PAULO JOSÉ LAGE ALVARENGA

UM ESTUDO SOBRE REFERÊNCIAS BIBLIOGRÁFICAS NA ÁREA DE CIÊNCIA DA COMPUTAÇÃO

Belo Horizonte
29 de março de 2007

Universidade Federal de Minas Gerais
Instituto de Ciências Exatas
Programa de Pós-Graduação em Ciência da Computação

UM ESTUDO SOBRE REFERÊNCIAS BIBLIOGRÁFICAS NA ÁREA DE CIÊNCIA DA COMPUTAÇÃO

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

PAULO JOSÉ LAGE ALVARENGA

Belo Horizonte
29 de março de 2007

UNIVERSIDADE FEDERAL DE MINAS GERAIS

FOLHA DE APROVAÇÃO

Um estudo sobre referências bibliográficas na área de Ciência da Computação

PAULO JOSÉ LAGE ALVARENGA

Dissertação defendida e aprovada pela banca examinadora constituída por:

Ph. D. Berthier Ribeiro de Araújo Neto – Orientador
Universidade Federal de Minas Gerais

Ph. D. Marcos André Gonçalves
Universidade Federal de Minas Gerais

Ph. D. Nivio Ziviani
Universidade Federal de Minas Gerais

Belo Horizonte, 29 de março de 2007

Resumo

Resultados de pesquisa científica e tecnológica são normalmente publicados como artigos em periódicos (em Inglês, "*journals*"), conferências e seminários. Nas áreas de ciências naturais, como Matemática, Física e Biologia, esses artigos são publicados predominantemente em periódicos ("*journals*"). Este mesmo padrão é observado em certas áreas de pesquisa tecnológica como as engenharias. Na área de Ciência da computação, entretanto, há razões para suspeitar que publicações em conferências têm adquirido um destaque crescente. De fato, correntemente o volume de artigos em conferências supera em muito aquele de artigos em periódicos. Se assim é, há de se especular se o impacto de artigos de conferências na área de Ciência da Computação é comparável ao impacto de artigos de periódicos. Este trabalho é um estudo das publicações dos principais autores da área de Ciência da Computação nos últimos 30 anos, tanto em conferências quanto em periódicos. Nosso estudo sugere que, na área de Ciência da Computação, conferências tiveram um crescimento grande o suficiente para tornar os trabalhos nelas veiculados tão importantes quanto os trabalhos publicados em periódicos.

Abstract

The results from technological and scientific research are normally discussed in articles published in journals, conferences and seminaries. In the areas of natural sciences, such as Mathematics, Physics and Biology, research results are normally published in journals. The same pattern is observed in certain technological areas such as engineering. In the Computer Science area, however, there are reasons to suspect that publications in conferences have acquired an increasing prominence. For instance, the number of articles published in conferences now surpasses the number of articles in journals. This work presents a study of the publications from the main authors in Computer Science in the last 30 years, both in conferences and in journals. Our study suggests that conferences have grown large enough to make their impact as important as that of journals, in the area of Computer Science.

Aos meus pais, Socorro e Virgílio, e à minha noiva, Denise.

Agradecimentos

Depois de um período repleto de aprendizado, trabalhos, muitos erros, novas experiências, vários experimentos, muitas vezes refeitos muitas vezes (sim, a redundância faz parte do aprendizado), e depois disso tudo até mesmo alguns acertos, é chegado o momento de concluir mais uma etapa. Mas este trabalho é apenas um dos resultados deste intenso período, que contou com o apoio de várias pessoas por mim muito queridas, a quem homenageio com estes agradecimentos.

Primeiramente agradeço a Deus, por tornar tudo possível e por me guiar em todos os momentos.

Agradeço a meu orientador, Berthier Ribeiro-Neto, pela sugestão do tema, pelas novas e tantas vezes surpreendentes idéias, pelo conhecimento transmitido em cada reunião e disciplina presenciada, pelo apoio dado e pelas palavras certas em cada momento;

ao professor Marcos Gonçalves e ao doutorando Denilson Pereira, pela disponibilidade em ajudar neste trabalho, em participar de reuniões e pelas várias críticas, sugestões e opiniões que contribuíram para a melhoria do trabalho;

aos professores Nivio Ziviani, Virgílio Almeida e Wagner Meira Júnior, pelo engrandecimento proporcionado em suas disciplinas, e também ao Fabiano Botelho, cuja monitoria em P.A.A. tornou um pouco menos doloroso o "sofrimento"[1] passado, e aos vários colegas que me acompanharam nas disciplinas estudadas;

ao professor Luis Zárate, da PUC-Minas, por, ainda na graduação, me introduzir à pesquisa acadêmica, transmitir seus importantes ensinamentos, pela confiança e orientação que me levaram a tentar e conseguir passar pelo teste de seleção para Mestrado na UFMG, assim como aos demais professores, amigos e colegas de Ciência da Computação da PUC-Minas, principalmente David, Joel, Lucas, Matheus, Rodrigo e Vitor, que contribuíram para meu crescimento acadêmico, profissional e pessoal;

aos grupos de pesquisa LATIN, LBD e SPEED e seus professores responsáveis, pela disponibilidade não apenas de um (no caso, três) laboratório(s), mas do ambiente que

[1] Segundo o professor Wagner Meira Júnior, o diferencial, o que realmente fica após o mestrado ou doutorado, é o "sofrimento" que passamos para concluir o estudo.

possibilitou tantas trocas de idéias, essenciais para decidir qual caminho trilhar em vários momentos deste trabalho;

aos amigos e colegas do DCC, que serão sempre lembrados pelos diversos momentos desses três anos de mestrado, principalmente mas não exclusivamente: Alan, André Bigonha, André Silva, Barroca, Hélio, Humberto Nigri, Marco Modesto e Thyerson.

aos funcionários do DCC, por serem sempre tão prestativos e atenciosos;

aos amigos e colegas de trabalho, desde o começo na Trajeto, Receita Federal, JLP, XPRO, até atualmente, na FUNCESI, pela experiência profissional obtida;

à CAPES e ao CNPq, que me auxiliaram por quase 2 anos com a bolsa de estudos;

aos amigos da Anatomia da Dança (ICB), pelos vários momentos de descontração e reânimo em vários momentos difíceis, principalmente ao Felipe pelas aulas à noite, e à Karin pelas aulas no horário de almoço e pelas massagens nos momentos de grande tensão;

aos amigos diversos que, apesar de não se enquadrarem em nenhum dos grupos acima desempenham um papel importante em minha vida, principalmente Carlos, Daniel, Demétrius e Jean;

aos familiares, que souberam entender minha ausência em diversas festas, férias (praticamente inexistentes) e demais eventos, e serem por algumas vezes praticamente ignorados em visitas, em decorrência a vários prazos e compromissos inadiáveis do mestrado, agradeço pelo carinho dado, por torcerem pelo meu sucesso e pelo apoio dado nos diversos momentos da minha vida;

e finalmente, mas de maneira nenhuma menos importantes, agradeço aos meus pais, Socorro e Virgílio, pela criação, por seus sábios ensinamentos durante minha vida, e à minha noiva, Denise, que, juntos, souberam tantas vezes me compreender, entender (e às vezes não entender) o quão difícil foi essa jornada, por muitas vezes aguentarem meu mal humor, e mesmo assim nunca deixarem de me apoiar, incentivar, de acreditar em mim, de me dar carinho e amor, por serem esse pilar triplo, minha base, meu chão, e, cada um à sua maneira, contribuir para tornar possível mais esta conquista.

A vocês, que tanto me auxiliaram não apenas no mestrado, mas na minha vida, muito obrigado.[2]

[2]Durante os agradecimentos citei nominalmente cada pessoa apenas uma vez, e dei preferência para a ordem alfabética. Muitas dessas pessoas deveriam estar em mais de um desses grupos, e outras foram removidas para que os agradecimentos não ocupassem mais espaço que a dissertação em si. Estendo, aqui, meus agradecimentos a estas pessoas.

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

Os resultados de pesquisa científica e tecnológica são normalmente divulgados através de artigos veiculados em periódicos (em Inglês, "*journals*"), conferências e seminários. Nas áreas de ciências naturais, como Matemática, Física e Biologia, os resultados de pesquisa são, contudo, veiculados em artigos publicados predominantemente em periódicos ("*journals*"). Este mesmo padrão é observado em certas áreas de pesquisa tecnológica como as engenharias.

Em outras áreas de tecnologia, entretanto, é comum que novos resultados de pesquisa sejam primeiramente veiculados em anais de conferências de grande reputação. Este é particularmente o caso da Ciência da Computação. Trata-se de uma tendência relativamente recente, fortemente influenciada pela continuada e rápida evolução da tecnologia na área e pelo comparativamente longo tempo requerido para publicação em periódicos.

Apesar desta particularidade da área de Ciência da Computação, organismos vários de suporte à pesquisa, como a Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (Capes) e o Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), consideram, para efeito de avaliação de projetos e de currículo de pesquisadores, basicamente publicações em periódicos. Isso se deve à tradição nas áreas de ciências naturais e à relativamente recente história da pesquisa em Ciência da Computação.

Tal critério de avaliação não faz distinção entre as áreas do conhecimento, o que pode comprometer a avaliação da excelência acadêmica, particularmente em áreas de tecnologia muito dinâmica como a área de Ciência da Computação. De fato, uma vez que os melhores periódicos de Ciência da Computação podem levar de 2 a 3 anos para publicar resultados recentes de pesquisa, temos que considerar que tais resultados podem não ser mais de grande interesse da comunidade, uma vez que já foram largamente discutidos nas principais conferências. No pior cenário, resultados veiculados em periódicos podem se encontrar obsoletos quando de sua publicação. Como resultado,

vários autores na área de Ciência da Computação publicam predominantemente em conferências.

Este trabalho tem como objetivo estudar o perfil de publicação dos principais autores da área de Ciência da Computação, estabelecendo uma análise comparativa de suas publicações em conferências e periódicos ao longo do tempo.

1.1 Contribuições

As principais contribuições deste trabalho são:

- Criação de um banco de dados contendo informações referentes às publicações acadêmicas dos 1.000 autores mais citados na área de Ciência da Computação, de acordo com o *site* CiteSeer, bem como as referências a estas publicações, categorizadas de acordo com o veículo em que foram publicadas.
- Identificação do tipo de veículo predominantemente utilizado por estes autores, se são periódicos ou conferências.
- Verificação do impacto destes veículos no decorrer do tempo, impacto aqui definido como o número de citações recebidas por publicação, assim como estimativa de aumento ou queda de impacto.

