determinada localização geográfica. Cada linha da tabela indica um elemento do fato publicação, que é composto por cinco membros de dimensões distintas (*Tempo*, *Localização geográfica*, *Programa*, *Docente*, *Veículo*) e duas métricas (*Artigos* e *Citações*).

Uma informação muito importante referente aos fatos é a maneira como eles devem ser agrupados e calculados (sumarizados) nos diversos níveis das dimensões. As métricas dos fatos podem utilizar funções diferentes para as totalizações nos níveis hierárquicos superiores das dimensões ou até não permitir totalizações. Na Tabela 2.1, por exemplo, podemos utilizar uma função de soma para compor os totais das métricas *Artigos* e *Citações*. Mas qual o valor da informação "número total de citações dos programas do Brasil"? Talvez seja mais interessante utilizar uma função de média geral ou média ponderada (em relação a número de artigos), para se ter uma ideia melhor do impacto médio por artigo (número de citações recebidas pelo artigo).

Tabela 2.1. Exemplo da composição do fato publicação.

Dimensões					Métricas	
Tempo	*Loc. Geog.*	*Programa*	*Docente*	*Veículo*	*Artigos*	*Citações*
12/2008	Brasil	Puc-Rio	Lucena	SIGMOD	3	30
12/2008	Brasil	Puc-Rio	Lucena	SIGIR	2	25
12/2008	Brasil	UFPE	João	WWW	1	442
12/2008	Am. Norte	Caltech	Jorge	SIGGraph	5	34
12/2008	Am. Norte	Caltech	Jorge	WSDM	4	13
12/2008	Am. Norte	Caltech	Antônio	SPIRE	4	14
12/2008	Am. Norte	MIT	Luís	TOIS	3	41

A Tabela 2.2 apresenta os mesmos fatos da Tabela 2.1 sumarizados em alguns níveis mais altos, com a métrica *Artigos* utilizando a função soma e a métrica *Citações* utilizando a função média geral. Todas as dimensões possuem o nível *Todos*, que caracteriza o nível mais alto de qualquer dimensão e que geralmente é suprimido das representações das árvores hierárquicas. O nível *Todos* significa desconsideração da dimensão em questão para a consulta realizada, pois não existem classificações distintas neste nível que é formado somente por um membro.

Tabela 2.2. Fato publicação sumarizado em níveis mais altos.

Dimensões					Métricas	
Tempo	*Loc. Geog.*	*Programa*	*Docente*	*Veículo*	*Artigos (soma)*	*Citações (média)*
12/2008	Brasil	TODOS	TODOS	TODOS	3534	4,59
12/2008	Brasil	Puc-Rio	TODOS	TODOS	645	7,97
12/2008	Brasil	UFPE	TODOS	TODOS	465	5,60
12/2008	Am. Norte	TODOS	TODOS	TODOS	15455	10,04
12/2008	Am. Norte	Caltech	TODOS	TODOS	564	8,09
12/2008	Am. Norte	Waterloo	TODOS	TODOS	1203	11,53
12/2008	Am. Norte	MIT	TODOS	TODOS	609	6,12

2.2.3 Cubos de Dados

O cubo de dados representa um conjunto de métricas que compartilham o mesmo conjunto de dimensões [Kimball et al., 2008]. O cubo é formado por um conjunto ordenado de células, onde cada célula é localizada pela intersecção de suas três dimensões (altura, largura e profundidade). A célula armazena os valores das métricas para a respectiva localização da célula. Obviamente, a maioria dos fatos de um armazém de dados possui mais de três dimensões, não podendo ser corretamente representados na imagem de um cubo (elemento 3D), de forma que alguns autores utilizam a expressão "hipercubo". Neste trabalho é utilizada a expressão "cubo".

As ferramentas utilizadas para as consultas OLAP utilizam-se desta representação tridimensional dos dados para definir suas principais operações de consulta. Estas operações são baseadas na manipulação e visualização de um cubo físico (real) por uma pessoa, por exemplo:

- rotações de 90° do objeto (mudança da posição dos eixos x, y, z);
- visualização plana de um dos lados (esconder uma das dimensões);
- corte de pedaços (seleção de partes do cubo);

- focalização de detalhes por meio da aproximação do objeto (ver dados com mais detalhes) .

A maioria dos operadores das consultas OLAP tem como base estas operações simples e intuitivas, só que adaptadas para a manipulação de dados multidimensionais (n dimensões) de forma que o usuário não precise de muito conhecimento para formalizar suas consultas.

2.3 Consultas OLAP

As consultas analíticas feitas em um armazém de dados têm como objetivo verificar o comportamento de métricas do negócio (dados numéricos, medidas) ao longo do tempo. Essas consultas são normalmente chamadas de consultas OLAP (*Online Analytical Processing*), tendo como principais características o cálculo de valores contidos em uma enorme quantidade de registros e a apresentação dos resultados em formatos de nível mais alto, como tabelas ou gráficos.

A característica principal dos sistemas OLAP é permitir uma visão conceitual multidimensional dos dados armazenados. A visão multidimensional é mais útil para os analistas do que a visão tabular tradicional utilizada nos sistemas de processamento de transação. Ela é mais natural, fácil e intuitiva, permitindo uma visão dos negócios da organização em diferentes perspectivas e, dessa maneira, torna o analista um explorador de informações [Chaudhuri & Dayal, 1997; Shoshani, 1997; Campos & Rocha Filho, 1997].

As ferramentas OLAP são projetadas para apoiar análises e consultas *ad hoc* em um armazém de dados, além de ajudar analistas e executivos a sintetizar informações sobre a organização, por meio de comparações, visões personalizadas, análise histórica e projeção de dados em vários cenários. Ferramentas OLAP são implementadas para ambientes multiusuário, arquitetura cliente-servidor, e oferecem respostas rápidas e consistentes às consultas interativas executadas pelos analistas, independentemente do tamanho e complexidade de um armazém de dados [Codd et al., 1993; Chaudhuri & Dayal, 1997; Inmon, 1996].

A fim de permitir uma visualização e manipulação multidimensional dos dados, as ferramentas OLAP oferecem diferentes funções, a saber:

- *Pivot*: muda a orientação dimensional de uma pesquisa. Por exemplo, *pivot* pode consistir na troca de linhas e colunas, ou mover uma das dimensões da linha, para a dimensão da coluna;

- *Roll-up*: os bancos de dados multidimensionais geralmente têm hierarquias ou relações de dados baseadas em fórmula dentro de cada dimensão. Então, a execução do *roll-up* computa todas essas relações para uma ou mais dimensões;
- *Slice*: um *slice* é um subconjunto da estrutura multidimensional que corresponde a um valor simples em lugar de um ou mais atributos das dimensões. É como fixar um valor de uma das dimensões de um cubo e considerar para pesquisa o subcubo formado por esse valor e pelas outras dimensões do cubo inicial;
- *Drill-down/up*: consiste em fazer uma exploração em diferentes níveis de detalhe das informações, como por exemplo, analisar uma informação por continente, país ou estado, partindo da mesma base de dados;
- *Drill-across*: é o processo de unir duas ou mais tabelas-fato de mesmo nível de detalhes, ou seja, tabelas com o mesmo conjunto de colunas e restrições dimensionais;

Essas funções podem ser utilizadas à vontade pelos usuários de um ambiente de armazém de dados, conforme as suas necessidades de informações.

Capítulo 3

Dados Utilizados

Neste capítulo é apresentada a construção da base de dados de publicações científicas. A construção da base de dados foi realizada em três etapas apoiadas em entidades básicas da base: docentes dos programas, publicações e citações.

Este capítulo está estruturado da seguinte forma: primeiro, na Seção 3.1 apresentamos o Projeto Perfil-CC; na Seção 3.2, como foram coletados os docentes dos programas analisados; na Seção 3.3 é descrita a recuperação das publicações; na Seção 3.4 é abordado o processo de obtenção das citações das publicações e, por fim, na Seção 3.5 são apresentadas as considerações finais para este capítulo.

3.1 Dados do Projeto Perfil-CC

O Projeto Perfil-CC nasceu com o objetivo de realizar uma avaliação qualitativa da produção gerada pelos principais programas da área no Brasil, bem como de sua inserção internacional, de modo que se possa estabelecer o papel que a área vem desempenhando no desenvolvimento científico e tecnológico do país. O projeto foi criado na UFMG em 2006 para estudar o perfil da produção científica dos principais programas brasileiros de modo a compará-los com seus congêneres da América do Norte e da Europa.

Neste contexto, diversos trabalhos foram feitos, como um estudo sobre a produção em Ciência da Computação no Brasil [Laender et al., 2008], uma pesquisa sobre redes sociais em computação [Menezes et al., 2008, 2009] e um trabalho sobre construção de *rankings* para conferências [Martins et al., 2009]. Os resultados obtidos nesta dissertação complementam o estudo realizado no contexto do projeto Perfil-CC com uma análise do impacto das citações dos artigos publicados pelos programas e o estudo da distribuição das publicações por subáreas da Ciência da Computação. Além disto,

este trabalho disponibiliza um ambiente de análise multidimensional para uma melhor análise dos dados do Projeto Perfil-CC.

A Figura 3.1 apresenta o processo completo da construção da base de dados. Inicialmente são coletados os docentes dos programas analisados. Em seguida são recuperadas as publicações de cada docente e as citações recebidas por cada publicação. Por fim, os dados são inseridos na base de dados. Cada uma das etapas é detalhada nas seções a seguir.

3.2 Docentes

Esta é a primeira etapa da construção do banco de dados utilizado em nossos experimentos. Inicialmente selecionamos os oito principais programas, segundo a classificação CAPES vigente no período de coleta de dados (junho de 2007). Os vinte e dois programas estrangeiros foram selecionados entre os mais importantes da América do Norte e da Europa. Os programas selecionados são:

- *Brasil*: PUC-Rio, UFRJ/COPPE, UFMG, UFPE, UFRGS, UNICAMP, USP / São Paulo e USP / São Carlos;
- *América do Norte*: Brown University, Caltech - California Technology Institute, CMU - Carnegie Mellow University, Cornell University, Harvard University, MIT - Massachusetts Institute of Technology, Princeton University, Stanford University, University of British Columbia, University of California at Berkeley, University of Illinois, University of Texas at Austin, University of Toronto, University of Washington, University of Waterloo e University of Wisconsin;
- *Europa*: Cambridge University, École Polytechnique de Paris, ETH Zürich, Imperial College, Oxford University e Université Pierre et Marie Curie - Paris VI.

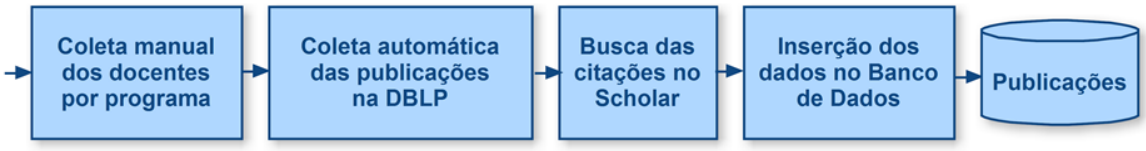


Figura 3.1. Processo de construção da base de dados de publicações.

A lista dos docentes foi obtida a partir da página Web dos respectivos programas. Esta coleta foi feita de forma manual, pois não havia padrão nas estruturas das páginas, impossibilitando a criação de coletores automáticos por expressão regular. A Figura

3.2 mostra um exemplo de página com a listagem dos docentes dos programas de pós-graduação.

Além do nome dos docentes, coletou-se a titulação (professor titular, adjunto, *full professor*, etc) e a URL de sua página Web, quando existente. A coleta dos 30 programas resultou numa lista com 2.027 docentes. Esta lista foi utilizada como entrada para a obtenção das publicações, processo detalhado na próxima seção.

Figura 3.2. Página com listagem dos docentes do programa.

3.3 Publicações

As publicações de cada docente foram obtidas a partir da *DBLP Computer Science Bibliography*[1] [Ley, 2002; Ley & Reuther, 2006]. A DBLP é a principal biblioteca digital da área de Ciência da Computação, mantida por um grupo da Universidade de Trier, Alemanha, coordenado por Michael Ley. Em junho de 2007, período em que foi feita a coleta de dados, a DBLP registrava mais de 910.000 artigos publicados nos anais de 3.636 diferentes conferências e em 613 diferentes periódicos, cobrindo os principais campos da área, com maior ênfase naqueles de cunho tecnológico, tais como

[1] http://dblp.uni-trier.de

Bancos de Dados, Engenharia de Software, Linguagens de Programação, Sistemas de Computação e Redes de Computadores. A DBLP tem uma representação maior do que repositórios bibliográficos de uso geral, como o mantido pelo ISI - *Institute for Scientific Information*[2].

Outros repositórios de publicações estudados foram o Citeseer[3] e o Google Scholar[4]. Tais repositórios possuem mais artigos catalogados do que a DBLP. Porém a coleta de dados é feita de forma automática, ficando suscetível a diversos erros como artigos duplicados, erros nos nomes dos autores e dos veículos de publicações. Já a DBLP possui informações bem estruturadas, organizadas de forma parcialmente manual por seus administradores Ley & Reuther [2006]. Seus dados possuem alta qualidade: baixa taxa de homônimos entre os autores, identificação e classificação exata dos diversos veículos de publicação.

Na DBLP foram encontrados as publicações de 1.760, ou 87%, dos 2.027 docentes da lista inicial. Conforme apresentado na Tabela 3.1, foram registrados no banco de dados 52.596 artigos de autoria de 1.760 docentes dos 30 programas analisados. Esses artigos foram publicados entre 1954 e 2007 em 456 periódicos e nos anais de 1.622 conferências distintas. Desses veículos, 241 periódicos e 605 conferências aparecem classificados no Qualis de Ciência da Computação[5].