Nosso estudo permite concluir que:

- Os autores mais prolíficos na área de Ciência da Computação publicam eqüitativamente em conferências e periódicos. Ademais, o número de citações por publicação tem crescido mais rapidamente para artigos publicados em conferências.
- Os números absolutos indicam ainda um maior número de artigos muito citados ocorrendo em periódicos, o que acreditamos ser devido à tradição dos periódicos.
- A partir de meados da década de 90, verifica-se uma mudança no comportamento dos autores, resultando em um número maior de publicações em conferências.
- Não há mais indícios indicando que a produção científica veiculada em periódicos é mais importante do que a produção científica veiculada em conferências, na área de Ciência da Computação.

O trabalho está organizado em 6 capítulos. Este primeiro capítulo, a introdução, descreve o problema, a abordagem e os objetivos deste trabalho. O segundo capítulo contém a revisão bibliográfica, tratando sobre a metodologia utilizada no trabalho,

suas áreas e como a bibliografia se insere no trabalho. O terceiro capítulo trata sobre a caracterização do problema, a necessidade de se coletar dados, as fontes de dados, a metodologia e a forma com que a coleta foi realizada. O quarto capítulo trata da categorização dos dados coletados por tipo de artigo. O quinto capítulo inclui uma análise dos dados coletados, bem como uma comparação entre periódicos e conferências, ao longo do tempo. O sexto e último capítulo conclui o trabalho com uma análise final dos resultados encontrados e uma descrição de trabalhos futuros.

Capítulo 2

Revisão Bibliográfica

O estudo de referências bibliográficas na área da Ciência da Computação envolve basicamente três etapas. A primeira delas é a coleta de meta-informação sobre as publicações que permita a criação de um banco de dados central. Nesta etapa são usadas estratégias de coleta de dados na Web, assim como inferências estatísticas para determinar o tamanho amostral necessário para a coleta.

A seguir, é feita uma categorização dos dados referentes aos veículos de publicação, sendo estes separados em publicações de conferências, periódicos, *workshops*, livros, relatórios técnicos. Identificamos ainda quais artigos são da área de Ciência da Computação e quais artigos são de outras áreas.

A terceira e última etapa é uma análise e interpretação dos dados obtidos. Este capítulo trata da revisão bibliográfica referente às três etapas do trabalho, descritas anteriormente.

2.1 Modelo Vetorial

De acordo com Baeza-Yates e Ribeiro-Neto [3], o modelo vetorial é um dos três modelos clássicos de recuperação de informação. Nos modelos clássicos de recuperação de informação, cada documento é descrito por um conjunto de palavras-chave representativas chamadas *termos indexados*. Um *termo indexado* é uma palavra, existente em um ou mais documentos. A pesquisa por documentos de interesse é feita a partir da submissão, pelo usuário, de uma consulta composta por termos indexados.

Seja t o número de termos no sistema e k_i um termo indexado genérico. $K = \{k_1, \ldots, k_t\}$ é o conjunto de todos os termos indexados. Um peso $w_{i,j} > 0$ é associado com cada termo indexado k_i de um documento d_j. Para um termo indexado que não aparece no texto do documento, $w_{i,j} = 0$. Com o documento d_j é associado um vetor de termos indexados $\overrightarrow{d_j}$ representado por $\overrightarrow{d_j} = (w_{1,j}, w_{1,j}, \ldots, w_{t,j})$. Além disso, seja g_i

uma função que retorne o peso associado com o termo indexado k_i em qualquer vetor t-dimensional.[3]

Para o modelo vetorial, o peso $w_{i,j}$ associado com o par (k_i, d_j) é positivo e não-binário. Além disso, os termos indexados da consulta também possuem pesos. Seja $w_{i,q}$ o peso associado com o par $[k_{i,j}, q]$, onde $w_{i,q} \geq 0$. Então, o vetor da consulta $\overrightarrow{q}$ é definido como $\overrightarrow{q} = (w_{1,q}, w_{2,q}, \ldots, w_{t,q})$ onde t é o número total de termos indexados no sistema. Como anteriormente, o vetor para um documento d_j é representado por $\overrightarrow{d_j} = (w_{1,j}, w_{1,j}, \ldots, w_{t,j})$.[3]

Portanto, um documento d_j e uma consulta q são representados como vetores em um espaço t-dimensional. A proposta do Modelo Vetorial é avaliar o grau de similaridade entre o documento d_j e a consulta q como a correlação entre os vetores $\overrightarrow{d_j}$ e $\overrightarrow{q}$. A proposta é quantificar esta correlação através do cálculo do cosseno dado pelo ângulo entre os dois vetores. Matematicamente:

$$sim(d_j, q) = \frac{\overrightarrow{d_j} \bullet \overrightarrow{q}}{\left|\overrightarrow{d_j}\right| \times \left|\overrightarrow{q}\right|} = \frac{\sum_{i=1}^{t} w_{i,j} \times w_{i,q}}{\sqrt{\sum_{i=1}^{t} w_{i,j}^2} \times \sqrt{\sum_{i=1}^{t} w_{i,q}^2}}$$

onde $\left|\overrightarrow{d_j}\right|$ e $\left|\overrightarrow{q}\right|$ são as normas dos vetores do documento e da consulta. Como será calculada a similaridade entre a mesma consulta e cada um dos documentos, o fator referente à norma da consulta $\left|\overrightarrow{q}\right|$ não afetará a ordenação dos resultados, podendo ser omitido da fórmula.

Uma estratégia muito freqüente para geração dos pesos $w_{i,j}$ e $w_{i,q}$ é conhecida como tf-idf (*term frequency-inverse document frequency*, ou freqüência do termo no documento, inverso da freqüência do termo na coleção). O fator *tf* mede a importância do termo para o documento através da quantidade de vezes que o termo k_i ocorre no documento d_j, ou seja, provê uma medida de quão bem o termo descreve o conteúdo do documento. O fator *idf* mede o quão discriminante é o termo através do inverso da freqüência do termo k_i em todos documentos da coleção. O objetivo de utilizar o *idf* é atribuir peso maior a termos mais raros e reduzir o peso de termos muito freqüentes na coleção, pois estes últimos são menos importantes na caracterização da necessidade de informação do usuário.

Seja N o número total de documentos no sistema e n_i o número de documentos no qual o termo de índice k_i aparece. Seja $freq_{i,j}$ a freqüência do termo k_i no documento d_j (ex. o número de vezes que o termo k_i é mencionado no texto do documento d_j). Caso o termo k_i não apareça no documento d_j, então $freq_{i,j} = 0$. A freqüência normalizada $tf_{i,j}$ do termo k_i no documento d_j é dada por $tf_{i,j} = \frac{freq_{i,j}}{max_l freq_{i,j}}$ onde o máximo é computado sobre todos os termos mencionados no texto do documento d_j. Se o termo k_i não ocorre no documento d_j então $tf_{i,j} = 0$. Além disso, seja o fator *idf* dado por $idf_i = \log \frac{N}{n_i}$ O esquema de ponderação de termos conhecido como *tf-idf* usará pesos dados por:

$w_{i,j} = tf_{i,j} \times idf_i; w_{i,q} = idf_i$

O uso do modelo vetorial neste trabalho será melhor detalhado no Capítulo 4.

2.2 Índice Jaccard

O índice Jaccard[5], também conhecido como coeficiente de similaridade de Jaccard, é um método estatístico utilizado para medir a similaridade e a diversidade entre conjuntos de elementos do mesmo tipo. O coeficiente de Jaccard é definido como $J(A,B) = \frac{|A \bigcap B|}{|A \bigcup B|}$, onde A e B são dois conjuntos de elementos do mesmo tipo. Neste estudo, o índice Jaccard é utilizado de forma a medir a similaridade entre os nomes dos veículos das publicações aqui coletadas e os nomes dos veículos presentes em uma base de referência.

Tal base de referência foi construída tendo como fonte bibliotecas digitais da área da computação, como a DBLP e ACM-DL. A base de referência de veículos da Ciência da Computação foi indexada, de modo a possibilitar a submissão de consultas compostas pelo nome do veículo das publicações coletadas, e é a partir destes resultados que é feito o cálculo de Jaccard entre o veículo do artigo e os resultados da consulta. Desta forma, a partir de um limite mínimo de similaridade, calculado empiricamente, os veículos da resposta são aceitos ou rejeitados. Sendo assim, o cálculo é feito utilizando os conjuntos A =palavras da consulta e B =palavras indexadas. O valor retornado é um número entre 0, quando não existem palavras em comum, e 1, quando todas as palavras estão presentes nos dois conjuntos.

Os valores mínimos aceitos, assim como a forma como são encontrados, serão discutidos com maiores detalhes no Capítulo 4.

2.3 Trabalhos Relacionados

Em nossa pesquisa não encontramos trabalhos que comparam publicações de conferências com publicações de periódicos. Vários trabalhos analisam o fator de impacto de periódicos [7] [8][9], calculado pelo *Institute for Scientific Information* (ISI) e mostram que não é adequado avaliar um artigo apenas baseado no periódico onde foi publicado.

O *Nature Editorial*[7] mostra que o fator de impacto é influenciado por uma pequena minoria de artigos, e uma análise feita com base em artigos individuais no periódico *Nature* mostrou que 89% do resultado referente a 2004 foi gerado por apenas 25% de seus artigos. Dentre os artigos de 2002-03, o artigo mais citado recebeu 522 citações em 2004, o segundo mais citado recebeu 351 citações, e dentre os demais 1800 artigos, apenas 50 receberam mais de 100 citações, tendo a grande maioria recebido menos de

20 citações. "O fator de Impacto não nos diz tanto quanto algumas pessoas podem pensar sobre a qualidade respectiva de ciência que periódicos estão publicando". Eles concluem dizendo que o problema é as organizações regulamentadoras de pesquisa avaliarem a qualidade científica de nações e instituições, e até mesmo julgarem indivíduos baseados no fator de impacto. Nós pensamos que um fato similar ocorre com artigos de conferências, e artigos de conferências são muitas vezes prejudicados por muitas organizações.

O *PLoS Medicine Editorial*[8] diz que o fator de impacto de periódicos pode ser substancialmente afetado por publicações contendo revisões de uma sub-área de artigos, que normalmente recebem mais citações do que artigos de pesquisa, ou publicações de apenas alguns artigos muito citados. O editorial também diz que existem várias maneiras de "jogar o jogo do fator de impacto". Editores de muitos periódicos encorajam autores a citar artigos publicados em periódicos ou publicar revisões que irão coletar um grande número de citações. Em nossos dados coletados é possível verificar este cenário, sendo possível notar muitos artigos publicados em conferência e que foram posteriormente revisados ou apenas re-publicados em periódicos, o que contribui para a vantagem de publicações em periódicos. Além disso, estas publicações revisadas de periódicos receberam mais citações de outras publicações de periódicos, e então parte dessas citações podem existir devido ao jogo do fator de impacto.

Seglen [14] critica o fator de impacto de periódicos e mostra alguns casos nos quais instituições em diversos países utilizam-no para avaliar indivíduos e instituições. No Brasil, as agências de pesquisa governamentais CNPq e Capes definem métricas para avaliar publicações nas quais publicações de periódicos recebem peso cerca de 3 vezes maior que publicações de conferências.