Tabela 3.1. Corpus de publicações gerado.

# Artigos	52.596
# Docentes na DBLP	1.760
# Programas	30
# Periódicos	456
# Conferências	1.622

3.4 Citações

A partir das publicações levantadas foi coletado o número de citações de cada artigo por meio do repositório de publicações Google Scholar. Para isso, submetemos consultas com o título completo do artigo e usamos uma opção avançada que retorna apenas os artigos que tenham em seu título todas as palavras consultadas. A Figura 3.3 mostra a consulta para recuperar o número de citações do artigo "Characterizing a spam traffic"

[2]http://scientific.thomson.com
[3]http://citeseer.ist.psu.edu
[4]http://scholar.google.com
[5]avaliação da qualidade do veículo fornecida pela CAPES.

no Google Scholar. Esse artigo possui 62 citações. As consultas foram automatizadas através de *scripts* escritos na linguagem Perl. Entretanto, o Google Scholar possui limitação quanto ao número de consultas submetidas, sob pena de bloquear o acesso ao repositório através do número IP da máquina que originou a consulta. Para evitar tal problema, foi inserido um intervalo aleatório de 45 ± 10 segundos entre as consultas, valor obtido empiricamente com sucesso no trabalho Silva et al. [2006]. Este intervalo possibilita consultar apenas 1.920 páginas por dia. Para aumentar esta taxa, foram utilizadas oito máquinas com endereços IPs distintos do laboratório onde este trabalho foi desenvolvido.

Eventualmente o Google Scholar retorna artigos distintos com títulos parecidos com os consultados distorcendo o resultado. Por exemplo, ao buscar o artigo "Concept-based interactive query expansion", o Google Scholar retorna os artigos "A study of user interaction with a concept-based interactive query expansion support tool" e o próprio "Concept-based interactive query expansion". Para evitar esse tipo de problema, realizamos um casamento aproximado de caracteres usando a métrica de distância de edição[6], descartando os artigos com título diferente do título original. Foram considerados os artigos cuja proximidade entre os títulos fosse maior que 75%, valor que também foi obtido no trabalho Silva et al. [2006]. Adicionalmente, para obter uma melhor qualidade dos dados, também fizemos uma inspeção manual nos artigos com mais de 40 citações, retirando os artigos associados incorretamente.

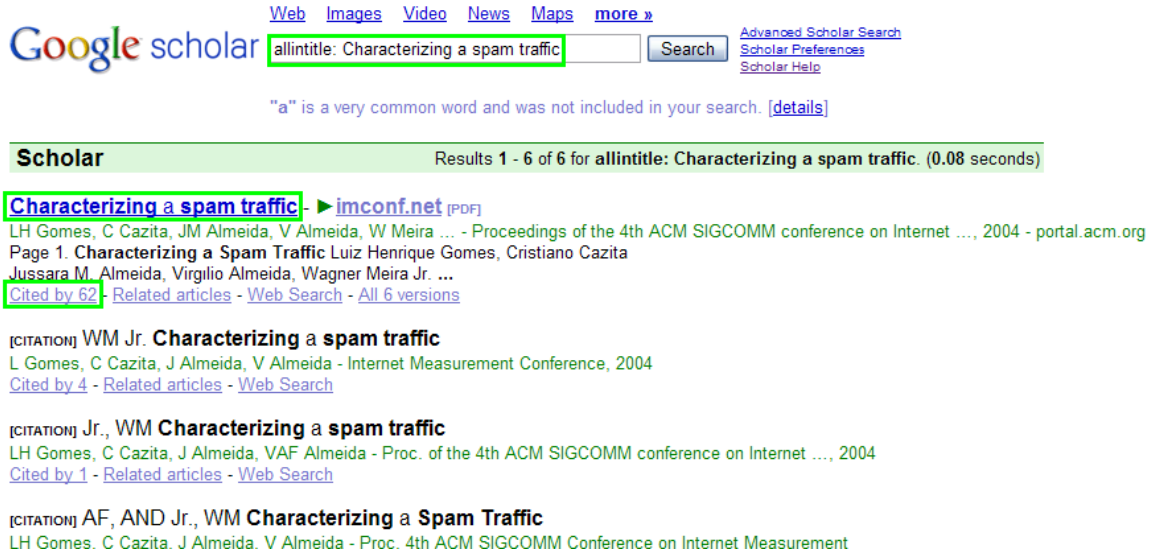


Figura 3.3. Interface de consulta do Google Scholar.

[6]Número mínimo de operações necessárias para transformar uma cadeia de caractere na outra. Por exemplo, a distância de edição entre *spam* e *slam* é de uma substituição de *p* por *l*, ou seja, 3/4 ou 75% de proximidade entre as cadeias.

Foram coletadas as citações de todos os artigos publicados entre 1998 e 2006 dos 30 programas analisados. Foram considerados apenas os artigos cuja classificação Qualis seja A Internacional, ou seja, publicados em veículos de circulação internacional e que têm maior impacto científico. No total foram coletados 15.795 artigos por oito máquinas em aproximadamente 24 horas. Os resultados da coleta foram salvos em arquivos-texto e posteriormente os dados foram inseridos no banco de dados bibliométrico. A Tabela 3.2 mostra taxa de artigos encontrados por esse método. Esse método apresentou uma boa cobertura com taxa de retorno de 97,46% dos artigos pesquisados, possibilitando a análise numa base representativa.

Tabela 3.2. Taxa de artigos encontrados no Google Scholar.

	Artigos	*%*
Encontrados	15.398	97,46%
Não encontrados	397	2,54%
Total	15.795	100,00%

3.5 Considerações Finais

Com os passos apresentados neste capítulo podemos construir o banco de dados utilizado no armazém de dados e consequentemente em nossas análises. A partir do banco de dados gerado, podem ser obtidas diversas estatísticas sobre os 30 programas abordados, tais como: média de publicações por docente por programa, distribuição das publicações entre as subáreas da Ciência da Computação, subáreas mais populares por programa, média de citações por artigo por programa, indicadores de produtividade por docente, por programa ou por país, entre outras. Para melhor flexibilidade, agilidade e eficiência na análise faz-se necessária a utilização de ferramentas adequadas, as quais serão abordadas no capítulo seguinte.

Capítulo 4

Construção do Armazém de Dados

Este capítulo trata da construção do armazém de dados de publicações científicas. A construção desse armazém de dados consistiu de três etapas: a modelagem do problema, a escolha da ferramenta a ser utilizada e a construção propriamente dita do armazém de dados. A Seção 4.1 apresenta o problema e a modelagem da sua solução. A Seção 4.2 apresenta as ferramentas utilizadas como apoio. A Seção 4.3 descreve a construção do armazém de dados. A Seção 4.4 apresenta as considerações finais para este capítulo.

4.1 Modelagem do Problema

Esta seção discute a modelagem do banco de dados de publicações científicas. O primeiro passo é inserir os dados num banco de dados relacional. Para fins didáticos, apresentaremos o esquema de forma simplificada e, em seguida, apresentaremos o esquema completo com todas as entidades do problema.

4.1.1 Esquema Simplificado

A Figura 4.1 mostra o esquema relacional com as tabelas e os relacionamentos. O esquema possui quatro tabelas: *author*, para designar os autores dos artigos; *program*, que representa o programa de pós-graduação ao qual o autor esteja filiado; *article*, que representa o artigo que o autor publicou; e *venue*, para representar o veículo de publicação ao qual o artigo pertence. A relacionamento de *author* para *article* é do tipo muitos para muitos (representada pela tabela *author_has_article*): um autor possui um ou mais artigos, e um artigo possui um ou mais autores. Os demais relacionamentos são do tipo um para muitos.

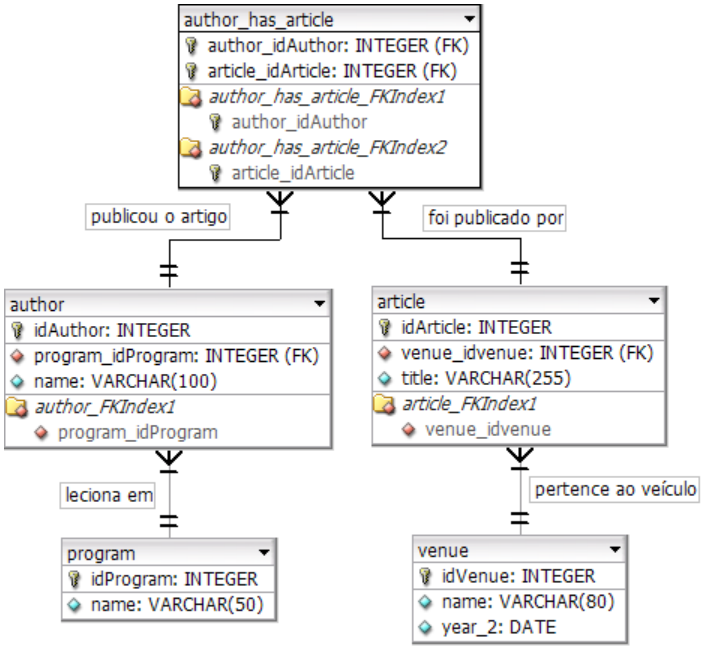


Figura 4.1. Esquema relacional simplificado para representar o problema.

Conforme vimos no Capítulo 2, o esquema dimensional possui uma estrutura mais intuitiva quando comparado ao relacional, permitindo uma consulta mais fácil para usuários não especializados [Kimball & Ross, 2002]. Nesse esquema, as tabelas representam Fatos e Dimensões. Os dados são desnormalizados para um melhor desempenho na leitura, visto que as junções entre as entidades só ocorrem entre as tabelas Fato e Dimensões.

A conversão entre o esquema relacional e o dimensional consiste, basicamente, em ligar todas as tabelas do primeiro esquema a uma tabela central: a tabela fato. O custo para o ganho no desempenho da leitura é o espaço adicional utilizado por essa tabela, já que a desnormalização leva a uma redundância dos dados. A Figura 4.2 ilustra a conexão entre a tabela fato (*fact table*) e as tabelas dimensão.

A tabela fato é composta por quatro chaves: programa, autor, artigo e veículo. Cada registro representa uma publicação de um artigo: a publicação de um artigo possui uma dimensão *article*, uma dimensão *author*, uma dimensão *program* e uma dimensão *venue*. Assim, cada linha da tabela fato representa o fato de um autor, que é vinculado a um programa, publicar um determinado artigo num determinado veículo. Podemos dizer que o grão mínimo do nosso problema é o artigo, ou seja, não há uma divisão ou quebra a partir de artigo. Dessa forma, o tamanho do banco de dados (número de linhas da tabela fato) é dado pelo número de artigos.

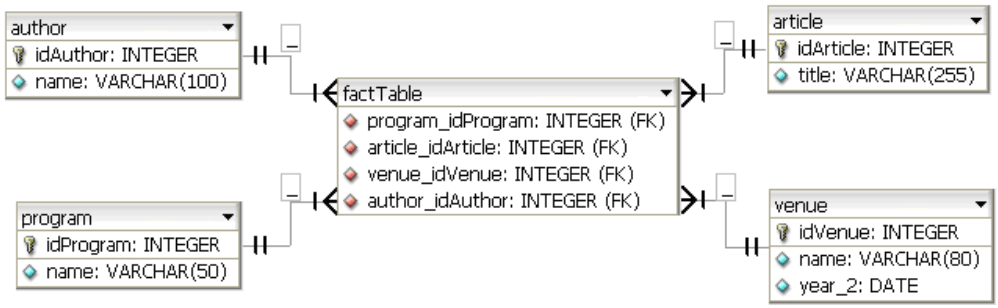


Figura 4.2. Esquema dimensional simplificado criado a partir do esquema relacional.

4.1.2 Esquema Completo

Na seção anterior, apresentamos o esquema proposto de forma simplificada. O esquema completo possui mais alguns dados, que são listados a seguir:

- *professor* (docentes) - precisamos separar as entidades autor e docente, pois um docente pode ou não ser um autor de um artigo, e um autor de um artigo pode ou não ser um docente;

- *name* (nome do docente) - um docente pode assinar um artigo com nomes distintos;

- *venue_ instance* (instância do veículo) - um veículo pode ter várias edições ocorrendo por exemplo, anualmente;

- *Qualis* (Qualis do veículo) - avaliação da qualidade do veículo fornecida pela CAPES.

- *citation* (citações) - informações relativas às citações dos artigos.

A Figura 4.3 apresenta o esquema relacional completo do problema. Esse esquema permite realizar uma série de análises interessantes em seus dados. Porém, para uma análise mais eficiente é necessário converter esse esquema para o formato dimensional. Novamente, para construir um esquema dimensional, ligamos todas as tabelas à tabela fato, conforme ilustrado na Figura 4.4. Além disto foi adicionada uma dimensão exclusiva para tempo (*time*) que representa o ano da publicação do artigo. Obtivemos assim um esquema com dez dimensões e uma tabela fato. Finalizada a modelagem do problema, partimos para o estudo das ferramentas disponíveis, a seguir.

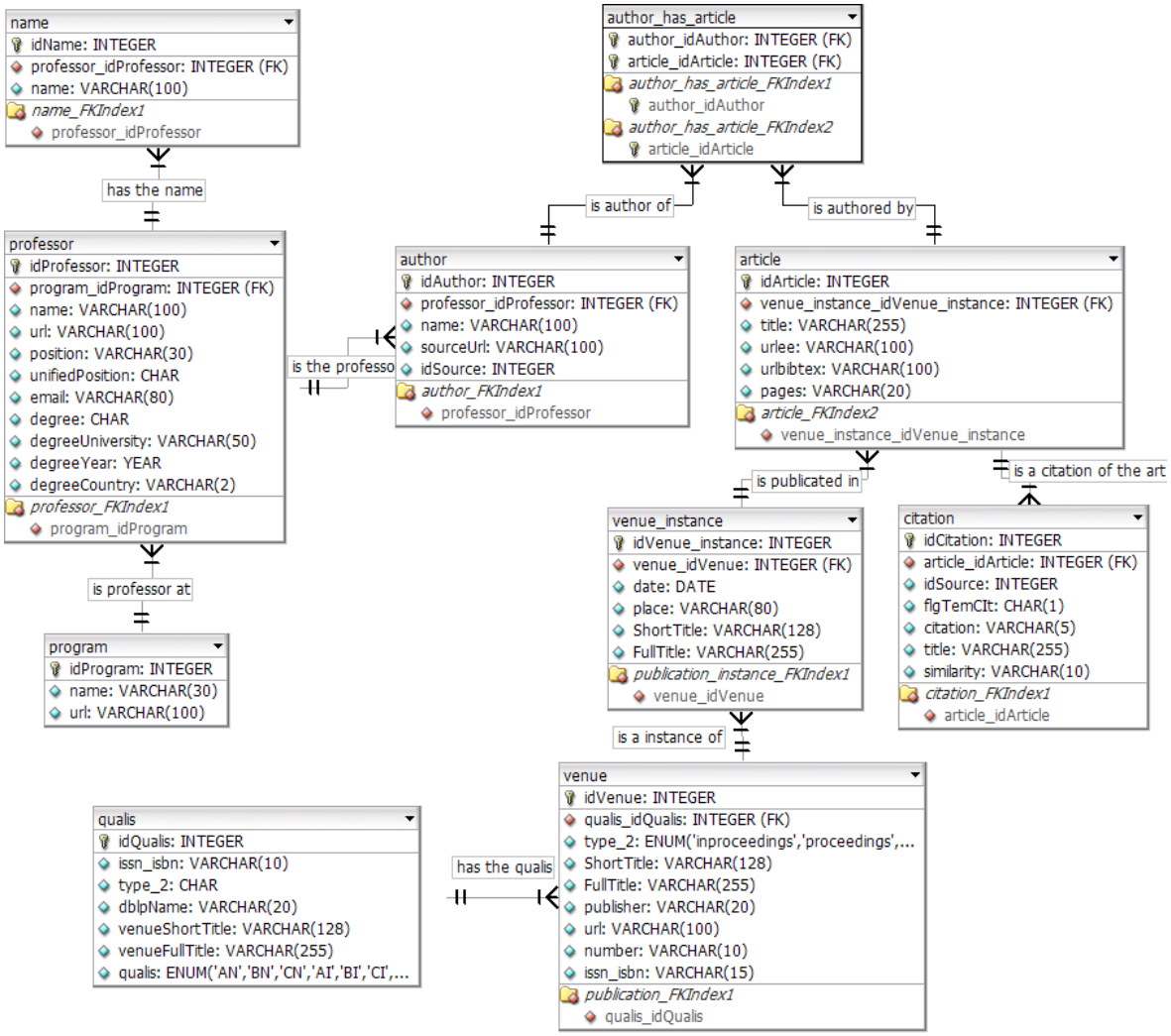


Figura 4.3. Esquema relacional completo que representa o problema.

4.2 Estudo de Ferramentas

Nesta seção apresentamos o processo de escolha da ferramenta utilizada para suportar o armazém de dados. Após consultas com especialistas no assunto, pré-selecionamos três ferramentas: FreeOLAP[1], Oracle Discoverer[2] e Pentaho[3].

A primeira ferramenta analisada foi o FreeOLAP, uma ferramenta OLAP com interface Web escrita em Java. O FreeOLAP executa em qualquer servidor de aplicação Web como, por exemplo, o Tomcat, e conecta a qualquer sistema gerenciador de bancos de dados relacional como o MySQL por meio do conector JDBC. Na época da análise,

[1] http://freeolap.com
[2] http://www.oracle.com/technology/products/discoverer/
[3] http://www.pentaho.com

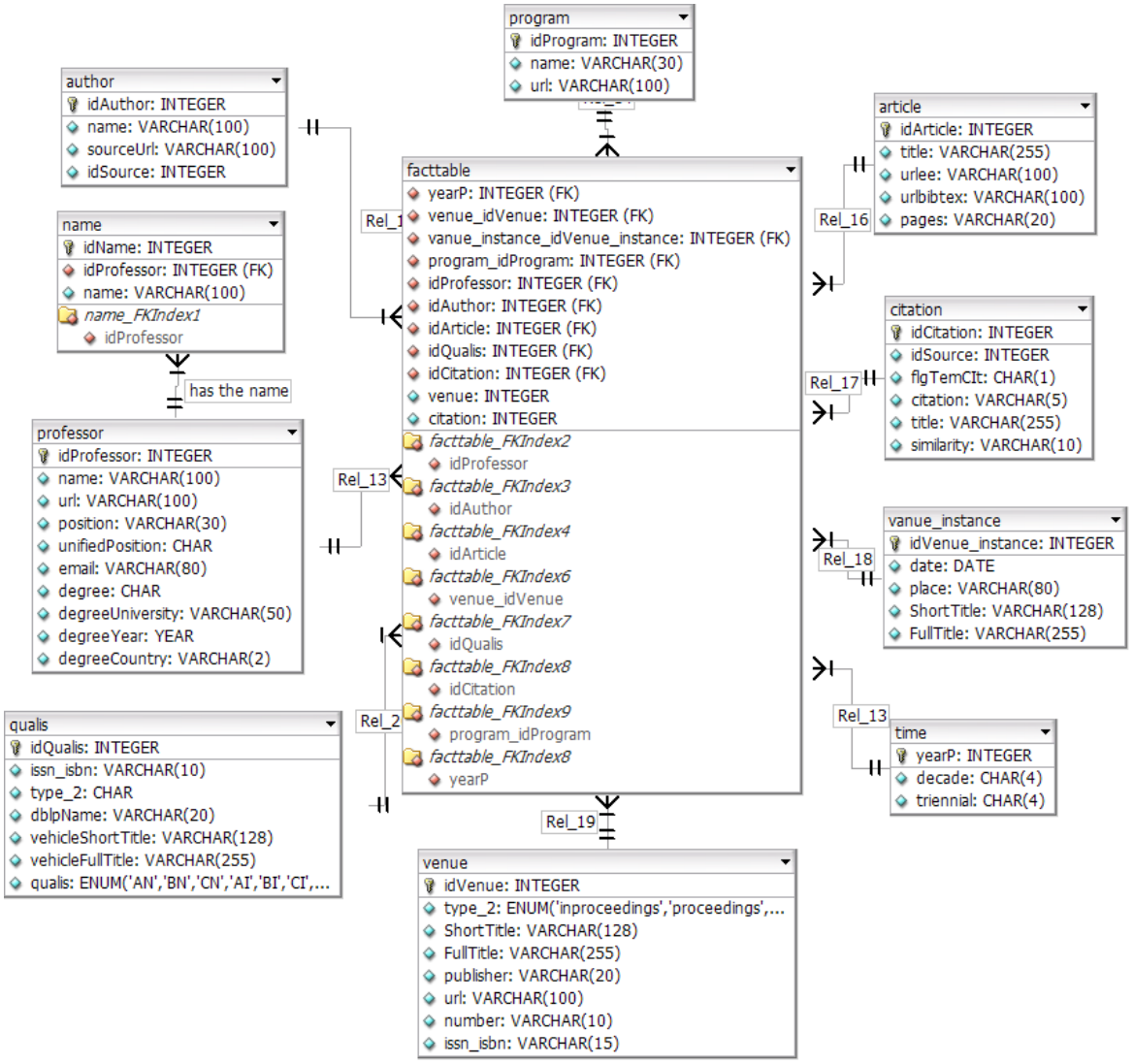


Figura 4.4. Esquema dimensional completo do problema.

esta ferramenta não estava com toda sua funcionalidade completa e seu projeto tinha sido descontinuado. Por falta de documentação técnica não foi possível fazer uma análise mais profunda da viabilidade da implementação do ambiente proposto usando esta ferramenta.

A segunda ferramenta avaliada foi o Oracle Business Intelligence Discoverer ou simplesmente Discoverer, que compreende na verdade um conjunto de ferramentas para consulta, geração de relatórios, análise de dados e publicação Web *ad hoc*. Diferentemente da primeira, esta ferramenta possui uma documentação técnica completa, o que facilitou a sua avaliação. Fizemos sua instalação, implementamos uma versão simplificada do armazém de dados de publicações e fizemos os testes iniciais. Esta ferramenta teve um bom desempenho em nossa análise.

Por último analisamos o pacote de Business Intelligence Pentaho. O projeto Pentaho é uma aplicação de código aberto de Business Intelligence composta de ferramentas de relatórios gerenciais, análise de dados, mineração de dados, *workflow* e tratamento de dados (ETL). A comunidade responsável pelo desenvolvimento da aplicação é composta por mais de 8.000 membros. Seu modelo de negócio consiste em gerar receita por meio de serviços de suporte técnico e gerenciamento para consumidores corporativos. Instalamos o produto, implementamos uma primeira versão do armazém de dados e testamos seu desempenho. Consideramos uma boa opção de ferramenta.

Por ser uma ferramenta de código aberto e consequentemente se adequar melhor aos recursos do laboratório onde o trabalho foi desenvolvido e pelos bons resultados nos testes de executados, decidimos implementar o armazém de dados de publicações científicas com a ferramenta Pentaho. Para contruir o armazém de dados, utilizamos uma série de outras ferramentas, listadas a seguir:

Pentaho Data Integration - O Pentaho Data Integration (anteriormente denominado Kettle) é uma ferramenta ETL (Extração, Transformação e Carga dos dados). Ele permite a extração dos dados do sistema-fonte, a execução de transformações nos dados como limpeza, unificação, cálculos e a carga dos dados para o banco de dados do armazém de dados.

Pentaho Cube designer - O Cube Designer é uma ferramenta de auxílio para a construção da definição dos cubos.

Pentaho Mondrian - O Mondrian é o servidor OLAP. Ele recebe solicitações de um cliente e as submete a um banco de dados tradicional, mapeando as consultas multidimensionais para um banco de dados relacional.

Pentaho Reporting - O Reporting (anteriormente conhecido como JPivot) provê a interface do sistema. É usado na visualização das consultas retornadas, contendo operações básicas de manipulações de cubos multidimensionais e de apresentação dos resultados em tabelas e gráficos.

Tomcat[4] - O Apache Tomcat é uma servidor Web para aplicações Java e de código aberto. O Tomcat é necessário para executar as aplicações Mondrian e Reporting. Utiliza a linguagem SQL (Structured Query Language - Linguagem de Consulta Estruturada) como interface.

MySQL[5] - O MySQL é um sistema de gerenciamento de bancos de dados (SGBD) de código aberto. É utilizado para armazenar os dados do armazém de dados e prover os dados requisitados pelo Mondrian nas consultas executadas no Reporting.

[4] http://tomcat.apache.org
[5] http://www.mysql.com

DBDesigner[6] - O DBDesigner é um editor visual para criação, modelagem e manutenção de bancos de dados. Também é do tipo código aberto.

Eclipse[7] - O Eclipse é uma IDE (ambiente integrado para desenvolvimento de *software*) de código aberto para a construção de códigos. Foi utilizada para construir as consultas em JSP e ajustar os cubos em XML.

4.3 Implementação do Armazém de Dados

Dividimos a implementação do armazém de dados em três etapas: construção da tabela fato, construção dos cubos e construção das consultas. Estes passos serão detalhados a seguir.

4.3.1 Construção da Tabela Fato

A primeira parte da nossa implementação foi a construção da tabela fato a partir do banco de dados de publicações já existente. As dimensões do novo esquema são as tabelas *program*, *professor*, *author*, *article*, *publication_instance*, *publication* e *qualis*, conforme ilustrado anteriormente na Figura 4.4. A tabela fato é composta das chaves estrangeiras para as dimensões e um indicador *publication* que tem sempre valor 1 para cada publicação do banco de dados. Este indicador é utilizado para calcular o número de publicações por professor, por programa, por veículo, entre outras visões.

A Figura 4.5 mostra a tela da ferramenta Pentaho Data Integration[8] com o *script* que gera os dados da tabela fato. A primeira parte do *script* cria a tabela fato no banco de dados por meio de comandos SQL. Em seguida extraímos os dados do banco de dados de publicações por meio de várias junções entre as tabelas conforme ilustrado na Figura 4.6. Finalmente, no passo seguinte, carregamos os dados na tabela fato. Com a tabela fato finalizada, foi possível trabalhar na construção dos cubos, o que será detalhado na próxima seção.

4.3.2 Construção dos Cubos

Os cubos foram construídos na linguagem MDX (*Multidimensional Expressions*)[9] utilizando as ferramentas Cube Designer e Eclipse, ilustradas nas Figuras 4.7 e 4.8,

[6] http://fabforce.net/dbdesigner4/

[7] http://www.eclipse.org

[8] ferramenta ETL (Extração, Transformação e Carga) da plataforma Pentaho.

[9] linguagem de consulta para bancos de dados multidimensionais, assim como SQL é uma linguagem de consulta para bancos relacionais.