Hecht et al. [9] concluem que não existe nada errado com o fator de impacto propriamente dito, porém ele possui um nome errado que leva a conclusões erradas. Dessa forma, ele tem sido usado erroneamente. Ele tem sido usado como uma medida da importância de um artigo específico de um periódico e do periódico no qual o artigo apareceu. Por extensão, o fator de impacto é usado erroneamente para medir a importância relativa de pesquisadores individuais, programas de pesquisa e até mesmo a instituição na qual a pesquisa é feita. Eles recomendam que o termo fator de impacto seja abolido e que a medida seja renomeada de modo a manter sua função atual, que é apenas a de medir a taxa de citações em um intervalo de tempo específico, e nada além disso.

Lawrence [10] mostra que artigos livremente disponibilizados online são mais citados. Ele não comparou publicações de periódicos com publicações de conferências, ele analisou apenas artigos de conferências em Ciência da Computação e disciplinas relacionadas. Ele disse que na Ciência da Computação um percentual substancial da

literatura encontra-se disponibilizado integralmente na Internet, artigos de conferências são tipicamente publicações formais e muitas vezes têm mais prestígio que artigos de periódicos, com a taxa de aceitação para algumas conferências abaixo de 10%.

O impacto estimado de veículos de publicação na área da Ciência da Computação feito pelo CiteSeer[15], em Maio de 2003, apresenta um cálculo de fator de impacto utilizando uniformemente todas as conferências e periódicos cadastrados em sua base de dados. Nessa estimativa, 8 conferências estão entre os 10 veículos de maior impacto, sendo os 2 restantes periódicos, e se expandirmos o foco para os primeiros 50 veículos, um total de 32 veículos são conferências, sendo os 18 restantes periódicos. Dessa forma, caso fosse considerado o fator de impacto como uma boa métrica para este tipo de julgamento, seria possível inferir que conferências possuem maior impacto no topo, tendendo a igualar-se ao impacto dos periódicos à medida que avançamos na ordenação realizada.

Rousseau e Rousseau [13] analisaram citações para publicações de periódicos na Informetrics 87/88 e Informetrics 89/90 e mostram que elas são distribuições de Zipf, por exemplo, existem poucos periódicos muito citados e muitos periódicos recebendo poucas citações. Esta característica também foi identificada em nosso trabalho, não apenas para periódicos, mas também para as conferências.

Capítulo 3

Coleta de Dados

3.1 O problema

Na área de Ciência da Computação, publicações são feitas de forma diferente de várias outras áreas de conhecimento, pois notoriamente os artigos publicados em conferências são muito citados, e muitas vezes não possuem sua versão em periódicos. Como forte exemplo, podemos citar o artigo mais citado de S. Brin e J. Page[4], de acordo com o *Google Scholar*, que possui sua publicação submetida apenas em uma conferência (Proceedings of the Seventh World Wide Web Conference), embora vários artigos publicados nesta conferência tenham sido veiculados em um periódico (Computer Networks and ISDN Systems). Também podemos usar como indício uma análise feita pelo CiteSeer, na qual é demonstrado o fator de impacto de todos veículos da área de Ciência da Computação. Neste estudo, 8 dos 10 primeiros veículos com maior fator de impacto são conferências.

Para analisar esta tendência, uma sugestão seria verificar a importância dos artigos não considerando apenas o veículo de publicação, mas sua importância real no meio acadêmico. Esta verificação será feita com base no conceito de autoridade, ou seja: se um autor é muito citado, ele é uma fonte confiável em sua área de conhecimento.

Analisando a forma de publicação em que os principais autores da área veiculam seus trabalhos, torna-se possível identificar a forma de publicação de artigos de maior impacto.

Para a realização de nosso estudo, utilizamos várias bibliotecas digitais específicas da área de Ciência da Computação, que passamos a descrever.

CiteSeer

O CiteSeer[18] é auto-denominado

> "...uma biblioteca digital de literatura científica e máquina de busca que

> enfoca primariamente a literatura em Ciência da Computação. O objetivo do CiteSeer é melhorar a disseminação e *feedback* da literatura científica e prover melhorias em funcionalidade, usabilidade, avaliabilidade, custo, compreensão, eficiência e rapidez no acesso de conhecimento científico e acadêmico."[16]

Foi desenvolvido pelo NEC Research Institute pelos pesquisadores Steve Lawrence, Lee Giles e Kurt Bollacker, e está atualmente hospedado pela Penn State's School of Information Sciences and Technology, aos cuidados do professor Lee Giles.

Contendo artigos de uma quantidade estimada entre 243.837 e 411.032 autores[1], todos automaticamente indexados a partir de versões eletrônicas dos mesmos, encontrados na *Web*, possui diversos recursos que tornariam interessantes análises a partir de sua base, sendo estes recursos, principalmente, a disponibilidade de estatísticas baseadas em todos documentos citados, não apenas os que estão indexados; disponibilidade de metadados do artigo (como ano de publicação, nome do veículo, nomes dos autores, entre outros); e disponibilidade gratuita dos dados. No entanto, durante a primeira tentativa de obtenção de seus dados, o CiteSeer apresentava uma grande instabilidade, permanecendo fora do ar por semanas ininterruptas, e quando apresentava-se operacional, sua máquina de busca não retornavam resultados condizentes com a consulta submetida. Além disso, a inexistência destes dados em outra API tornou inviável a sua utilização para a obtenção dos dados dos artigos a serem analisados.

DBLP

A DBLP[11] é um *site* de bibliografia da área da Ciência da Computação, mantido pela Universität Trier, na Alemanha, e existe desde a década de 1980. Em Novembro de 2004 ela listava 566.666 artigos na área de Ciência da Computação, e dados de Outubro de 2006 informam que existem mais de 800.000 artigos indexados. Apesar de ser altamente organizado, possuir uma ampla lista de documentos de diversos autores, assim como uma precisa categorização entre periódicos e conferências, a DBLP não inclui dados importantes para nossa pesquisa, como dados referentes a citações de artigos (tanto quais artigos são citados por um artigo quanto quais artigos citam um artigo). Deste modo, seu uso como base para este trabalho tornou-se impossível.

A *Association for Computing Machinery Digital Library* (ACM-DL[1]) é autodenominada:

> "...uma vasta coleção de citações e texto completo de artigos d e periódicos e boletins de notícias da ACM , assim como publicações de conferências".[2]

[1] Conforme cálculos que serão demonstrados no Capítulo 3

Apesar de possuir todos os dados necessários, como citações feitas por um artigo, citações recebidas por um artigo, ano de publicação, veículo de publicação, entre outros, seu domínio é limitado a artigos publicados em veículos relacionados à *ACM*, restringindo-o de modo que podem haver distorções nas conclusões obtidas através destes dados.

Google Scholar

Por fim, uma solução encontrada foi utilizar uma máquina de busca vertical, o *Google Scholar* (GS), que foi posteriormente lançado também em uma versão brasileira, chamada *Google Acadêmico*. Trata-se de uma máquina de busca acessível gratuitamente na *Web*, que indexa o texto completo de publicações acadêmicas de formatos e disciplinas diversos. Lançado em Novembro de 2004, o GS inclui publicações ordenadas de acordo com citações recebidas, e disponibiliza o texto completo de publicações abertas, ou o *site* de compra de artigos pagos, como os da Elsevier, maior editora de artigos acadêmicos do mundo. O *Google Scholar* disponibiliza também diversos filtros para permitir maior precisão em sua busca, tais como o filtro que restringe a busca a nomes de autores. O sistema encontra-se em fase *Beta*, e por isso vem sofrendo diversas alterações no decorrer deste trabalho. Um grande problema do GS é sua política de não disponibilizar uma lista de abrangência de Periódicos e Conferências, assim como não prover informação sobre quão atualizado é seu acervo. Apesar disso, por seus dados abrangerem meta-dados do CiteSeer, DBLP, ACM-DL e as principais editoras da área da computação, assim como de outras áreas, seus dados foram considerados os mais próximos da realidade.

3.2 A Solução

Portanto, como o foco do trabalho é a análise dos principais artigos dos autores mais citados na área de Ciência da Computação, assim como quais artigos os citam, tornou-se inviável utilizar o CiteSeer, por sua instabilidade e pela dificuldade na obtenção de seus dados no momento inicial da pesquisa, e o DBLP, por não apresentar informações referentes às citações dos artigos indexados. Apesar de possuir informações referentes a citações, a biblioteca digital ACM DL não é uma alternativa para nosso estudo por indexar apenas as publicações que foram submetidas em veículos da *ACM*, excluindo publicações muito importantes, como as veiculadas nas conferências e periódicos da *Institute of Electrical and Electronics Engineers* (IEEE).

O Google Scholar, por sua vez, indexa todo o conteúdo atualizado do CiteSeer, da DBLP e das principais editoras, como Elsevier, e levando isso em consideração, seu conteúdo foi considerado como o mais vasto, sendo assim o mais próximo da realidade.

Para a utilização do Google Scholar, é necessária a submissão de consultas, para a obtenção dos resultados relevantes à sua consulta. Neste trabalho foi utilizada como chave desta consulta, uma lista contendo os 1.000 autores mais citados na área da Ciência da Computação, a ser descrita a seguir.

3.3 Compilação da lista de autores

Para efetuar a coleta utilizando o Google Scholar, devem ser feitas consultas objetivas para obter os resultados esperados. Neste trabalho, as consultas mais objetivas são obtidas quando procuramos pelo nome do autor, obtendo todas suas publicações. Essa necessidade leva à compilação de uma lista contendo os 1.000 autores que receberam a maior quantidade de citações em suas publicações.

Para a obtenção da lista de autores, recorremos à biblioteca CiteSeer. Porém, as referências a nomes de autores sofrem variações, tornando necessário utilizar métodos para normalizá-las. De acordo com experimentos de Newman [12], em bases não padronizadas em que um autor é mencionado de várias formas diferentes, a quantidade real de autores pode ser definida como um valor entre um limite superior, representado pelo total de diferentes nomes encontrados, e um limite inferior, obtido através do agrupamento de nomes que possuem a primeira inicial e o último sobrenome em comum. Uma das ocorrências encontradas no banco pode ser usada como exemplo: As ocorrências *J. D. Ullman*, *J Ullman*, *Jeffrey D. Ullman* e mesmo *J D Ullman*, para o limite superior, são considerados, cada ocorrência, como um autor diferente, portanto neste caso foram identificados 4 autores distintos. Para um limite inferior, todas as ocorrências são consideradas o mesmo autor, sendo este *J Ullman*.

Este cálculo, com números não disponibilizados pelo CiteSeer, não poderia ter sido realizado no início da pesquisa devido à indisponibilidade destes dados naquele momento, e após a possibilidade de download de seus dados através da *Web* de acordo com a *OAI* [17], com dados de Agosto/2006, foi calculado o limite inferior de 243.837 autores, e o limite superior de 411.032 autores.

O CiteSeer fornece uma lista com 10.000 autores mais citados. A forma em que os nomes informados nesta lista de autores está disponível consta apenas da inicial do primeiro nome e do último sobrenome, tornando possível inferir que a geração desta lista de 10.000 autores citados é baseada no limite inferior de quantidades de autores. Esta é a forma considerada mais adequada, pois reunirá vários artigos do mesmo autor que seriam considerados de outro autor, pela forma em que seu nome foi cadastrado. No entanto, esta forma gera o agrupamento de vários autores que possuem a mesma inicial do primeiro nome e o mesmo último sobrenome, mesmo sendo autores distintos.

Apesar destes autores não serem homônimos em seu sentido literal, no trabalho eles serão considerados como homônimos, pois são autores diferentes considerados como o mesmo autor por apresentar similaridade em seus nomes. Como exemplo deste fato, podemos apresentar o autor *D Johnson*, autor considerado o mais citado na área da Ciência da Computação pelo CiteSeer, porém apresenta 41 ocorrências para este nome, sendo que a maioria das ocorrências mostra que os autores são notoriamente diferentes, como *David B. Johnson*, *Don H. Johnson*, *Douglas Johnson*, *Damian Johnson*, *Deanne Johnson*, *Diana Johnson* e muitos outros.

Tal fato tornou necessária uma verificação manual desta lista, removendo ocorrências que apresentavam autores homônimos entre os resultados retornados pelo Google Scholar. Dessa forma, foram selecionados os 1.000 primeiros autores que estivessem de acordo com as seguintes restrições:

- A consulta, baseada no nome do autor (normalmente em sua forma abreviada de acordo com a forma utilizada em seus artigos, porém muitas vezes expandida para melhor identificação do mesmo) não pode retornar publicações relevantes nas quais não houve participação do autor em questão (consideradas publicações de autores homônimos);

- Os nomes referentes a indexação errada do CiteSeer, como P. Thesis (supostamente do autor Ph. D. Thesis, que notoriamente refere-se a erro na identificação de nome do autor) foram removidos;

- Pelo menos 1 artigo relevante do autor deve ser categorizado como publicação em periódico ou publicação em conferência.

- Foram considerados como artigos relevantes aqueles que obtiveram no mínimo 10 citações de acordo com o *Google Scholar*. Tal definição tornou-se necessária para minimizar o tamanho da coleta, visto que existe um grande número de publicações pouco citadas e o grande número de acessos ao servidor do Google Scholar necessário para coletar os artigos que as citam.

Desta forma foi concluída a geração da lista de consultas a serem submetidas ao Google Scholar a fim de coletar dados referentes aos artigos mais relevantes de cada um dos 1.000 autores selecionados. Esta seleção alcançou aproximadamente 2.000 autores da lista de autores mais citados, portanto aproximadamente 1.000 autores dentre os primeiros 2.000 mais citados não se apresentaram de acordo com as restrições citadas acima.

3.4 Coleta dos dados dos artigos

A partir desta lista de autores, tornou-se necessário, então, coletar e indexar as informações referentes aos artigos de cada um dos 1.000 autores selecionados.

Para isso foi criado um aplicativo, chamado Scholector, capaz de navegar automaticamente pelas páginas do Google Scholar, salvando suas páginas e organizando as ocorrências encontradas em um arquivo XML. Além de agir como um *crawler* para resultados do Google Scholar, o Scholector possui, ainda, funções para identificar cada uma das informações contida em cada registro das consultas retornadas pelo Google Scholar, e também a categorização dos veículos em que foram publicados os artigos identificados no processo.

Vários problemas foram encontrados para a realização da coleta. Há um limite imposto pelo Google Scholar em que a consulta retorna no máximo 1.000 resultados. Sendo assim, se um autor possui mais de 1.000 publicações, ou se um artigo possui mais de 1.000 citações, serão apresentados apenas os 1.000 artigos mais citados. Isso gerou a necessidade de efetuar uma poda na coleta, e como os resultados são apresentados de forma ordenada de acordo com a quantidade de citações recebidas pela publicação, a poda efetuada foi desconsiderar artigos que, segundo o Google Scholar, possuem menos de 10 citações. Este corte traz implicações no resultado, pois remove um grande número de publicações menos expressivas academicamente. Como este trabalho analisa apenas os autores mais citados na área da Ciência da Computação, este corte não torna-se brusco, pois desta forma estudamos apenas os artigos mais significativos destes autores, sendo possível deduzir qual é o meio mais notável que integra suas publicações.

Existe, ainda, um limite imposto pelo Google Scholar em que são permitidos apenas 1.000 acessos diários ao seu servidor, por IP. Devido à magnitude da coleta proposta, tornou-se necessário contornar esta limitação a partir de outros métodos. Houve uma tentativa de obter permissão de acesso diretamente à base de dados, porém, devido a dificuldades técnicas que seriam ao mesmo tempo demoradas e dispendiosas ao Google Scholar, esta liberação não pôde ser feita. Foi criada, então, uma política para utilização de proxies públicos com alta anonimidade, para fazer a coleta. Esta escolha foi feita pois, através do uso de um proxy, a requisição da página é feita para o Proxy, e este, por sua vez, efetua a nova requisição para o servidor da página a ser acessada. Desta forma, o Google Scholar recebe um acesso através do IP do proxy, e não do IP da máquina que está efetuando a coleta. A fim de não prejudicar a prestação de serviço dos servidores do Google Scholar com vários acessos simultâneos, foi feito um escalonamento de forma que apenas um proxy acessaria o Google Scholar por vez, como em uma fila *round robin*, ou seja, assim que um proxy termina seu acesso, há um pequeno tempo de espera e outro proxy inicia seu acesso, um a cada vez. Quando a lista de proxies chega ao

seu fim, a busca volta a realizar uma consulta utilizando o primeiro proxy, e continua percorrendo a lista seqüencialmente. Também houve a preocupação de não prejudicar a prestação de serviço dos servidores públicos de Proxy, limitando-se a quantidade de consultas em apenas 200 consultas diárias.

Além disso, o Google Scholar encontra-se em fase Beta, portanto, passou por várias modificações no decorrer do tempo. Isso tornou necessário retroceder alguns passos da coleta para terminá-la com sucesso, assim como modificações no *parser* do *Scholector*, pois alguns novos dados relevantes eram incluídos e havia a necessidade de identificar tais dados. Alguns dos problemas que implicaram em maior impacto na coleta foram:

- *Adição do agrupamento da mesma ocorrência de artigos em clusters.* Anteriormente, vários artigos eram exibidos várias vezes, o que denominamos como *splitted citations*. Este problema é de difícil solução, e como ele foi minimizado pelo Google de uma forma considerada satisfatória, foi considerado que haveria menor esforço em refazer a coleta do que implementar um outro tratamento de *splitted citations*.

- *Migração do* Google Scholar *para a versão brasileira,* Google Acadêmico. Tal mudança acarretou indisponibilidade do serviço por aproximadamente 1 mês (neste período o serviço era automaticamente direcionado para o servidor brasileiro, que ainda se encontrava indisponível) além de alterar, em pequena proporção, a exibição de seus resultados.

- *Adição da opção de baixar citações completas no formato BibTeX.* Apesar de se mostrar útil na solução de alguns problemas, como será descrito posteriormente, esta opção não foi utilizada por ser necessário fazer um novo acesso ao *Google Acadêmico*, e a quantidade de acessos diários foi um problema difícil de ser contornado.

- *Filtrar dados por área.* Foi adicionada, apenas no *Google Scholar*, a possibilidade de exibir somente resultados das áreas de Engenharia, Ciência da Computação e Matemática. Esta opção, que seria de muita ajuda, não foi utilizada pois a coleta já havia sido completamente finalizada quando esta opção se tornou disponível.

3.5 Extração dos dados obtidos a partir das consultas submetidas ao Google Scholar

O resultado das consultas realizadas é um conjunto de informações referente a cada artigo publicado pelo autor em questão. Estes resultados são ordenados por quantidade

de citações recebidas, e são compostos por:

- *Tipo de publicação.* Consta de uma categorização feita pelo *Google Acadêmico*, que categoriza a publicação como, por exemplo, um livro. No entanto, nem todos os livros possuem esta categorização.
- *Título.* Título da publicação.
- *Link para Publicação.* Link diretamente para uma versão eletrônica do artigo, ou, caso o artigo seja pago, para o *site* onde o artigo é vendido.
- *Agrupamento.* O *Google Acadêmico* agrupa ocorrências semelhantes que tratam possivelmente do mesmo artigo em 1 ocorrência, e disponibiliza uma outra página web com cada uma das ocorrências que foram omitidas em detalhes.
- *Autores.* Uma lista dos autores da publicação. Caso existam vários autores, um ou mais são ocultos; estes são substituídos por reticências para indicar que não são mostrados todos os autores.
- *Ano de Publicação.* Ano em que a publicação foi veiculada. Existem casos onde este dado não está disponível.
- *Veículo.* Nome da conferência, periódico, workshop, ou, em caso de livros, editora que veicula a publicação. Este dado não encontra-se presente em todos resultados, e em muitos casos possui palavras abreviadas, ou mesmo oculta algumas palavras, substituindo-as por reticências para indicar que foram ocultas.
- *Outras informações.* Outras informações sobre a procedência da publicação, como *site* referente ao veículo de publicação.
- *Resumo.* Um breve resumo, parágrafo ou descrição da publicação.
- *Citado por.* Informa quantas publicações referenciaram esta publicação, assim como um *link* para uma página que contem as publicações que a citam (limitando-se às 1.000 publicações mais citadas, caso seja citado por mais de 1.000 publicações).
- *BibTeX. Link* para informações padronizadas no formato BibTeX, contendo, de forma organizada, dados como título, nome dos autores, veículo, volume, páginas e ano de publicação. Estes dados não são resumidos, porém para obtê-los é necessário fazer um acesso extra ao servidor do Google Scholar.

- *Outros Links.* Além disso o *Google Acadêmico* disponibiliza *links* para outras páginas, como versão HTML da publicação, link para pesquisa na web utilizando o Google, serviço de Pesquisa em bibliotecas e artigos relacionados ao artigo em questão, porém nenhum desses links é utilizado, embora armazenado no banco de dados.

Figura 3.1: Exemplo de resultados retornados pelo *Google Acadêmico*

Para realizar as consultas automaticamente, assim como identificar cada informação obtida em uma página de resultados do *Google Scholar*, foi criado um programa capaz de submeter consultas à máquina de busca, assim como navegar pela página de resultados, ao mesmo tempo em que identifica e indexa cada parte do resultado.