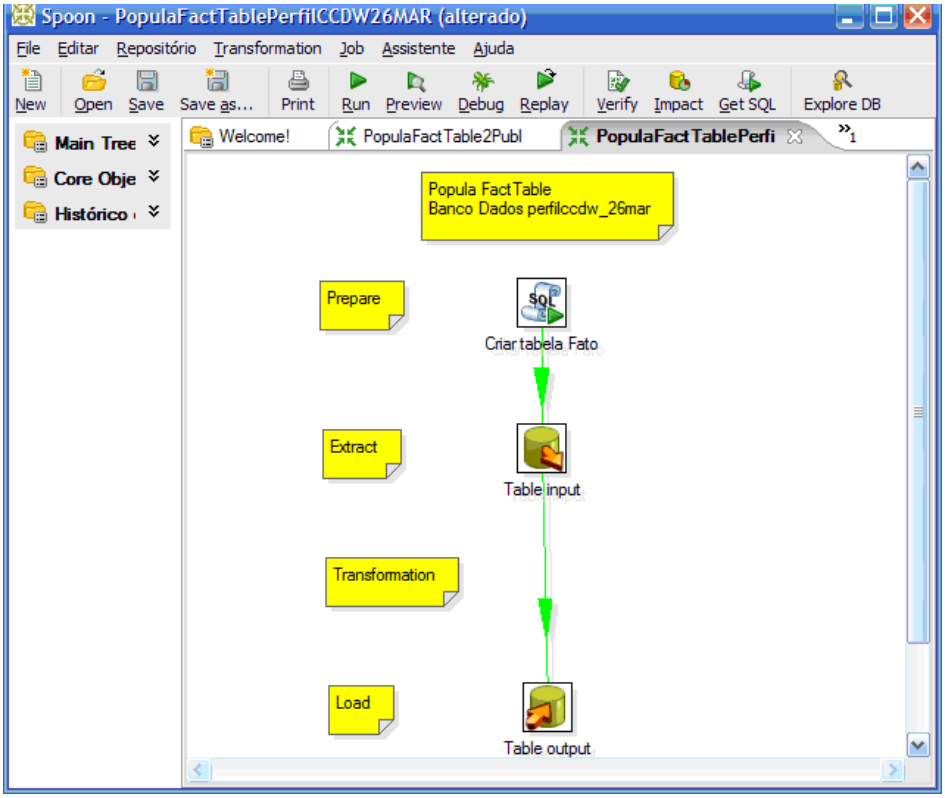


Figura 4.5. Criação do *script* no Pentaho Data Integration para gerar os dados da tabela fato.

repectivamente. Para uma maior flexibilidade das análises decidimos inserir no cubo todas as dimensões disponíveis no esquema dimensional como, por exemplo, *docente* e *publicação*. O Cube Designer gera como formato de saída um arquivo MDX com as definições do cubo. Este arquivo é ligado às configurações das consultas que serão descritas na próxima seção.

4.3.3 Geração das Consultas

Geramos as consultas com auxílio do editor Eclipse (Figura 4.9). Embora a ferramenta ofereça flexibilidade quanto à geração de consultas, oferecemos aos usuários do ambiente consultas pré-definidas que podem ser utilizadas como ponto de partida. A Figura 4.9 mostra as consultas disponíveis no ambiente gerado, acessado pelo endereço http://www.latin.dcc.ufmg.br:8080/perfilccDW/. Por exemplo, a consulta "4 - Artigos por Programa" lista as publicações dos programas sem restrição de datas. Ela é facilmente convertida para a consulta "5 - Artigos por Programa (Filtro: 2006)" através do uso de filtros, que considera apenas as publicações do ano de 2006.

A Figura 4.10 mostra a consulta "5 - Artigos por Programa (Filtro: 2006)". Po-

```
SELECT pr.faculty_idFaculty AS idFaculty,
       pr.idProfessor,
       au.idAuthor,
       ar.idArticle,
       pu.idPublication,
       pi.idPublication_instance,
       pu.qualis_idQualis AS idQualis,
       c.idCitation,
       pu.area              AS idSubarea,
       YEAR(pi.date)        AS yearP,
       1                    AS publication,
       1                    AS citation
FROM   professor            AS pr,
       author               AS au,
       author_has_article   AS aha,
       article              AS ar,
       publication_instance AS pi,
       publication          AS pu,
       citation             AS c
WHERE  au.professor_idProfessor                        = pr.idProfessor
   AND au.idAuthor                                     = aha.author_idAuthor
   AND aha.article_idArticle                           = ar.idArticle
   AND ar.publication_instance_idPublication_instance  = pi.idPublication_instance
   AND pi.publication_idPublication                    = pu.idPublication
   AND c.article_idArticle                             = ar.idArticle;
```

Figura 4.6. Código utilizado para gerar os dados da tabela fato.

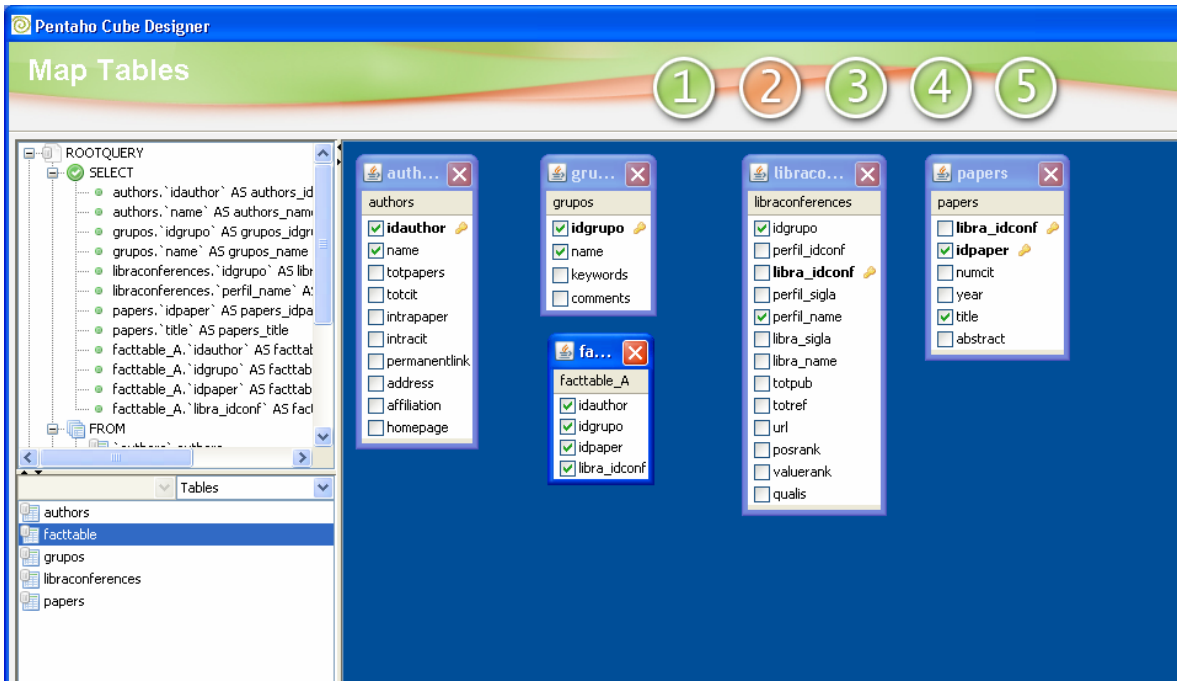


Figura 4.7. Tela do programa utilizado para modelar os cubos (Cube Designer).

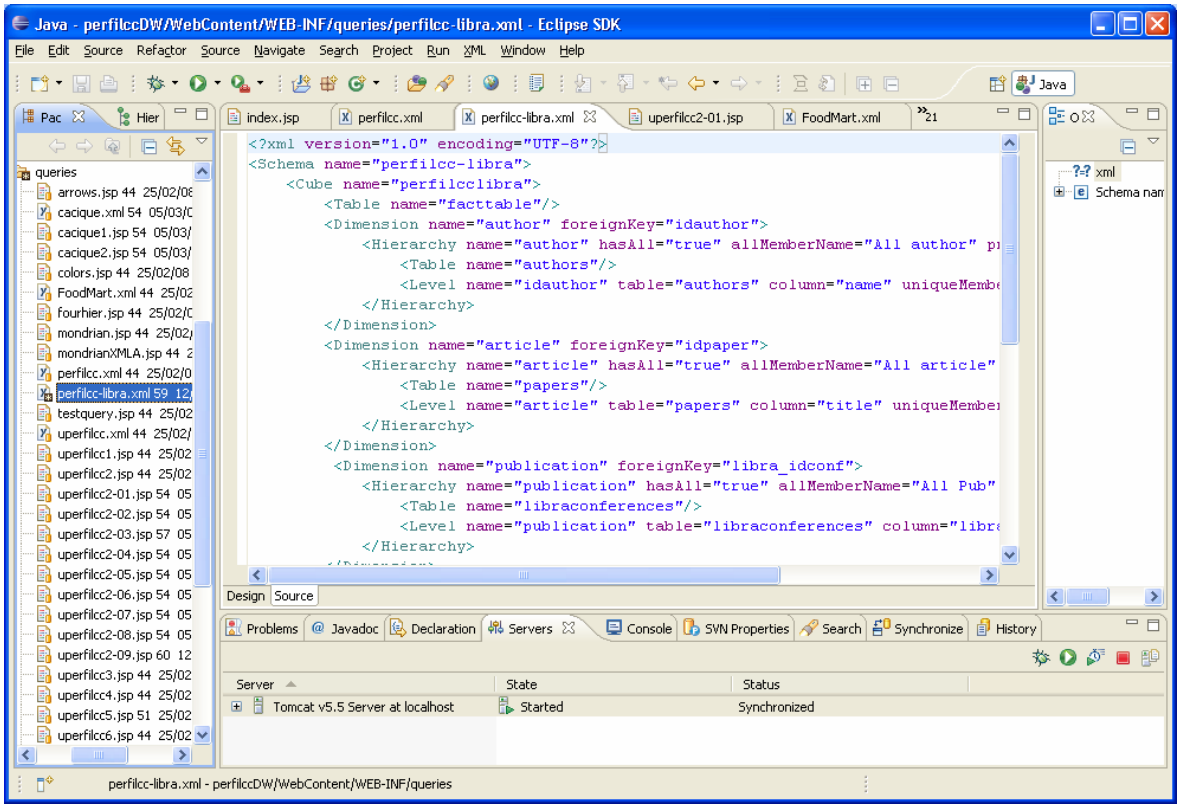


Figura 4.8. Tela do programa utilizado para editar os cubos e consultas (Eclipse).

Figura 4.9. Consultas disponibilizadas no ambiente Web gerado.

demos observar que há uma barra com vários recursos, como, por exemplo, a edição visual da consulta e a exibição do respectivo código, a ordenação dos dados, a exibição do gráfico, versão para impressão e exportação para a planilha.

Artigos por Programa (Filtro: 2006)

		Measures
programGroup	program	DArticles
-All programGroup	+All program	4191
+Brazil	-All program	499
	PUC-Rio	52
	UFMG	43
	UFPE	68
	UFRGS	111
	UFRJ/COPPE	75
	Unicamp	60
	USP/SC	67
	USP/SP	35
+Europe	-All program	778
	Cambridge	216
	ETH Zürich	165
	Imperial College	137
	Oxford	79
	Paris VI	126
	École Polytechnique	62
+N. America	-All program	2706
	British Columbia	89
	Brown	105
	Caltech	27
	CMU	195
	Cornell	125
	Harvard	66
	Illinois	489
	MIT	177
	Princeton	116
	Stanford	275
	Toronto	243
	UC Berkeley	188
	UTexas Austin	179
	Washington	121
	Waterloo	299
	Wisconsin	105

Figura 4.10. Consulta criada na ferramenta OLAP com dados bibliográficos.

A geração automática de gráficos a partir dos dados exibidos nas tabelas é ilustrada na Figura 4.11. A figura mostra de uma forma mais clara a distribuição do número de artigos por programa para o ano de 2006. Podemos configurar vários parâmetros do gráfico, como a fonte, o tipo do gráfico (exemplo: de barras ou de linhas), o tamanho e as cores.

Outro recurso disponível no ambiente desenvolvido é a exibição de detalhes de um dado "macro" ou agregado (*Drill Through*). A Figura 4.12 mostra o número de publicações por ano do professor Marcos André Gonçalves. Se quisermos saber por que o número de publicações em 2004 é 11, clicamos na coluna correspondente ao ano e uma nova tela é exibida com detalhes das publicações, como o título, o veículo, a classificação Qualis e o ano.