Este programa, o qual foi dado o nome de *Scholector*, foi desenvolvido na linguagem C++, e utiliza a API do Microsoft Internet Explorer para navegar, de forma transparente, no *Google Acadêmico*, a fim de proporcionar, quando necessária, intervenção manual na etapa de coleta do algoritmo.

O funcionamento do *Scholector* se dá da seguinte forma:

1. Utilizando a lista de consultas, gerada previamente, o *Scholector* navega para a página de consultas do *Google Scholar* e submete a consulta.

2. A partir do documento *html* retornado, enquanto existirem resultados, cada campo é identificado e organizado um arquivo com estrutura *XML*, de forma seqüencial.

3. Enquanto houver páginas com resultados, navega para a próxima página e, para a nova página, segue os procedimentos descritos no passo anterior.

4. Enquanto houver consultas não submetidas, efetua todos os passos anteriores.

Em todos momentos em que há um novo acesso ao *Google Acadêmico*, existe, também, a opção de utilizar um proxy ou alterar o proxy atual, sendo esta troca automaticamente feita a partir de uma lista de proxies válidos pré-definidos. Esta opção foi utilizada no trabalho devido à restrição de 1.000 acessos diários ao servidor do *Google Scholar*, sendo necessário, portanto, obter novas máquinas para submeter consultas. Para que não houvesse bloqueio nos proxies públicos, cada proxy foi limitado a realizar apenas 200 consultas diárias, portanto o uso de 5 proxies duplica a capacidade de coleta diária. Outra ação feita para não haver bloqueio foi a substituição periódica dos *Cookies* utilizados pelo *Google Acadêmico* por um *Cookie* gerado quando ainda não haviam sido feitas consultas, porém já possuía configurações como idioma e quantidade máxima de resultados por página.

Dessa forma, é coletado o primeiro nível dos dados, chamados Publicações P. O segundo nível consiste em publicações que citam cada uma das publicações válidas, chamadas Citações C. Neste momento, a coleta das citações não é viável, pois ainda não foram identificadas quais publicações foram veiculadas em conferências ou periódicos da área de Ciência da Computação. Desta forma, a quantidade de acessos necessários é drasticamente reduzida, pois não serão coletados os citadores de publicações que não serão analisadas.

Após feita a categorização dos artigos de Publicações P, a coleta das Citações C será feita a partir do acesso aos *links* armazenados no registro *Citado Por*, que são coletados como descrito na coleta do nível 1, porém utilizando diretamente cada um destes *links*, ao invés de efetuar novas consultas. É feita a coleta de todas as Citações C de cada Publicação P que foi publicada em Conferência ou Periódico (identificados na categorização descrita no Capítulo na área de Ciência da Computação, desde que apresente um número mínimo de 10 citações, de acordo com o *Google Acadêmico*.

Com excessão da troca de consulta por palavra-chave pelo acesso direto via *links* fornecidos pelo *Google Acadêmico*, o processo de coleta de Citações C a serem coletadas seguirá da mesma forma que a coleta de Publicações P. Além disso não serão desconsideradas as Citações C que possuírem menos que 10 citações de acordo com o Google Acadêmico, pois esta tarefa não implicará em maiores custos no processo de coleta.

Capítulo 4

Categorização de veículos

A categorização de veículos, ou seja, a separação de veículos em seus respectivos tipos (ex.: Livros, Conferências, Periódicos, Workshops, Relatórios técnicos, entre outros) não é uma tarefa fácil, principalmente tendo em consideração que os resultados retornados pelo *Google Acadêmico* possuem artigos de várias fontes possíveis, sendo elas de todas as áreas de ciência. É necessário, portanto, identificar não apenas o tipo de veículo, mas também se este veículo é relacionado à área de Ciência da Computação ou a outra área.

Nos resultados do *Google Acadêmico* existem campos referentes ao veículo de publicação. Não existe, no entanto, informação referente à origem da publicação (se é uma conferência ou um periódico), nem se este veículo pertence à área da Ciência da Computação.

Isto tornou necessário categorizar os veículos informados. Não é necessário, no entanto identificar o veículo propriamente dito, mas simplesmente separá-lo entre conferências, periódicos e outros.

Para isso, foi utilizada como base de referência uma listagem disponibilizada pela *DBLP*, contendo todas as conferências e todos os periódicos na área da Ciência da Computação, separadamente. Esta base foi indexada e utilizada como base para a categorização realizada.

Além desta base, foram utilizadas quatro outras bases criadas empiricamente, sendo elas:

- *Abreviações a expandir.* Constitui de várias abreviações amplamente utilizadas, e suas respectivas formas expandidas. Como exemplo é possível transformar a palavra abreviada *conf.*, em sua forma expandida *conference.*

- *Indicadores de tipo.* Palavras ou frases que, por si só, já identificam o tipo do veículo, porém não identificam se o veículo é ou não da área da Ciência da

Computação. Como exemplo é possível correlacionar a palavra *transactions* como uma palavra que está sempre ligada a periódicos.

- *Indicadores de área.* Palavras ou frases que, por si só, já identificam a área, porém não identificam o tipo de veículo. Como exemplo podemos inferir que veículos que possuem a palavra *ACM* são da área de Ciência da Computação.

- *Siglas de veículos.* São siglas que, por si só, identificam um veículo da área da ciência da computação. Estas siglas estão presentes na listagem da DBLP e foram manualmente identificadas. As siglas que poderiam ser confundidas com palavras freqüentes, como *SEE* não foram incluídas nesta lista, porém cadastradas como sendo um nome completo de conferência, possibilitando ainda encontrar resultados que possuem como identificador apenas sua sigla.

A partir destas bases, é necessário criar uma metodologia capaz de categorizar eficientemente os veículos informados pelo Google Scholar.

Em [6] foi utilizado o Modelo Vetorial em um problema semelhante, obtendo grande precisão. No entanto, o problema identificado em [6] possuía um domínio fechado de publicações, enquanto neste trabalho podem existir casos não identificados, sendo estes publicações de outras áreas ou publicações não categorizadas na base utilizada como referência. O algoritmo vetorial retornará, então, várias respostas, sendo que em algumas vezes não haverá resposta válida, por exemplo, ao tentarmos categorizar o veículo "*British Dental Journal*", da área da Odontologia, existirão resultados possíveis de acordo com o modelo vetorial, devido à ocorrência da palavra Journal, no entanto, por não se tratar de um veículo da área da Ciência da Computação, não existe uma resposta correta à consulta, sendo toda resposta retornada considerada uma resposta inválida.

Para separar respostas válidas de inválidas, foi utilizado o Coeficiente de Jaccard. Este coeficiente foi utilizado de duas maneiras distintas. A primeira, quando não existem palavras ocultas no resultado do *Google Acadêmico*, e a segunda quando existem palavras ocultas em seu resultado. Desta forma, resultados que se possuem uma similaridade mínima dada pelo Coeficiente de Jaccard encontrado serão consideradas respostas válidas, e as que possuem similaridade abaixo da mínima aceita serão consideradas inválidas.

Para determinar um limite do coeficiente de Jaccard capaz de separar as respostas válidas das respostas inválidas, foi utilizada uma base de testes para determinar um valor de poda nos dois casos. Este teste consistia em seguir a categorização com intervenção humana, identificando como categorizado corretamente ou erroneamente. Ambos os valores corretos e errados do coeficiente de Jaccard foram armazenados, e

foi considerado como valor de poda o valor em que a maior quantidade de veículos foi categorizada corretamente, com a menor quantidade de veículos categorizada erroneamente. A partir de uma amostragem e teste de precisão, os valores 0.8 e 0.6 foram utilizados para veículo completo e veículo com omissão de palavras, respectivamente, identificando corretamente 98% dos veículos classificáveis e identificando aproximadamente 1% de veículos erroneamente, sendo que os últimos deveriam estar em outra categoria ou não ser classificado. O processo de Jaccard foi feito com algumas adaptações para melhor encaixe no problema, sendo estes:

- Remoção de palavras com duas ou menos letras, a fim de diminuir o impacto da remoção de artigos e outras palavras pequenas dos nomes das publicações que freqüentemente são removidas;
- Casamento de uma palavra com parte inicial de outra palavra é considerado como a mesma palavra, para diminuir o impacto de abreviaturas no cálculo do coeficiente de Jaccard;
- O Jaccard é calculado normalmente, sendo este definido pela interseção das palavras válidas dividida pela união das palavras válidas, de acordo com as considerações acima definidas.

O processo de categorização, segue, então, as seguintes etapas:

1. É feita a expansão das abreviações mais comuns, de acordo com a base pré-definida.
2. É feita uma busca pelos identificadores de tipo, a fim de limitar as buscas posteriores em ocorrências relacionadas ao tipo encontrado. Caso seja encontrado Livro ou Relatório técnico, o mesmo é categorizado com sucesso, não sendo necessário seguir nas próximas etapas.
3. É feita uma busca pelas siglas de veículos, restringindo apenas às siglas do veículo identificado no passo anterior, caso algum tipo tenha sido identificado. Caso seja identificada alguma sigla, o veículo é categorizado com sucesso, não sendo necessário seguir nas próximas etapas.
4. É feita uma consulta à base previamente compilada constituída por veículos da área da computação, obtida a partir dos veículos cadastrados na *DBLP*. Neste passo, a busca utilizada é baseada no modelo vetorial, tendo como resultado os veículos ordenados de acordo com a maior semelhança entre o veículo do artigo, fornecido no resultado do *Google Acadêmico*, com abreviações expandidas e o veículo cadastrado nesta base de veículos compilada.

5. Para cada resultado retornado, considerando restrições de tipo de veículo, caso exista alguma identificada no passo 2, é calculado o coeficiente de Jaccard, e caso algum resultado possua este coeficiente acima do limite mínimo identificado, este veículo é categorizado com sucesso, não sendo necessário seguir nas próximas etapas.

6. Caso exista uma categorização do tipo de veículo (etapa 2), é feita uma busca por palavras que identifiquem a área do veículo. Caso seja identificada, o veículo é categorizado como um veículo correspondente da área da computação. Caso não seja identificada, o veículo é categorizado como um veículo correspondente, porém de outra área.

7. Caso nenhum dos passos tenham identificado o veículo, o veículo é categorizado como *não identificado*.

8. Caso o campo esteja vazio, este processo não é realizado, e o veículo é categorizado como *sem veículo*.

Este processo de categorização é válido tanto para artigos dos autores mais citados (artigos do primeiro nível) quanto para os artigos que os citam (artigos do segundo nível). A coleta do segundo nível é feita apenas para artigos categorizados como conferências ou periódicos na área da Ciência da Computação.