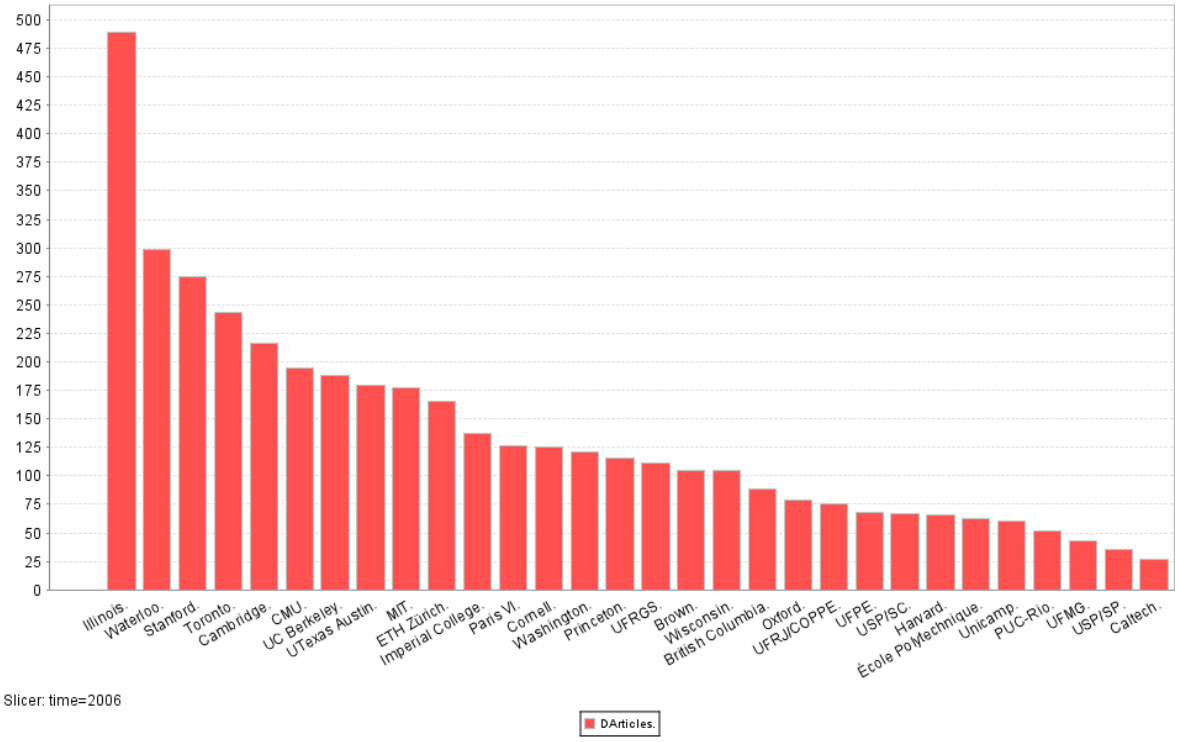
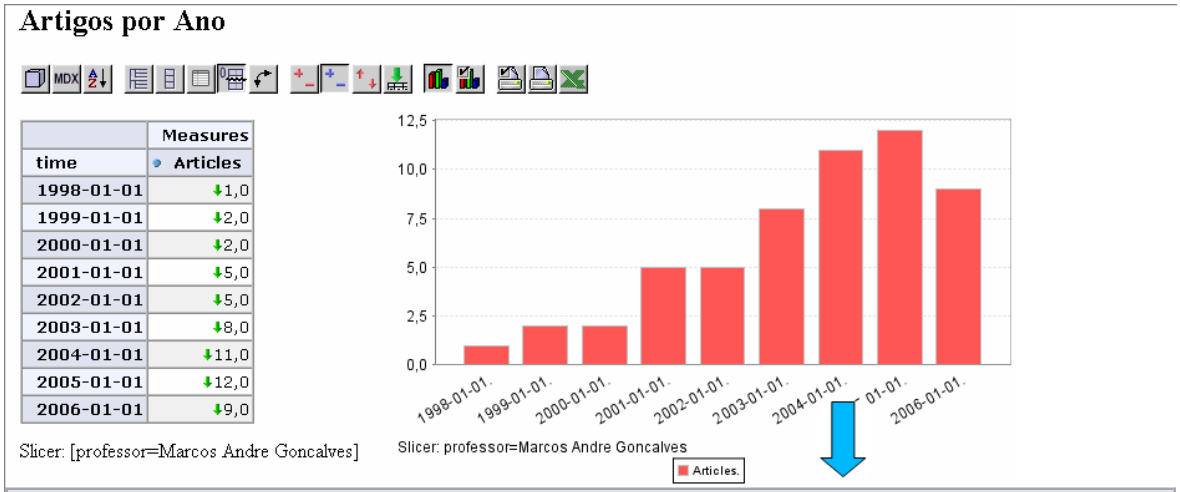


Figura 4.11. Gráfico gerado a partir da consulta Número de Artigos por Programa (2006).

4.4 Considerações Finais

Vimos que para realizar uma implementação bem-sucedida é fundamental uma análise criteriosa do problema a ser estudado. Para isso é importante conhecer bem o domínio a ser analisado (publicações científicas) e a necessidade de informação dos utilizadores do ambiente. É natural que surjam novas demandas de alterações das consultas. Podemos realizar estas alterações com rapidez e baixo custo por meio do conjunto de ferramentas apresentado neste capítulo. Após apresentar o processo de desenvolvimento do armazém de dados de publicações, a próxima seção abordará o estudo do perfil de publicação em Ciência da Computação.



Drill Through Table for publication

title	vehicleShortTitle	qualis	date
An OAI compliant content-based image search component.	JCDL	AI	01/01/2004
BDBComp: building a digital library for the Brazilian computer science community.	JCDL	AI	01/01/2004
Combining structural and citation-based evidence for text classification.	CIKM	AI	01/01/2004
ETANA-DL: a digital library for integrated handling of heterogeneous archaeological data.	JCDL	AI	01/01/2004
ETANA-DL: managing complex information applications -- an archaeology digital library.	JCDL	AI	01/01/2004
Modelagem de Bibliotecas Digitais usando a Abordagem 5S: Um Estudo de Caso.	SBBD	AN	01/01/2004
Prototyping Digital Libraries Handling Heterogeneous Data Sources - The ETANA-DL Case Study.	ECDL	NA	01/01/2004
Recommender Systems Research: A Connection-Centric Survey.	J. Intell. Inf. Syst.	AI	01/01/2004
Streams, structures, spaces, scenarios, societies (5s): A formal model for digital libraries.	ACM Trans. Inf. Syst.	AI	01/01/2004
The effectiveness of automatically structured queries in digital libraries.	JCDL	AI	01/01/2004

Page 1/2 Goto Page 1 Rows/page 10

Figura 4.12. Exemplo de *Drill Through*.

Capítulo 5

Análise dos Dados

O objetivo deste capítulo é apresentar uma análise dos dados utilizando o armazém de dados de publicações científicas. A Seção 5.1 apresenta uma análise do número de publicações. A Seção 5.2 apresenta uma análise de citações recebidas pelos artigos publicados no triênio 2004-2006. A Seção 5.3 apresenta uma análise das subáreas do conhecimento em Ciência da Computação. A Seção 5.4 apresenta considerações finais sobre este capítulo.

5.1 Análise do Número de Publicações

O objetivo desta seção é apresentar uma análise do número de publicações para o triênio 2004-2006. A análise apresenta uma comparação do número de publicações dos programas em conferências e periódicos de qualidade. Para determinar a qualidade do veículo de publicação utilizamos a classificação Qualis criada pela CAPES[1]. Por exemplo, a classe AI corresponde aos veículos internacionais que têm maior impacto científico. O Qualis utilizado foi também referente ao período 2004-2006. Usamos como métrica de produtividade o número de publicações por docente, já que o tamanho do corpo docente dos 30 programas é heterogêneo.

A Tabela 5.1 mostra o resultado do número de publicações Qualis A Internacional por docente no triênio 2004-2006. O desvio padrão do número de publicações por docente está indicado na coluna *DesvPad*. Os dados foram dispostos graficamente na Figura 5.1 com um intervalo de confiança de 95%. Considerou-se apenas os docentes dos 30 programas analisados com pelo menos uma publicação listada na DBLP totalizando 1.760 docentes (87% do total). Os programas nacionais (PUC-Rio, UFRJ/COPPE,

[1]Órgão vinculado ao governo brasileiro responsável pela regulamentação dos cursos de pós-graduação no país

Tabela 5.1. Média de Publicações por Docente no Triênio 2004-2006 dos Programas Analisados (Qualis AI).

	Programa	*Artigos*	*Docentes*	*Art/Doc*	*DesvPad*
1	Illinois	1.005	114	8,82	8,60
2	UC Berkeley	421	54	7,80	5,71
3	Stanford	593	77	7,70	5,54
4	CMU	470	67	7,01	7,26
5	Harvard	122	19	6,42	3,95
6	Princeton	248	40	6,20	4,59
7	MIT	418	68	6,15	6,21
8	Wisconsin	239	40	5,98	4,28
9	Brown	221	37	5,97	5,09
10	UTexas Austin	344	59	5,83	4,67
11	Cornell	285	49	5,82	4,14
12	Washington	267	50	5,34	3,30
13	Toronto	496	95	5,22	5,01
14	ETH Zurich	262	53	4,94	3,70
15	Imperial College	226	48	4,71	5,81
16	Waterloo	474	123	3,85	5,00
17	Caltech	49	14	3,50	2,46
18	PUC-Rio	81	24	3,38	3,59
19	Oxford	125	40	3,13	2,57
20	British Columbia	178	57	3,12	2,24
21	UFMG	103	35	2,94	3,00
22	UFRGS	183	65	2,82	2,39
23	UFPE	110	43	2,56	2,54
24	UFRJ/COPPE	92	36	2,56	2,09
25	Unicamp	90	42	2,14	1,48
26	USP/SC	77	40	1,93	2,50
27	Ecole Polytechnique	91	53	1,72	1,83
28	Paris VI	148	98	1,51	1,20
29	Cambridge	274	182	1,51	1,67
30	USP/SP	57	38	1,50	1,42
-	Total	7.749	1760	4,40	-

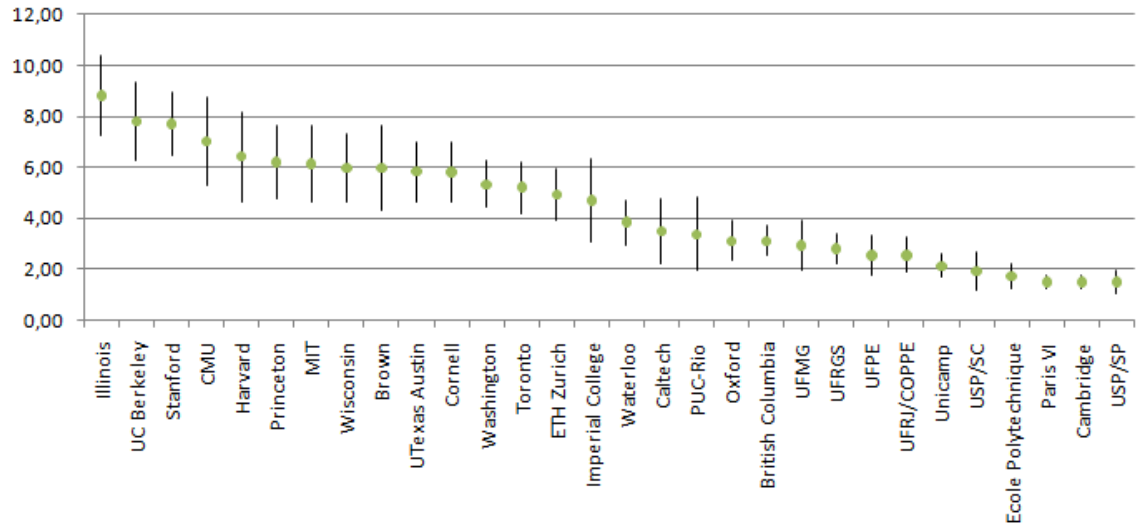


Figura 5.1. Média de Publicações por Docente no Triênio 2004-2006 dos Programas Analisados (Qualis AI).

UFMG, UFPE, UFRGS, UNICAMP, USP/SP e USP/SC) possuem uma média entre 1,50 e 3,38 publicações por docente. Esse volume de publicações é comparável ao de pelo menos seis dos programas internacionais analisados (Waterloo, Caltech, British Columbia, École Polytechnique, Cambridge e Paris VI), os quais possuem uma média entre 1,51 e 3,85 publicações por docente.

Tais resultados mostram que a produção científica dos programas nacionais, representada por artigos publicados em periódicos e conferências internacionais, é comparável em volume e qualidade a de alguns dos principais programas da América do Norte e da Europa.

5.2 Citações Recebidas pelos Artigos

O objetivo desta seção é realizar uma análise do impacto da produção de artigos publicados pelos 30 programas, considerando o número de citações recebidas por cada artigo. A Seção 3.4 apresenta a metodologia utilizada para recuperar as citações dos artigos no Google Scholar.

A Tabela 5.2 mostra a média de citações por artigo dos programas analisados. O desvio padrão do número de citações por artigo está indicado na coluna *DesvPad*. Os dados foram dispostos graficamente na Figura 5.2 com um intervalo de confiança de 95%. O programa de British Columbia possui a maior média de citações por artigo, com 47,48 citações por artigo. Entretanto, cabe ressaltar que esse resultado está suportado

principalmente por um único artigo com 3.959 citações, o que corresponde a quase metade de todas as 8.166 citações do programa. Se retirarmos esse artigo, a média cai para 23,25 e o programa vai para a 14^a posição. Esse é um caso isolado, além dele, há apenas mais um artigo com mais de 1.000 citações dentre todos os 7.749 artigos da amostra considerada.

Tabela 5.2. Citações por artigo no período 2004-2006

#	*Programa*	*Artigos*	*Citações*	*Cit/Art*	*DesvPad*
1	British Columbia	172	8.166	47,48	301,97
2	UC Berkeley	356	15.840	44,49	90,44
3	MIT	418	18.428	44,09	80,11
4	Washington	219	9.318	42,55	66,68
5	Cornell	274	11.146	40,68	77,29
6	Stanford	541	20.040	37,04	57,78
7	ETH Zurich	179	5.671	31,68	34,28
8	Wisconsin	206	6.464	31,38	52,35
9	Princeton	236	7.036	29,81	42,65
10	Harvard	112	3.174	28,34	49,89
11	Illinois	1.000	26.675	26,68	92,95
12	UTexas Austin	332	8.763	26,39	47,57
13	CMU	455	11.840	26,02	40,57
14	Brown	206	5.103	24,77	44,65
15	Oxford	116	2.675	23,06	49,27
16	Caltech	47	1.062	22,6	34,69
17	Toronto	453	9.562	21,11	38,30
18	Cambridge	268	5.290	19,74	33,57
19	Waterloo	400	7.572	18,93	28,25
20	Ecole Polytechnique	90	1.507	16,74	25,40
21	UFMG	95	1.396	14,69	23,94
22	PUC-Rio	80	1.155	14,44	26,23
23	Imperial College	222	2.867	12,91	15,94
24	Paris VI	143	1.631	11,41	25,67
25	USP/SP	53	486	9,17	16,90
26	UFRGS	181	1.555	8,59	17,64
27	Unicamp	85	691	8,13	13,09
28	UFRJ/COPPE	88	528	6	8,72
29	USP/SC	71	407	5,73	15,88
30	UFPE	100	543	5,43	10,19
	Total	7.198	196.591	27,31	-

Ainda considerando a Tabela 5.2, os seis programas mais bem colocados são norte-americanos (British Columbia, UC Berkeley, MIT, Washington, Cornell e Stanford). O programa do ETH Zürich vem logo em seguida e é o primeiro programa europeu. UFMG e PUC-Rio possuem valores muito próximos e ocupam as posições 21 e 22, respectivamente, seguidos pelos programas do Imperial College e de Paris VI. Logo

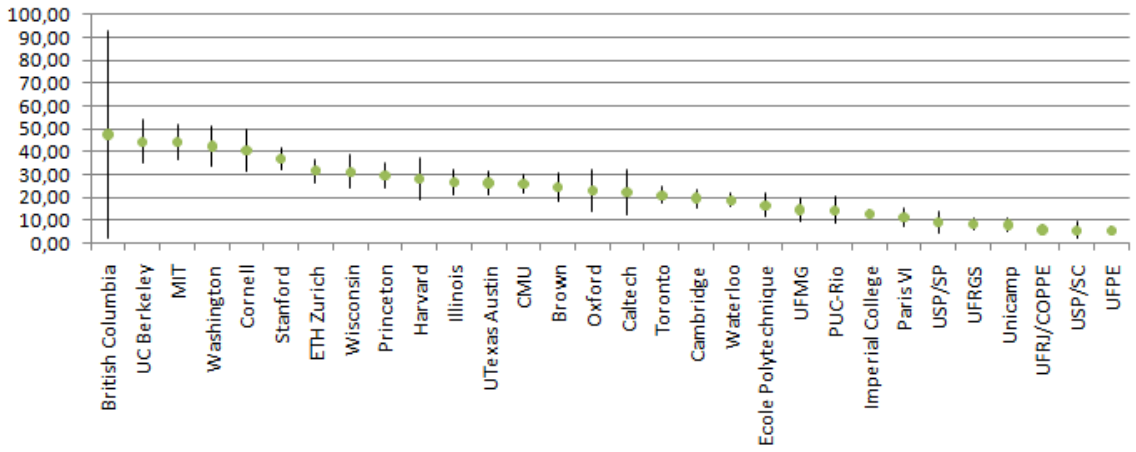


Figura 5.2. Citações por artigo no período 2004-2006.

a seguir estão os demais programas brasileiros (USP/SP, UFRGS, Unicamp, UFRJ, USP/SC, UFPE).

Agrupando os programas por região, os programas da América do Norte possuem em média 31,36 citações por artigo contra 19,29 dos europeus e 8,98 dos brasileiros. Se considerarmos os dois programas brasileiros mais bem colocados (UFMG e PUC-Rio), essa média vai para 14,57, próxima à dos programas de Waterloo, da École Polytechnique, do Imperial College e de Paris VI (citações na faixa 11,41 - 18,93), mostrando a inserção internacional desses dois programas brasileiros em termos do impacto de suas publicações.

Conforme discutido anteriormente, outro importante indicador bibliométrico é o índice h [Hirsch, 2005]. Um índice h igual a 5 indica que foram publicados cinco artigos com pelo menos cinco citações no período considerado. A vantagem do índice h em relação às citações por artigo é que ele não é influenciado por poucos artigos de grande visibilidade. Os dados obtidos de British Columbia comprovam este fato. Ressalva-se que os programas com maior número de docentes tendem a ter maior índice h por publicarem mais artigos em número absoluto. O índice h por programa pode ser relativizado calculando-se o índice h médio por docente.

A Tabela 5.3 mostra o índice h médio por docente dos programas analisados no triênio 2004-2006. O desvio padrão do índice h por docente está indicado na coluna *DesvPad*. Os dados foram dispostos graficamente na Figura 5.3 com um intervalo de confiança de 95%. O índice h de cada programa corresponde à média aritmética do índice h de cada docente. Novamente foram considerados apenas os artigos cuja classificação Qualis da CAPES seja A Internacional.

A classificação é encabeçada pelos programas da UC Berkeley e de Stanford, os

Tabela 5.3. Índice h médio por docente do programa - período 2004-2006

#	*Programa*	*Índice h*	*DesvPad*
1	UC Berkeley	5,80	4,50
2	Stanford	5,77	4,02
3	Princeton	5,17	3,37
4	Illinois	5,14	3,81
5	Harvard	4,89	2,89
6	Washington	4,82	2,80
7	Cornell	4,69	3,04
8	CMU	4,68	4,00
9	Wisconsin	4,62	2,93
10	MIT	4,54	3,70
11	Brown	4,43	2,94
12	UTexas Austin	4,39	2,61
13	Toronto	3,58	2,66
14	Imperial College	3,46	2,73
15	ETH Zurich	3,00	2,03
16	Waterloo	2,74	2,37
17	Caltech	2,72	1,87
18	British Columbia	2,68	1,88
19	UFMG	2,63	2,24
20	Oxford	2,23	1,57
21	PUC-Rio	2,04	1,87
22	UFRGS	1,96	1,15
23	Unicamp	1,81	1,12
24	USP/SC	1,48	0,98
25	Ecole Polytechnique	1,46	1,21
26	UFRJ/COPPE	1,33	1,37
27	UFPE	1,30	1,21
28	Paris VI	1,24	0,89
29	Cambridge	1,21	1,11
30	USP/SP	1,13	0,93
-	Total	3,18	-

quais ganharam várias posições em relação à métrica citações por artigo. British Columbia, que possui a maior taxa de citações por artigo, sustentado principalmente por um único artigo, foi para a posição 18. UFMG, Puc-Rio, UFRGS e Unicamp estão entre os programas brasileiros com melhor média do índice h. Os programas brasileiros possuem um índice h médio próximo ao de sete programas internacionais: Waterloo, Caltech, British Columbia, Oxford, École Polytechnique, Paris VI e Cambridge. Agrupando os programas por região, os docentes dos programas da América do Norte possuem um índice h médio de 4,42, contra 2,10 dos europeus e 1,71 dos brasileiros.

A Tabela 5.4 mostra os programas analisados em ordem decrescente. Os dados foram dispostos graficamente na Figura 5.4. A diferença desta medida para a anterior

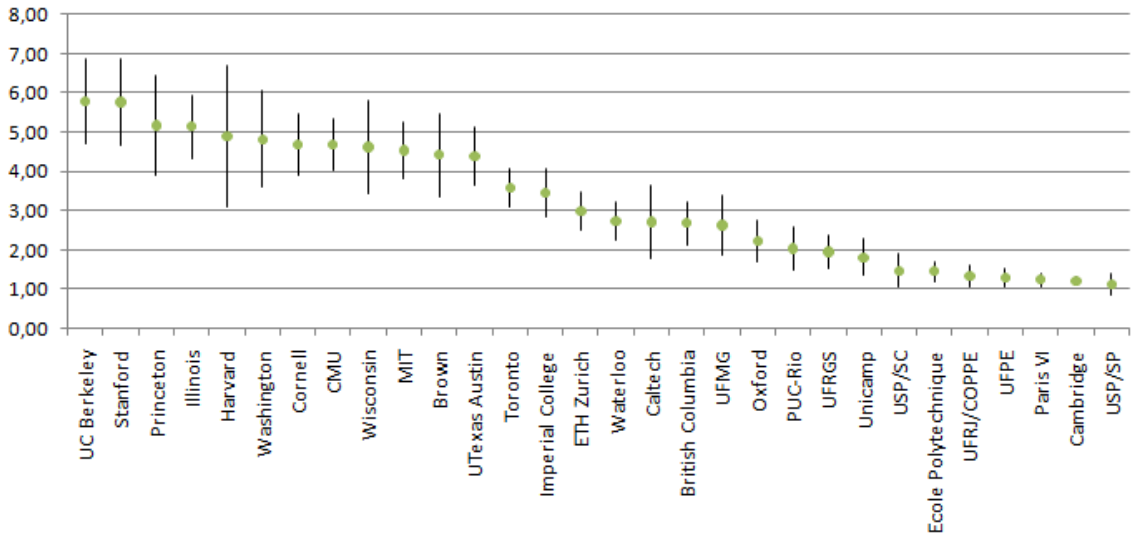


Figura 5.3. Índice h médio por docente do programa - período 2004-2006.

é que agora utiliza-se o índice h do programa e não a média dos índices h dos docentes dos programas.

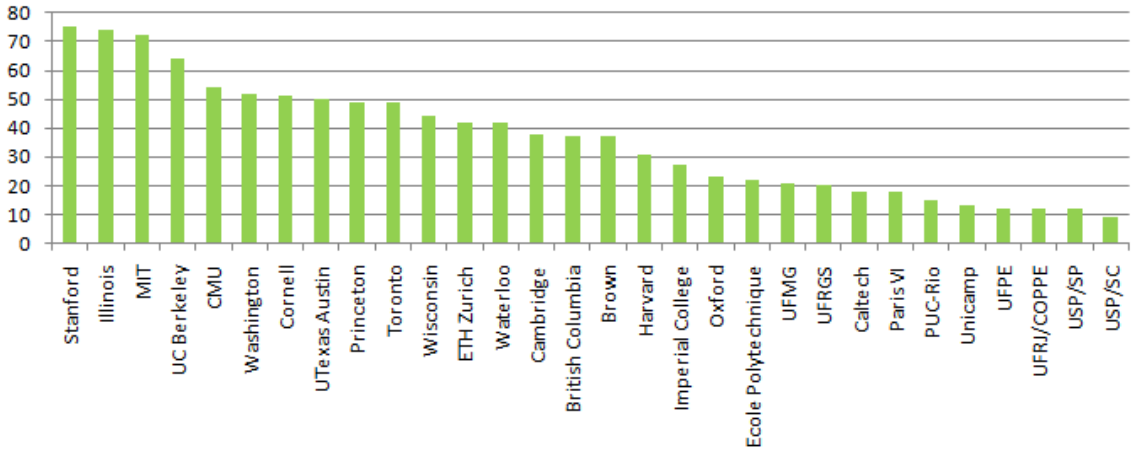


Figura 5.4. Índice h por programa - período 2004-2006.

O índice h de programas privilegia os programas com maior corpo docente, por terem mais publicações no total. A classificação é encabeçada pelos programas de Stanford e de Illinois, que ganharam várias posições em relação à taxa de citações por artigo (5 e 9 respectivamente). British Columbia, que possui a maior taxa de citações por artigo, sustentado principalmente por um único artigo, foi para a posição 15. ETH Zürich é o programa europeu com melhor resultado. UFMG e UFRGS são os brasileiros de maior impacto com índice h 21 e 20, respectivamente. Os dois programas brasileiros

Tabela 5.4. Índice h por programa - período 2004-2006

#	*Programa*	*Docentes*	*Índice h*
1	Stanford	77	75
2	Illinois	114	74
3	MIT	68	72
4	UC Berkeley	54	64
5	CMU	67	54
6	Washington	50	52
7	Cornell	49	51
8	UTexas Austin	59	50
9	Princeton	40	49
10	Toronto	95	49
11	Wisconsin	40	44
12	ETH Zürich	53	42
13	Waterloo	123	42
14	Cambridge	182	38
15	British Columbia	57	37
16	Brown	37	37
17	Harvard	19	31
18	Imperial College	48	27
19	Oxford	40	23
20	École Polytechnique	53	22
21	UFMG	35	21
22	UFRGS	65	20
23	Caltech	14	18
24	Paris VI	98	18
25	PUC-Rio	24	15
26	Unicamp	42	13
27	UFPE	43	12
28	UFRJ/COPPE	36	12
29	USP/SP	38	12
30	USP/SC	40	9
-	Total	1.760	36,10

mais bem posicionados possuem um índice h no mesmo patamar dos programas de Oxford, da École Polytechnique, da Caltech e de Paris VI.

5.3 Análise das Subáreas em Ciência da Computação

Nesta seção analisamos o perfil das publicações das subáreas de Ciência da Computação. Para tal usamos as 30 subáreas criadas no contexto do Projeto Perfil-CC. As 30 subáreas foram levantadas por meio de consultas a especialistas de diversas áreas da computação no país (para uma descrição da metodologia utilizada veja [Laender et al., 2008]). Para fazer a análise, criamos consultas que retornam a quantidade de artigos por subáreas e por programas de pós-graduação. Um exemplo de consulta seria a que obtém a quantidade de artigos de Inteligência Artificial publicados pelos programas da PUC-Rio e da COPPE/UFRJ.

O Apêndice A apresenta as subáreas mais fortes em cada programa. As Tabelas A.1, A.2 e A.3 apresentam os programas brasileiros, norte-americanos e europeus mais produtivos por subárea. Novamente foram considerados os artigos publicados no triênio 2004-2006 e cuja classificação de acordo com o Qualis da CAPES seja A Internacional.

A Figura 5.5 mostra um exemplo de consulta criada para análise das subáreas. Essa consulta lista o número de publicações por programa por subárea. Desta consulta podemos destacar a produção da COPPE/UFRJ na área de *Algoritmos e Teoria da Computação*, da UFPE em *Aprendizado de Máquina* e da UFRGS em *Arquitetura de Computadores, Processamento de Alto Desempenho*.

Sub-áreas mais frequentes (Filtro: Qualis AI, Triênio: 2004-2006)

		Measures	
subarea	program	DArticles	NumProfessors
Algoritmos e Teoria da Computação	PUC-Rio	5	1
	UFMG	2	2
	UFPE	2	1
	UFRGS	2	2
	UFRJ/COPPE	12	4
	Unicamp	6	3
	USP/SP	8	10
Aprendizado de Máquina	UFMG	3	1
	UFPE	36	7
	UFRGS	7	3
	UFRJ/COPPE	10	4
	Unicamp	1	1
	USP/SP	1	1
Arq Comput, Proc Alto Desempenho	PUC-Rio	3	1
	UFMG	5	5
	UFPE	4	5
	UFRGS	16	12
	UFRJ/COPPE	8	6

Figura 5.5. Consulta criada no ambiente para análise das subáreas.