As categorias utilizadas no ato de categorização foram:

1. Não Classificado (possui menos de 10 citações segundo Google Scholar
2. Conferência (casamento com Sigla)
3. Periódico (casamento com Sigla)
4. Conferência (casamento com Jaccard, veículo com nome completo)
5. Periódico (casamento com Jaccard, veículo com nome completo)
6. Conferência (casamento com Jaccard, veículo com palavras omitidas)
7. Periódico (casamento com Jaccard, veículo com palavras omitidas)
8. Conferência (de área diferente de Ciência da Computação)
9. Periódico (de área diferente de Ciência da Computação)
10. Livro (segundo Google Scholar)
11. Livro (segundo palavras-chave encontradas)

12. Vazio (Não apresentava texto no nome do veículo)
13. Não Identificado (não foi possível categorizar o texto)
14. Workshop (casamento com Sigla)
15. Workshop (casamento com Jaccard, veículo com nome completo)
16. Workshop (casamento com Jaccard, veículo com palavras omitidas)
17. Workshop (de área diferente de Ciência da Computação)
18. Relatório Técnico
19. Conferência (publicada pela ACM ou IEEE)
20. Periódico (publicado pela ACM ou IEEE)
21. Workshop (publicado pela ACM ou IEEE)

Estas categorias foram separadas desta forma para avaliar a precisão do algoritmo de categorização. No entanto não é necessário identificar a estratégia utilizada, mas somente os veículos, sendo então obtida a seguinte categorização:

1. Não Classificado
2. Conferência da área de Ciência da Computação
3. Periódico da área de Ciência da Computação
4. Workshop da área de Ciência da Computação
5. Livro
6. Relatório Técnico
7. Conferência de outras áreas
8. Periódico de outras áreas
9. Workshop de outras áreas
10. Vazio
11. Não Identificado

Capítulo 5

Análise das Publicações Coletadas

Este capítulo contém uma análise estatística dos dados coletados. Ao todo foram coletadas 51.941 páginas do *Google Scholar*, contendo dados referentes às publicações dos 1.000 autores mais citados na área da Ciência da Computação, segundo o CiteSeer, assim como as publicações que as citam. A seleção destes 1.000 autores, feita a partir do CiteSeer, considera todas as publicações do autor, independentemente do veículo de publicação.

As publicações coletadas estão divididas em dois grandes grupos. O primeiro será chamado *Publicações*, ou simplesmente P, sendo composto pelas publicações veiculadas em meios diversos de comunicação, que possuem participação direta de um ou mais autores selecionados. O segundo grupo será chamado *Citações*, ou simplesmente C, e contém as citações para as publicações do primeiro grupo que pertencem às categorias de Periódicos ou Conferências, ambos na área de Ciência da Computação.

É importante enfatizar de antemão que toda a análise apresentada neste capítulo não desconsidera auto-citações, ou seja, artigos relacionados nos grupos de Citações C que citam artigos do grupo Publicações P que possuem 1 ou mais autores em comum não são removidos. No entanto, dado um artigo pertencente ao grupo de Publicações P, se o mesmo foi encontrado em 2 ou mais autores dentre os 1.000 autores analisados, apenas uma ocorrência do mesmo artigo (a ocorrência do autor mais citado) foi considerada.

Nestes dois grupos, foi feita uma categorização das publicações, conforme descrito no Capítulo 4. Os "periódicos" são constituídos por *journals*, *transactions* e *magazines*. "Conferências" agrupam *conferences* e *symposiums*. Apesar de *workshops* serem muitas vezes agrupados com conferências, neste estudo eles não foram considerados como conferências, pois usualmente seus artigos diferem entre si, sendo que *Workshops* apresentam artigos normalmente menores, apesar de terem, também, sua importância. *Workshops*, assim como livros e relatórios técnicos identificados foram agrupados como

"outros". Há veículos cuja categorização não foi possível, seja por não apresentarem meta-informação sobre a publicação ou por não ter sido possível a identificação da categoria. Estes últimos foram agrupados como "não categorizados". Como o estudo é referente apenas à área da Ciência da Computação, *conferências* e *periódicos* foram separados entre periódicos da área e periódicos de outras áreas. Os números referentes a cada categoria podem ser vistos na Tabela 5.1.

Tipo de Publicação	*P*
Periódicos de Computação	12.570
Conferências de Computação	11.769
Periódicos/Conferências (Outras áreas)	9.112
Workshop de Computação	588
Workshop (Outras Áreas)	1.284
Livros	4.277
Relatórios Técnicos	177
Veículo vazio	4.864
Veículo não identificado	23.922
Total	**68.563**

Tipo de Citação	*C*
Periódicos de Computação	301.561
Conferências de Computação	302.450
Periódicos/Conferências (Outras áreas)	178.493
Workshop de Computação	19.791
Workshop (Outras Áreas)	49.407
Livros	43.562
Relatórios Técnicos	3.564
Veículo vazio	556.815
Veículo não identificado	363.615
Total	**1.819.258**

Tabela 5.1: Totais de publicações P dos 1.000 autores mais citados no CiteSeer, bem como o número específico de citações C, separados por categoria.

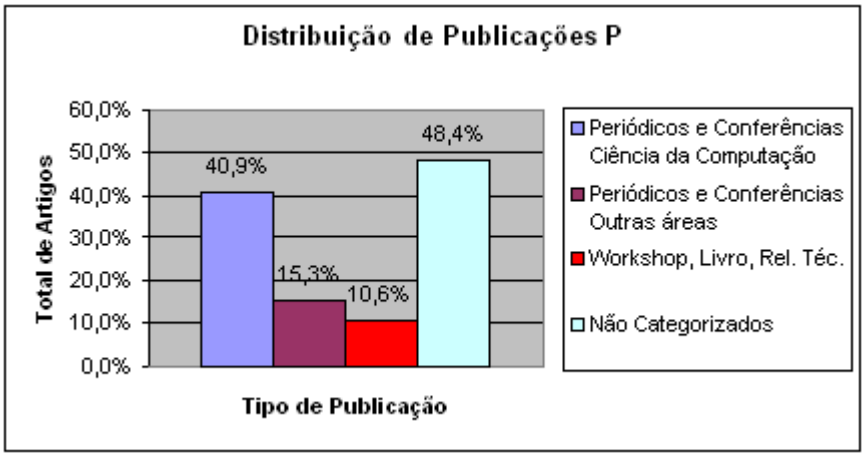

Figura 5.1: Publicações P separadas por seus tipos.

As Figuras 5.1 e 5.2 separam as publicações P e citações C, respectivamente, de acordo com os tipos das mesmas e suas respectivas áreas (Ciência da Computação ou Outras Áreas), para toda a coleta realizada. Como podemos observar, 40,9% das publicações P são periódicos ou conferências da área de Ciência da Computação, 15,3% são de outras áreas, 10,6% são "*Workshops*, livros e relatórios técnicos" e 48,4% não puderam ser categorizados por falta de informação. Com relação às citações C podemos observar que 33,2% das citações correspondem a periódicos e conferências da área da

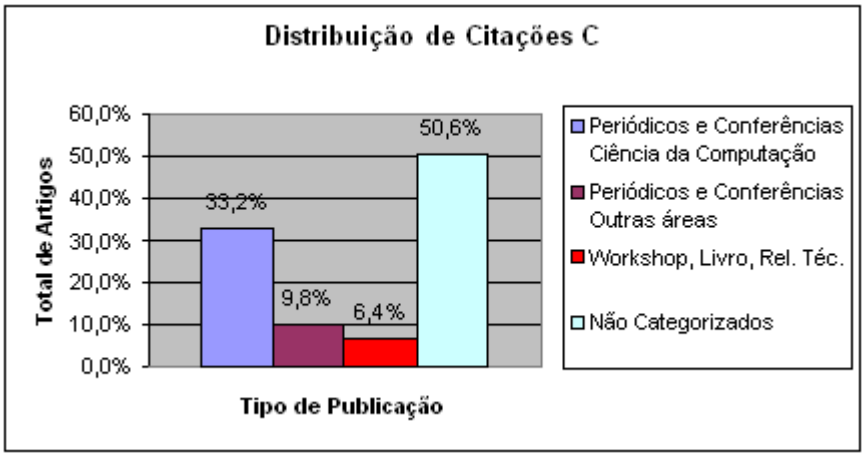

Figura 5.2: Citações C separadas por seus tipos.

Ciência da Computação, 9,8% são de outras áreas, 6,4% são "*Workshops*, livros e relatórios técnicos" e 50,6% não puderam ser categorizados por falta de informação.

É importante observar que estamos utilizando uma biblioteca digital de coleta automática, e que nela existem entradas duplicadas para as mesmas publicações, porém com alguns dados, como título, diferentes. Isso ocorre devido à falta de padronização em citações, gerando múltiplas entradas para a mesma publicação; títulos ou nomes de autores corrompidos; artigos de áreas que não são do interesse da pesquisa; ou mesmo artigos que foram veiculados apenas em páginas *Web*; o que torna, como resultado, impossível identificá-los. Como resultado, 7% das publicações P não possuem informação referente ao veículo de origem, pelo *Google Scholar*, e 34% possuem referência que não foi possível identificar. Como dito anteriormente, o objetivo deste estudo, nesse momento, é identificar apenas publicações na área da computação, e dada a diversidade de áreas encontradas em nossos resultados, é razoável aceitar que esta proporção não seja identificada como da área de estudo, ou mesmo possua pequenas falhas.

Dado o tamanho e a forma em que a coleta foi realizada, podemos afirmar que a identificação de periódicos na área da Ciência da Computação demonstra ser boa, pois a forma de citações das mesmas em artigos acadêmicos apresenta baixa variação, e a pequena taxa de de ruído apresentada torna possível obter maior precisão. Para conferências, no entanto, foi verificada uma variabilidade maior nas formas de citação das mesmas, o que inviabilizou uma precisão tão grande quanto a obtida em periódicos. Em várias ocorrências são apresentadas abreviaturas que muitas vezes torna difícil identificar sua precedência até mesmo manualmente. Para aumentar esta precisão, foram feitas várias análises e intervenções manuais a fim de identificar novos padrões de citações de conferências, e grande parte de nomes de veículos não identificáveis por

nossa base de conferências foi realimentada, de modo a possibilitar novas reclassificações automáticas, com menos erros e maior precisão. Temos consciência de que existem tanto artigos de periódicos quanto artigos de conferências que são relevantes ao nosso estudo e não foram categorizados, no entanto, devido à natureza aleatória das formas não identificadas, não existem evidências de qualquer tendência nos resultados. Além disso foram encontradas algumas publicações com cadastro errado de veículo, diferindo da realidade, o que não pôde ser tratado devido às limitações em analisar e identificar manualmente cada artigo.

É importante notar, também, que as citações C referem-se apenas a citações de artigos que foram identificados em conferências e periódicos, ambos na área da Ciência da Computação. Restringindo a citações apenas aos artigos dos veículos em estudo foi possível limitar o espaço de coleta, diminuindo o esforço despendido na mesma, assim como focar nas citações que trazem informação útil ao nosso estudo, ao invés de dispersar a atenção em citações que não dizem respeito ao trabalho.