Para uma melhor visualização dos dados retornados pelas consultas, utilizamos um programa de análise gráfica de dados, o TreeMap [Shneiderman & Wattenberg, 2001], desenvolvido sob a coordenação do professor Ben Shneiderman da Universidade de Maryland, EUA. O Treemap mostra linhas de dados como grupos de retângulos que podem ser organizados, dimensionados e coloridos para revelar graficamente padrões desconhecidos. Esta técnica de visualização de dados permite aos usuários um reconhecimento mais fácil de complexos relacionamentos de dados que não são óbvios por outras técnicas[2].

A Figura 5.6 mostra os dados da Figura 5.5 visualizados de forma gráfica com a distribuição das subáreas das publicações dos programas. Quanto maior o retângulo, maior o número de artigos do programa de pós-graduação. Quanto mais claro a cor do retângulo, maior é o corpo docente da subárea. Um mesmo professor pode publicar em mais de uma subárea. A partir da Figura 5.6 podemos concluir que a UFRGS possui a maior quantidade de artigos em números absolutos, seguida pela UFPE e pela UFMG.

A Tabela 5.5 lista as principais subáreas dos programas brasileiros e as principais subáreas considerando todos os programas. Algumas áreas entre os programas analisados destacam-se pelo número de artigos: a área de *Concepção de Circuitos Integrados* e *Inteligência Artificial* na UFRGS, *Aprendizado de Máquina* na UFPE e *Bancos de Dados e Bibliotecas Digitais* na UFMG, todos com mais de 25 artigos no triênio. Considerando todos os programas, as áreas com mais publicações são *Engenharia de Software*, *Métodos Formais*, *Aprendizado de Máquina* e *Inteligência Artificial*, todos com mais de 60 artigos no triênio.

A Figura 5.7 mostra a comparação do número de publicações e o número de docentes entre todos os 30 programas analisados. No geral, alguns programas norte-americanos se sobressaem em quantidade de publicações: Illinois, Waterloo e Stanford. Os europeus mais bem classificadas são Cambridge, ETH Zürich e Imperial College. Os brasileiros mais bem classificadas estão no mesmo patamar de Princeton, Paris VI, Wisconsin, Brown, British Columbia, Oxford, Harvard, École Polytechnique e Caltech. No geral, o número de docentes dos programas brasileiros é menor que o dos programas norte-americanos (o gráfico mostra uma coloração mais escura). Considerando apenas as publicações da subárea de *Arquitetura de Computadores* de Illinois, a quantidade de publicações é maior que todas as publicações da Caltech, ressaltando que o corpo docente da Caltech é oito vezes menor.

Podemos notar que cada programa "especializa-se" numa determinada subárea e que a distribuição das subáreas é heterogênea. Além disto, as áreas mais tradicionais

[2]Wikipedia. http://www.wikipedia.org. Acessado em Março/2008

Tabela 5.5. Principais subáreas por programa de pós-graduação.

Programa	*Área*	*Artigos*
TODOS	Engenharia de Software, Métodos Formais	65
	Aprendizado de Máquina	65
	Inteligência Artificial	64
UFRGS	Concepção de Circuitos Integrados	49
	Inteligência Artificial	29
	Arquitetura de Computadores, Processamento de Alto Desempenho	16
UFPE	Aprendizado de Máquina	36
	Sistemas de Informação	14
	Inteligência Artificial	14
	Engenharia de Software, Métodos Formais	14
UFMG	Bancos de Dados, Bibliotecas Digitais	27
	Redes de Computadores, Sistemas Distribuídos, Sistemas P2P	12
	Recuperação de Informação	11
UFRJ/COPPE	Engenharia de Software, Métodos Formais	14
	Algoritmos e Teoria da Computação	12
	Computação Aplicada	10
	Aprendizado de Máquina	10
Unicamp	Arquitetura de Computadores, Processamento de Alto Desempenho	10
	Sistemas de Informação	9
	Multitemáticas	9
PUC-Rio	Engenharia de Software, Métodos Formais	16
	Sistemas de Informação	14
	Web, Sistemas Multimídia e Hipermídia	8
USP/SC	Web, Sistemas Multimídia e Hipermídia	10
	Multitemáticas	10
	Engenharia de Software, Métodos Formais	10
USP/SP	Arquitetura de Computadores, Processamento de Alto Desempenho	10
	Computação Gráfica, Processamento de Imagens	8
	Algoritmos e Teoria da Computação	8

Figura 5.6. Subáreas - Programas Brasileiros. Filtros: Qualis AI, Anos 2004-2006.

da Ciência da Computação são as que possuem mais artigos: *Arquitetura de Computadores*, *Redes de Computadores*, *Bancos de Dados*, *Algoritmos* e *Inteligência Artificial*.

A Figura 5.8 faz uma análise dos programas europeus. Foram considerados os artigos de classificação Qualis A Internacional no triênio 2004-2006. Os programas com mais artigos são os de Cambridge, ETH Zürich e Imperial College. Em Cambridge notamos várias subáreas com muitas publicações e muitos docentes, em detrimento do Imperial College, onde as áreas com muitas publicações possuem poucos docentes (*Computação Gráfica*, *Concepção de Circuitos Integrados*, *Engenharia de Software* e *Métodos Formais*). Isso pode ser um indício que o grau de colaboração entre esses docentes seja menor no Imperial College do que em Cambridge.

Por fim, a Figura 5.9 mostra os dados dos programas norte-americanos. Foram considerados os artigos de classificação Qualis A Internacional no triênio 2004-2006. O programa com mais artigos é o de Illinois, que possui várias publicações nas áreas de

Figura 5.7. Subáreas - Todos os programas. Filtros: Qualis AI, Anos 2004-2006.

Arquitetura de Computadores, *Redes de Computadores*, *Bancos de Dados* e *Computação Gráfica*. Outros programas em destaque são os de Stanford, Toronto e UC Berkeley. A subárea *Algoritmos* é a de maior destaque em seis programas: Waterloo, MIT, Cornell, Washington, Princeton e Harvard. Outras áreas de destaque são *Arquitetura de Computadores*, *Computação Gráfica*, *Redes de Computadores* e *Bancos de Dados*. A subárea de *Arquitetura de Computadores* de Illinois possui mais artigos que todas as subáreas de Harvard, essa por sua vez possui um corpo docente bem menor que Illinois (6 vezes menor).

5.4 Considerações Finais

Os resultados obtidos mostram que o número de publicações dos programas nacionais é comparável a de alguns dos principais programas da América do Norte e da Europa. É possível perceber por meio das publicações que os programas se especializam em determinadas subáreas da Ciência da Computação. Quando analisamos os dados das

Figura 5.8. Subáreas - Programas Europeus. Filtros: Qualis AI, Anos 2004-2006.

citações, percebemos que o impacto da produção científica dos programas nacionais é comparável ao impacto dos programas estrangeiros analisados.

Figura 5.9. Subáreas - Programas norte-americanos. Filtros: Qualis AI, Anos 2004-2006.

Capítulo 6

Conclusões e Trabalhos Futuros

Neste trabalho, foi realizada uma análise multidimensional da produção científica em Ciência da Computação. Para isso foi desenvolvido um armazém de dados para publicações científicas. Os dados das publicações foram obtidos em três etapas: primeiramente foi realizada uma coleta manual dos docentes dos programas analisados, seguida da coleta automática dos artigos dos autores na biblioteca digital DBLP, e por fim, da coleta automática das citações no Google Scholar. O armazém de dados gerado permite uma série de análises sobre publicações científicas.

O repositório foi disponibilizado na Web[1] para permitir a sua utilização e consulta sem a necessidade de se instalar ferramentas adicionais. Uma ferramenta OLAP foi usada para permitir uma análise dos dados, trazendo flexibilidade e facilidade na geração das consultas.

O armazém de dados gerado foi utilizado para estudar o perfil de publicação dos principais programas de pós-graduação em Ciência da Computação do país. Os resultados obtidos complementam o estudo realizado no contexto do projeto Perfil-CC [Laender et al., 2008] com uma análise do impacto das citações dos artigos publicados pelos programas e o estudo da distribuição das publicações por subáreas da Ciência da Computação.

Considerando as publicações entre 2004-2006 cuja classificação Qualis é A Internacional, os programas nacionais (PUC-Rio, UFRJ/COPPE, UFMG, UFPE, UFRGS, UNICAMP, USP/SP e USP/SC) possuem uma média entre 1,50 e 3,38 publicações por docente. Esse volume de publicações é comparável ao de pelo menos seis dos programas internacionais analisados (Oxford, Caltech, British Columbia, École Polytechnique, Paris VI e Cambridge) que possuem uma média entre 1,51 e 3,85 publicações por docente.

[1] http://www.latin.dcc.ufmg.br:8080/perfilccDW/

Concluímos que o número de publicações dos programas brasileiros é comparável aos da América do Norte e da Europa, o que mostra a inserção internacional desses programas.

Analisamos também as citações recebidas pelas publicações geradas pelos programas. Considerando as publicações do período 2004-2006 cuja classificação Qualis é A Internacional, na época da coleta dos dados (junho de 2008), o número de citações por artigo é 8,98 para os programas brasileiros, 19,29 para os europeus e 31,36 para os norte-americanos. Se considerarmos os dois programas brasileiros mais bem colocados (UFMG e PUC-Rio), essa média vai para 14,57, próximo à de Waterloo, École Polytechnique, Imperial College e Paris VI (citações na faixa 11,41 - 18,93), mostrando que o impacto da produção científica desses programas, representada por artigos publicados em periódicos e conferências internacionais, é comparável ao impacto de alguns dos principais programas da América do Norte e da Europa.

Complementando a análise com o índice h, os programas brasileiros possuem um índice h médio próximo ao de sete programas internacionais: Waterloo, Caltech, British Columbia, Oxford, École Polytechnique, Paris VI e Cambridge. Em média, o índice h dos docentes dos programas brasileiros é 1,71, contra 2,10 dos europeus e 4,42 dos norte-americanos.

Analisamos ainda a distribuição das subáreas dos programas. Foi possível notar que cada programa "especializa-se" numa determinada subárea e que a distribuição dessas subáreas é heterogênea. Além disto, as subáreas mais tradicionais da Ciência da Computação, Arquitetura de Computadores, Redes de Computadores, Bancos de Dados, Algoritmos e Inteligência Artificial, são as que apresentam maior produtividade em termos de artigos publicados.

Com relação a trabalhos futuros que possam complementar o que foi desenvolvido, destacamos:

- A adição dos dados de outros programas de pós-graduação em Ciência da Computação ao armazém de dados, como, por exemplo, os programas brasileiros com conceitos 3 e 4 na CAPES ou programas de outros países, para ampliar o escopo da análise.

- Avaliação do perfil evolutivo dos programas de pós-graduação para avaliar a tendência de um programa ter seu conceito melhorado na CAPES.

- Utilização do arcabouço desenvolvido para análise de programas de outras áreas, como Física ou Biologia.

Referências Bibliográficas

Abelló, A.; Samos, J. & Saltor, F. (2001). A Framework for the Classification and Description of Multidimensional Data Models. In *Proceedings of the 12th International Conference on Database and Expert Systems Applications*, pp. 668–677, Munich, Germany.

Alvarado, R. U. (2009). Obsolescência da literatura sobre a Lei de Lotka. *Revista de Ciência da Informação*, 10(1):1–18.

Arruda, D.; Bezerra, F.; Neris, V. A.; de Toro, P. R. & Wainer, J. (2009). Brazilian Computer Science research: gender and regional distributions. *Scientometrics*, 79(3):651–665.

Barbieri, C. (2001). *BI - Business Intelligence - Modelagem & Tecnologia*. Axcel Books.

Campos, M. L. & Rocha Filho, A. V. (1997). Data warehouse. In *XVI Jornada de Atualização em Informática*, pp. 221–261, Brasília, DF.

Chaudhuri, S. & Dayal, U. (1997). An overview of data warehousing and OLAP technology. *ACM Sigmod Record*, 26(1):65–74.

Codd, E. F.; Codd, S. B. & Salley, C. T. (1993). Providing OLAP to user-analysts: An IT mandate. San Jose, California. E.F. Codd and Associates.

Garfield, E. & Merton, R. (1979). *Citation indexing: Its theory and application in science, technology, and humanities*. Wiley, New York.

Hirsch, J. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences*, 102(46):16569–16572.

Hüsemann, B.; J., L. & G., V. (2000). Conceptual Data Warehouse design. In *2nd. International Workshop On Design and Management of Data Warehouses*, pp. 3–9, Stockholm, Sweden.

Inmon, W. H. (1996). *Building the data warehouse*. Wiley.

Kimball, R. & Ross, M. (2002). *The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling*. Wiley.

Kimball, R.; Ross, M.; Thornthwaite, W.; Mundy, J. & Becker, B. (2008). *The Data Warehouse Lifecycle Toolkit*. Wiley, 2 edição.

Laender, A. H. F.; de Lucena, C. J. P.; Maldonado, J. C.; de Souza e Silva, E. & Ziviani, N. (2008). Assessing the research and education quality of the top Brazilian Computer Science graduate programs. *ACM SIGCSE Bulletin Inroads*, 40(2):135–145.

Ley, M. (2002). The DBLP Computer Science Bibliography: Evolution, Research Issues, Perspectives. In *Proceedings of the 9th International Symposium on String Processing Information Retrieval*, pp. 1–10, Lisboa, Portugal.