	Total	**Perc.(%)**
Publicações de Periódicos	12.570	52%
Publicações de Conferências	11.769	48%
Total de publicações	24.339	
Citações de Periódicos	332.498	39%
Citações de Conferências	411.918	48%
Citações de Livros, Workshops e Relatórios Técnicos	109.960	13%
Total de citações	854.376	

Tabela 5.2: Total de publicações P e citações C consideradas em nosso estudo.

A Tabela 5.2 mostra o total de publicações P e citações C consideradas em nosso estudo. Para as publicações P foram consideradas apenas publicações de Periódicos e Conferências na Área da Ciência da Computação, como descrito anteriormente. A quantidade de publicações P veiculadas em periódicos é pouco maior que a quantidade de publicações veiculadas em conferências (52% de periódicos e 48% de conferências), e de acordo com o erro assumido estatisticamente, esta proximidade de valores identificaos como iguais. Para citações C foram considerados não somente periódicos e conferências na área da Ciência da Computação, mas também em outras áreas, pois partimos do pressuposto que para avaliar a importância do artigo devemos levar em consideração também a relevância do mesmo em outras áreas de pesquisa. Além disso foram consideradas relevantes citações oriundas de Workshops, Livros e Relatórios Técnicos. Desse modo, há um número maior de citações vindas de conferências com diferença significativa, sendo 39% vindas de periódicos contra 48% vindas de conferências, e o grupo de citações feitas por workshops, livros e relatórios técnicos (também sem limitar

a área) totaliza 13% das citações.

	Total de Citações oriundas de:			
	Periódicos	Conferências	Livros, Rel. Tec. e Workshops	Totais
para Publicações em Periódicos	219.120	224.829	61.647	505.596
para Publicações em Conferências	113.378	187.089	48.313	348.780
Total de Citações	332.498	411.918	109.960	854.376

Tabela 5.3: Número total de citações por publicação para artigos em conferências e em periódicos.

	Média de Citações oriundas de:			
	Periódicos	Conferências	Livros, Rel. Tec. e Workshops	Totais
para Publicações em Periódicos	17,4	17,9	4,9	40,2
para Publicações em Conferências	9,6	15,9	4,1	29,6
Razão entre as médias (Per/Conf)	1,8	1,1	1,2	1,4

Tabela 5.4: Número médio de citações por publicação para artigos em conferências e em periódicos.

As Tabelas 5.3 e 5.4 mostram a relação entre as publicações P e as citações C, destacando para as Publicações P classificadas como Conferências e Periódicos, de quais veículos vêm suas Citações C. Verifica-se que 59% das Citações C referem-se a Publicações P veiculadas em periódicos, enquanto 41% de Citações C referem-se a Publicações P veiculadas em conferências. Nota-se também que as Publicações P veiculadas em periódicos recebem um valor bem próximo de Citações C oriundas de conferências e de periódicos, enquanto as Publicações P publicadas em conferências recebem em sua maioria Citações C veiculadas em conferências. Analisando a razão das médias, Apenas em relação à média das Citações C vindas de Periódicos são predominantemente maiores, enquanto a média de Citações C vindas de Conferências, Livros, Workshops e Relatórios Técnicos é próxima para Publicações P de periódicos e conferências.

Agrupando publicações de acordo com a quantidade de citações recebidas temos a Tabela 5.5. É possível, mais uma vez, verificar que enquanto aproximadamente 79% das publicações de periódicos possuem menos de 50 citações, aproximadamente 82% das publicações de conferências estão nessa situação. Portanto, tendo em consideração os números absolutos, existem mais publicações menos citadas em conferências do que em periódicos, e portanto existem mais periódicos com maior número de citações do que

Publicações separadas por número de citações recebidas						
	< 50	50 a 199	200 a 349	350 a 499	>= 500	Total
Nº. de publicações em Periódicos	9.901	2.300	230	85	54	12.570
Nº. de publicações em Conferências	10.054	1.560	102	35	18	11.769
Total de publicações	19.955	3.860	332	120	72	24.339

Tabela 5.5: Total de publicações, em conferências e periódicos, separadas por número de citações recebidas.

conferências, embora estes últimos não alcancem centenas de artigos, para os últimos agrupamentos.

Número médio de citações recebidas por publicação					
	< 50	50 a 199	200 a 349	350 a 499	>= 500
Publicações em Periódicos	17,0	92,1	254,4	404,0	609,5
Publicações em Conferências	15,9	87,2	264,2	408,9	625,7

Tabela 5.6: Médias agrupadas de citações por publicação.

No entanto, se obtivermos valores médios, como demonstrados na Tabela 5.6, podemos verificar que publicações de conferências apresentam valores médios ligeiramente maiores para publicações com 200 ou mais citações recebidas.

Figura 5.3: Distribuição das citações recebidas por publicação.

Traçando um gráfico das citações, como pode ser visualizado na Figura 5.3, podemos notar que a curva de citações para publicações de conferências tem uma queda mais rápida que a curva de citações para publicações de periódicos. Tal distribuição mostra como em números absolutos a predominância de citações favorece periódicos,

embora a cauda dos periódicos seja maior que a de conferências. Nota-se que existe um número muito baixo de publicações com grande número de citações recebidas tanto para periódicos quanto para conferências, e um número grande de publicações pouco citadas, mesmo tendo em consideração os principais autores da área da Ciência da Computação.

Em termos gerais, portanto, os artigos publicados em periódicos apresentaram maior número de citações recebidas em relação aos artigos publicados em conferências, mesmo que as últimas possuam em média um valor maior de publicações bem citadas. No entanto, é fato que o número de citações recebidas por um artigo tende a aumentar de acordo com o passar do tempo. Sendo assim, compararmos apenas artigos de conferências e artigos de periódicos desconsiderando o tempo de publicação do artigo não permite uma análise em detalhe.

Total de Publicações em grupos de 5 anos									
	N/D	<= 1980	81-85	86-90	91-95	96-00	01-05	06	Total
Publicações em Periódicos	94	1.077	1.030	1.978	2.977	3.404	2.005	5	12.570
Publicações em Conferências	485	301	467	1.274	2.659	4.031	2.551	1	11.769
Total de Publicações	579	1.378	1.497	3.252	5.641	7.445	4.572	12	24.339

Tabela 5.7: Total de publicações agrupadas a cada 5 anos.

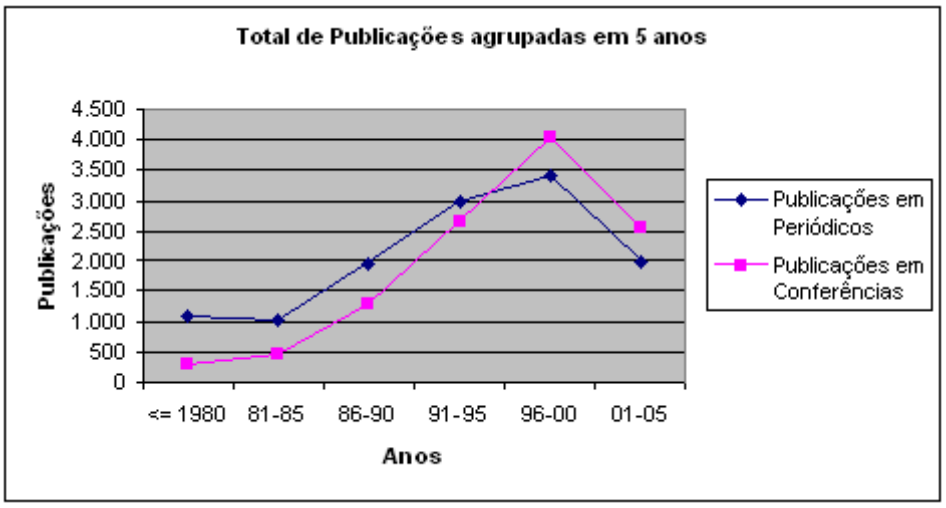

Figura 5.4: Distribuição das Publicações P agrupadas a cada 5 anos.

Desta forma, analisamos os artigos distribuídos em relação ao tempo de publicação. Para tal análise é necessário validar as datas das publicações. Publicações que não possuíam data de publicação, de acordo com o *Google Scholar* assim como Citações C que foram publicadas anteriormente às Publicações P as quais elas citam foram removidas desta análise, separadas nas categorias N/D das análises respectivas. Além

disso, Publicações P que não possuíam Citações C com data válida, por não possuírem citações, também foram classificadas em N/D. Foram feitas análises anuais e verificados agrupamentos de 2, 5 e 10 anos, por todo o período de coleta, e foi identificado que o agrupamento de 5 em 5 anos é o que melhor representa as mudanças identificadas. Além disso, devido à baixa representatividade de Conferências anteriores ao ano de 1981, os artigos referentes a este período foram agrupados em um só grupo, e os artigos publicados a partir de 1981 foram agrupados conforme definição acima. A Tabela 5.7 apresenta a distribuição encontrada para as Publicações P. Os valores referentes a N/D e *06* encontram-se apenas para apresentação de todos os valores, sendo N/D publicações de acordo com a definição acima e *06* as publicações do ano de 2006, e assim evitar divergências com os números exibidos nos gráficos anteriores, o que poderia causar a errada impressão de que há erros nos números exibidos. Estes grupos não serão analisados, pois um trata de inconsistências na base e outro de publicações muito recentes para representarem impacto suficiente para análise. Nota-se que até o ano de 1990 há predominantemente publicações em periódicos, e baixa presença de conferências. No entanto, a partir de 1991 há um grande salto na quantidade de publicações de conferências, chegando a apresentar um número bem próximo de publicações em conferências e em periódicos. A partir de 1996 é evidente a predominância de Publicações P de conferências entre os artigos mais citados dos autores mais citados na área da Ciência da Computação. A Figura 5.4 mostra mais claramente a distribuição destas publicações em cada agrupamento de 5 anos. Nela é evidenciada a superação numérica de publicações em conferências a partir de meados dos anos 90. Esta ampliação da abrangência de publicações em conferências, não apenas em um único pico, mas se mantendo com maior número de publicações a partir de então, demonstra a força obtida por este veículo. Anteriormente é possível notar que a veiculação de artigos concentrava-se com maior número em periódicos, e a partir do momento que inicia a popularização da Internet, também é iniciada a popularização das conferências, na área da Ciência da Computação. Este fenômeno mostra o retrato desta mudança, caracterizada pela necessidade de obter um meio capaz de divulgar a produção científica com a vazão demandada pela área da Ciência da Computação, que apresenta inovações de maneira tão rápida. É importante verificar que a queda visualizada no agrupamento 01-05 é relacionada ao fato de que as publicações ainda são recentes, tendo recebido ainda baixo número de citações, sendo também possível que muitas delas tenham sido descartadas pelo filtro inicial, por apresentarem menos de 10 citações de acordo com o *Google Scholar*.