Ley, M. & Reuther, P. (2006). Maintaining an Online Bibliographical Database: The Problem of Data Quality. In *Actes des Sixièmes Journées Extraction et Gestion des Connaissances*, pp. 5–10, Lille, France.

Martins, W. S. (2009). Abordagens para Avaliação Automática de Conferências Científicas: Um Estudo de Caso em Ciência da Computação. Master's thesis, Programa de Pós-Graduação em Ciência da Computação da Universidade Federal de Minas Gerais.

Martins, W. S.; Gonçalves, M. A.; Laender, A. H. F. & Pappa, G. L. (2009). Learning to Assess the Quality of Scientific Conferences: A Case Study in Computer Science. In *Proceedings of the 9th ACM/IEEE Joint Conference on Digital Libraries*, pp. 193–202, Austin, Texas.

Mattern, F. (2006). Bibliometric Evaluation of Computer Science - Problems and Pitfalls. *Invited Talk, SARIT 06: Swiss IT Professors' Day, Swiss Association for Research in Information Technology, Basel, Switzerland.*

Menezes, G. V.; Ziviani, N. & Laender, A. H. F. (2008). Um Estudo Comparativo de Redes Sociais em Ciência da Computação. In *Workshop on Information Visualization and Analysis in Social Networks*, pp. 1–10, Campinas, SP.

Menezes, G. V.; Ziviani, N.; Laender, A. H. F. & Almeida, V. (2009). A Geographical Analysis of Knowledge Production in Computer Science. In *Proceedings of the 18th International Conference on World Wide Web*, pp. 1041–1050, Madrid, Spain.

Messaoud, R. B.; Boussaid, O. & Rabaséda, S. (2004). A new OLAP aggregation based on the AHC technique. In *Proceedings of the 7th ACM International Workshop on Data Warehousing and OLAP*, pp. 65–72, Washington, DC.

Nascimento, M. A.; Sander, J. & Pound, J. (2003). Analysis of SIGMOD's co-authorship graph. *SIGMOD Record*, 32(3):8–10.

Newman, M. E. J. (2004). Coauthorship Networks and Patterns of Scientific Collaboration. *Proceedings of the National Academy of Sciences*, 101:5200–5205.

Pedersen, D.; Riis, K. & Pedersen, T. B. (2002). A powerful and SQL-compatible data model and query language for OLAP. *Australian Computer Science Communications*, 24(2):121–130.

Pedersen, T. B.; Shoshani, A.; Gu, J. & Jensen, C. S. (2000). Extending OLAP querying to external object databases. In *Proceedings of the 9th International Conference on Information and Knowledge Management*, pp. 405–413, McLean, VA.

Petricek, V.; Cox, I. J.; Han, H.; Councill, I. G. & Giles, C. L. (2005). Comparison of On-line Computer Science Citation Databases. In *Proceedings of 9th European Conference on Research and Advanced Technology for Digital Libraries*, pp. 438–449, Vienna, Austria.

Shneiderman, B. & Wattenberg, M. (2001). Ordered Treemap Layouts. In *Proceedings of the IEEE Symposium on Information Visualization 2001*, pp. 73–78, San Diego, California.

Shoshani, A. (1997). OLAP and statistical databases: similarities and differences. In *Proceedings of the 16th ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*, pp. 185–196, Tucson, Arizona.

Silva, A. J.; Modesto, M. A.; Gonçalves, M. A.; Cristo, M.; Laender, A. H. & Ziviani, N. (2006). Busca pelo Texto Completo de Artigos Catalogados em uma Biblioteca Digital. In *II Workshop de Bibliotecas Digitais*, pp. 71–80, Florianópolis, SC.

Trujillo, J.; Palomar, M. & Gómez, J. (2000). Applying Object-Oriented Conceptual Modeling Techniques to the Design of Multidimensional Databases and OLAP applications. In *Proceedings of 1st International Conference on Web-Age Information Management*, pp. 83–94, Shanghai, China.

Wainer, J.; Xavier, E. C. & Bezerra, F. (2009). Scientific production in Computer Science: A comparative study between Brazil and other countries. *Scientometrics*, 81:535–547.

Apêndice A

Tabelas

Tabela A.1. Programas brasileiros mais produtivos por subárea (período 2004-2006).

Subárea	*Programa*	*Art*	*Doc*	*Subárea*	*Programa*	*Art*	*Doc*
Algoritmos e Teoria da Computação	PUC-Rio	5	1	Geoinformática	UFRGS	3	4
	UFRJ/COPPE	12	4	Informática na Educação	UFRGS	5	3
	Unicamp	6	3	Inteligência Artificial	PUC-Rio	6	3
	USP/SP	8	10		UFPE	14	6
Aprendizado de Máquina	UFMG	3	1		UFRGS	29	11
	UFPE	36	7		UFRJ/COPPE	3	4
	UFRGS	7	3		USP/SC	5	3
	UFRJ/COPPE	10	4		USP/SP	5	3
	USP/SC	8	3	IHC, Sistemas Colaborativos	PUC-Rio	5	2
Arq Comput, Proc Alto Desempenho	PUC-Rio	3	1	Linguagens de Programação	PUC-Rio	6	2
	UFMG	5	5		UFPE	5	2
	UFPE	4	5	Multitemáticas	PUC-Rio	7	5
	UFRGS	16	12		UFMG	4	4
	UFRJ/COPPE	8	6		UFPE	4	4
	Unicamp	10	7		UFRGS	5	8
	USP/SP	10	4		UFRJ/COPPE	5	5
Bancos de Dados, Bibliot Digitais	PUC-Rio	3	3		Unicamp	9	7
	UFMG	27	6		USP/SC	10	6
	UFRGS	10	8	Pesq Operac e Otimiz Combinatória	UFRJ/COPPE	6	5
	UFRJ/COPPE	3	3		Unicamp	8	5
	Unicamp	7	4		USP/SP	5	5
	USP/SC	4	3	Proc de Língua Natural	UFRGS	3	2
Biologia Computacional	USP/SP	6	4		USP/SC	7	2
Computação Aplicada	UFPE	3	2	Recuperação de Informação	UFMG	11	4
	UFRJ/COPPE	10	5	Redes, Sist Distrib, Sistemas P2P	UFMG	12	6
	Unicamp	5	4		UFPE	4	4
	USP/SC	7	2		UFRGS	10	6
Computação Gráfica, PD Imagens	UFRGS	12	7		Unicamp	8	4
	Unicamp	4	4	Simulação e Modelagem	UFMG	3	3
	USP/SC	4	3	Sistemas de Informação	PUC-Rio	14	9
	USP/SP	8	4		UFMG	3	4
Computação Ubíqua	UFMG	5	4		UFPE	14	9
Concepção de Circuitos Integrados	UFPE	4	2		UFRGS	11	5
	UFRGS	49	9		Unicamp	9	5
	Unicamp	3	5		USP/SC	3	4
Eng Soft, Métodos Formais	PUC-Rio	16	3	Sist Emb, T. Real, Tol. Falhas	UFRGS	4	2
	UFMG	6	6		Unicamp	3	3
	UFPE	14	5	Web, Sist Multimíd e Hipermíd	PUC-Rio	8	4
	UFRGS	5	4		UFMG	7	7
	UFRJ/COPPE	14	3		UFRGS	3	5
	Unicamp	3	2		UFRJ/COPPE	3	2
	USP/SC	10	10		Unicamp	3	3
Formalismos, Lógica, Semântica	UFRJ/COPPE	8	4		USP/SC	10	6

Tabela A.2. Programas da América de Norte mais produtivos por subárea (período 2004-2006).

Subárea	*Programa*	*Art*	*Doc*	*Subárea*	*Programa*	*Art*	*Doc*
Algoritmos e Teoria da Computação	CMU	35	9	Concepção de Circuitos Integrados	Stanford	27	7
	Cornell	44	10		Toronto	22	5
	Harvard	24	4	Eng Soft, Métodos Formais	CMU	22	6
	Illinois	46	12		Illinois	37	22
	MIT	68	11		Toronto	38	7
	Princeton	52	10		Waterloo	35	11
	Stanford	38	13	Inteligência Artificial	CMU	48	14
	Toronto	22	12		Harvard	24	7
	UC Berkeley	37	12		Illinois	63	18
	Washington	31	6		Stanford	46	18
	Waterloo	79	30		Toronto	48	20
Aprendizado de Máquina	CMU	35	12		UTexas Austin	44	9
	Illinois	39	7		Washington	24	10
	Stanford	42	16		Waterloo	44	19
	Toronto	28	14	IHC, Sistemas Colaborativos	Toronto	23	4
	UTexas Austin	44	9	Linguagens de Programação	Illinois	30	16
Arq Comput, Proc Alto Desempenho	CMU	32	14		MIT	27	10
	Illinois	147	36		Stanford	26	13
	MIT	38	15		UC Berkeley	29	10
	Princeton	42	13		UTexas Austin	26	9
	Stanford	55	16		Waterloo	28	13
	Toronto	33	13	Mineração de Dados	Illinois	30	7
	UC Berkeley	40	18	Multitemáticas	UC Berkeley	35	15
	UTexas Austin	45	13	Redes, Sist Distrib, Sistemas P2P	Cornell	27	11
	Wisconsin	49	12		Illinois	102	25
Bancos de Dados, Bibliot Digitais	CMU	44	11		MIT	48	20
	Cornell	24	6		Stanford	42	15
	Illinois	87	11		Toronto	29	12
	Stanford	55	13		UC Berkeley	51	14
	Toronto	62	10		UTexas Austin	34	11
	UC Berkeley	31	8		Wisconsin	37	12
	Washington	24	3	Segurança e Privacidade	Illinois	23	11
	Waterloo	46	11		Stanford	30	9
	Wisconsin	33	7	Sist Emb, T. Real, Tol. Falhas	Illinois	53	14
Biologia Computacional	Stanford	22	7	Visão Computacional	CMU	29	5
Computação Gráfica, PD Imagens	British Columbia	35	13		Illinois	65	8
	Brown	29	9		MIT	28	5
	Illinois	80	13		Toronto	28	11
	MIT	48	12	Web, Sist Multimíd e Hipermíd	Illinois	42	14
	Stanford	72	15				
	Toronto	28	11				
	UC Berkeley	34	8				
	Washington	25	9				
	Waterloo	60	13				

Tabela A.3. Programas da Europa mais produtivos por subárea (período 2004-2006).

Subárea	*Programa*	*Art*	*Doc*	*Subárea*	*Programa*	*Art*	*Doc*
Algoritmos e Teoria da Computação	Cambridge	17	12	Inteligência Artificial	Cambridge	4	2
	É. Polytechnique	16	19		ETH Zürich	5	5
	ETH Zürich	30	10		Imperial College	14	8
	Imperial College	13	10		Oxford	7	3
	Oxford	14	8		Paris VI	21	12
Aprendizado de Máquina	Cambridge	7	8	IHC, Sistemas Colaborativos	Cambridge	7	7
	ETH Zürich	10	4		ETH Zürich	4	2
	Imperial College	4	2		Oxford	4	3
	Paris VI	11	8	Linguagens de Programação	Cambridge	35	12
Arq Comput, Proc Alto Desempenho	Cambridge	22	20		É. Polytechnique	8	7
	ETH Zürich	25	10		ETH Zürich	4	4
	Imperial College	11	9		Imperial College	15	7
	Paris VI	9	5		Oxford	24	10
Bancos de Dados, Bibliot Digitais	Cambridge	5	4	Multitemáticas	Cambridge	12	10
	ETH Zürich	22	6		É. Polytechnique	7	3
	Oxford	10	3		ETH Zürich	5	5
	Paris VI	5	5		Imperial College	6	4
Biologia Computacional	Cambridge	7	5		Oxford	7	7
	É. Polytechnique	5	4	Pesq Operac e Otimiz Combinatória	É. Polytechnique	7	7
	ETH Zürich	6	3		ETH Zürich	6	2
	Imperial College	8	3		Oxford	4	3
Computação Aplicada	Cambridge	7	5		Paris VI	17	12
	ETH Zürich	5	5	Proc de Língua Natural	Cambridge	10	8
	Oxford	8	6	Recuperação de Informação	Cambridge	5	3
	Paris VI	4	4		Paris VI	6	3
Computação Gráfica, PD Imagens	Cambridge	7	7	Redes, Sist Distrib, Sistemas P2P	Cambridge	26	18
	ETH Zürich	26	8		É. Polytechnique	9	4
	Imperial College	50	3		ETH Zürich	30	8
Computação Ubíqua	Cambridge	12	13		Imperial College	9	10
	ETH Zürich	9	5		Paris VI	26	13
	Paris VI	5	4	Robótica, Controle e Automação	Paris VI	4	4
Concepção de Circuitos Integrados	Cambridge	8	8	Segurança e Privacidade	Cambridge	10	8
	ETH Zürich	4	2		ETH Zürich	11	5
	Imperial College	34	2	Simulação e Modelagem	Cambridge	4	4
Eng Soft, Métodos Formais	Cambridge	13	8		Paris VI	6	4
	ETH Zürich	9	4	Sistemas de Informação	ETH Zürich	10	3
	Imperial College	20	3		Imperial College	10	11
	Oxford	9	7		Paris VI	8	7
Formalismos, Lógica, Semântica	Cambridge	28	17	Sist Emb, T. Real, Tol. Falhas	Cambridge	13	14
	É. Polytechnique	25	11		É. Polytechnique	5	2
	ETH Zürich	5	1		ETH Zürich	7	4
	Imperial College	17	12		Paris VI	5	6
	Oxford	27	11	Visão Computacional	ETH Zürich	11	3
				Web, Sist Multimíd e Hipermíd	Cambridge	6	5
					ETH Zürich	11	5