Na Tabela 5.8 são apresentadas as Citações C referentes às Publicações P agrupadas como anteriormente. Nela pode-se perceber diferença semelhante para a quantidade de Citações C recebidas para as Publicações P dos principais autores da área da Ci-

ência da Computação. Publicações P veiculadas em periódicos no período anterior a 1980 receberam mais de 6 vezes mais Citações C que Publicações P veiculadas em conferências no mesmo período. Porém também torna-se claro que a partir de 1996 as Publicações P destes autores foram citadas de forma semelhante tanto em conferências quanto em periódicos, apresentando, inclusive, uma inversão de posições, tendo então as conferências superado o número de Citações obtidas por periódicos. Mesmo nos últimos 5 anos analisados, embora não houvesse tempo para acumular um alto número de citações, foram encontradas mais Citações C para Publicações P de conferências do que em Publicações P de periódicos. Essa distribuição pode ser melhor visualizada na Figura 5.5. Complementando estes dados com os valores referentes às médias de citações recebidas para os mesmos períodos, verificadas na tabela 5.9, vemos que a média de Citações C para conferências se aproxima da média de Citações adquiridas por periódicos publicados na mesma época, sendo que no último período são praticamente iguais. Portanto, na Tabela 5.8 tornou-se claro que a razão da média de Citações adquiridas por periódicos pela média de Citações adquiridas por conferências está diminuindo, o que demonstra o crescimento constante da importância de conferências na área de Ciência da Computação. Este crescimento gradativo da média de citações para conferências e conseqüente diminuição entre as diferenças da importância de publicações de conferências e de periódicos podem ser melhor visualizados na Figura 5.6.

Total de Citações em grupos de 5 anos								
	N/D	<= 1980	81-85	86-90	91-95	96-00	01-05	06	Total
para Publicações em Periódicos	61.092	52.164	42.371	86.683	114.586	113.009	35.686	5	505.596
para Publicações em Conferências	44.521	8.547	13.324	39.588	77.748	120.136	44.914	2	348.780

Tabela 5.8: Total de Citações (C) recebidas por Publicações (P) agrupadas a cada 5 anos

Média de Citações em grupos de 5 anos						
	<= 1980	81-85	86-90	91-95	96-00	01-05
para Publicações em Periódicos	48,43	41,14	43,82	38,49	33,20	17,80
para Publicações em Conferências	28,40	28,53	31,07	29,24	29,80	17,61
Razão entre as médias (Per/Conf)	1,7	1,4	1,4	1,3	1,1	1,0

Tabela 5.9: Médias de Citações C recebidas por Publicações P recebidas agrupadas a cada 5 anos

Além de agrupar as publicações de acordo com o ano de publicação, obtendo o crescimento de periódicos e conferências ao longo do tempo, é relevante analisar o impacto das publicações de cada veículo analisado independente da época de publicação, mas com o decorrer do tempo. Deste modo é possível verificar por quanto tempo uma publicação de conferência ou de periódico continua sendo citada. Para isso foi mantida

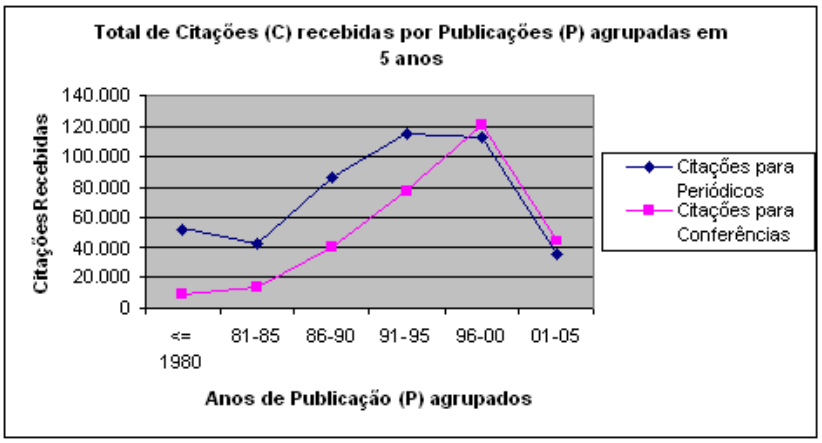

Figura 5.5: Distribuição do total de citações recebidas, agrupadas por ano de publicação em grupos de 5 anos

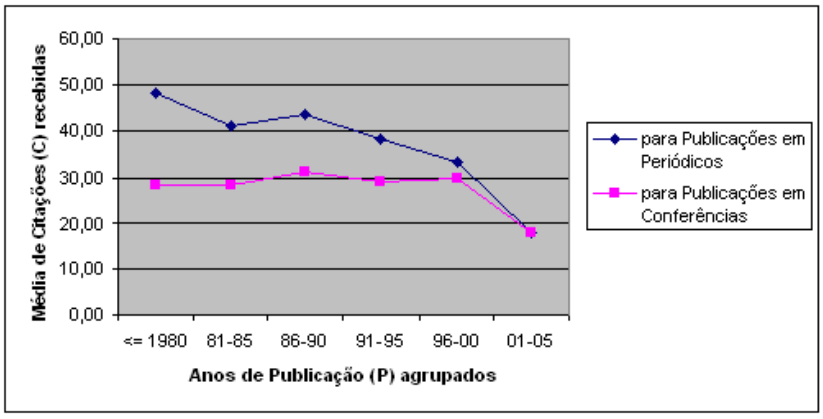

Figura 5.6: Distribuição da média de citações recebidas, agrupadas por ano de publicação em grupos de 5 anos

	Total de Citações C em tempo após publicação							
	N/D	<= 5	6 a 10	11 a 15	16 a 20	21 a 25	> 25	Total
Periódicos	61.092	196.588	126.826	62.848	28.376	15.142	14.724	505.596
Conferências	44.521	187.138	81.551	23.551	7.621	2.866	1.532	348.780

Tabela 5.10: Total de citações ao longo do tempo.

a janela de 5 anos utilizada nos outros agrupamentos, o que torna, por exemplo, o primeiro agrupamento a reunião das citações recebidas por um artigo a partir do momento de sua publicação até o quinto ano após publicado, o segundo agrupamento a reunião das citações recebidas por um artigo a partir de 6 anos de publicação até o décimo ano após publicado, sucessivamente. Estes valores são encontrados nas Tabelas 5.10 e 5.11, onde podemos observar que nos 5 primeiros anos de publicação artigos publicados em

	Média de Citações C em tempo após publicação					
	<= 5	6 a 10	11 a 15	16 a 20	21 a 25	> 25
Periódicos	15,7	12,1	8,9	6,9	7,2	13,7
Conferências	16,6	9,3	5,0	3,7	3,7	5,1
	0,9	1,3	1,8	1,9	1,9	2,7

Tabela 5.11: Média de citações ao longo do tempo.

conferências e periódicos mantém aproximadamente a mesma quantidade de citações, levemente superior em números absolutos para periódicos, e levemente superior para conferências em média de citações. No entanto, à medida que o tempo passa, periódicos tendem a ampliar sua vantagem de citações em relação às publicações de conferências, tanto no valor absoluto quanto na média. Nota-se claramente como a quantidade média de citações cai para artigos veiculados em conferências a partir, principalmente da primeira década (grupo de 11 a 15 anos e posteriores). Este fato vai ao encontro dos dados obtidos na Figura 5.4, que mostram que apenas a partir de meados da década de 90 que as conferências passam a ter importância semelhante à de periódicos. Desta forma, a quantidade de publicações que recebem quantidades expressivas de citações está limitada aos seus 10 primeiros anos de publicação. Podemos concluir, portanto, que periódicos possuem maior tradição, porém conferências vêm conseguindo espaço, principalmente após a expansão da Internet, o que tornou artigos de conferências mais acessíveis, e com isso mais citados, principalmente no período imediatamente posterior à sua veiculação (primeiros 5 anos).

A partir deste estudo é possível concluir, então, que há maior tradição em periódicos, mostrando-se superiores a conferências antes da década de 1990. Contudo, a partir de 1990, a participação de conferências na divulgação de conhecimento na área de Ciência da Computação se ampliou. O fato de outras áreas científicas valorizarem apenas periódicos e publicações veiculadas em conferências serem pouco valorizadas não se reflete na área da Ciência da Computação. No entanto, por existir tal tradição, é possível verificar que muitas publicações da área da Ciência da Computação são primeiramente veiculadas em conferências, mas por serem menos valorizadas pelas agências financiadoras de pesquisa, é muito comum existirem versões revisadas das melhores publicações de conferências publicadas em periódicos. Isso implica que estes artigos dividirão suas citações em um curto período de tempo, mas com o tempo as citações tenderão a apontar o artigo publicado no periódico, primeiramente porque a nova versão agrega novos resultados, e em segundo lugar porque há um certo interesse em publicações de periódicos terem um alto número de citações, também para periódicos, como constatado nos primeiros resultados da pesquisa. Todavia a importância das conferências na área é notável, principalmente como veículo de rápida difusão de novas idéias, dando a vazão necessária à Ciência da Computação, que tem evolução

singular, superando a taxa de evolução de outras áreas que já possuem sua base mais sólida por existirem a mais tempo. Dessa forma, podemos concluir que atualmente conferências atingem seu público mais rapidamente que periódicos por levarem menos tempo de submissão, sem com isso perder em citações recebidas quando comparamos com as citações recebidas por publicações submetidas em periódicos na mesma época. No entanto em relação ao tempo de vida de publicações em conferências, ou seja, o tempo que as mesmas continuam sendo citadas, ainda é cedo para afirmar que é tão grande quanto o de periódicos, dado que o fenômeno do crescimento da aceitação de conferências é ainda recente, e pode ser limitado pelo menor peso dado a elas vindo das agências financiadoras de pesquisa.

Capítulo 6

Conclusão e trabalhos futuros

Foi realizada uma coleta referente aos 1.000 autores mais citados na área de Ciência da Computação, segundo o CiteSeer. A partir destes autores, foi feita uma análise da forma em que as publicações desta área são publicadas. Foram avaliados os veículos de conferências e os veículos de periódicos. Como esperado, a partir da última década, com a facilidade de distribuição de artigos acadêmicos veiculados tanto em artigos de conferências quanto periódicos, os primeiros tornaram-se mais facilmente acessíveis, visto que anteriormente artigos veiculados em periódicos predominavam em bibliotecas, fato este que legitima periódicos como principal veículo para publicação de artigos. No entanto, como foi constatado, conferências vêm ganhando espaço a partir da última década, sendo um meio amplamente utilizado pelos principais pesquisadores da área de Ciência da Computação, tendo, inclusive, a característica de divulgarem mais rapidamente seus resultados.

Com este trabalho, foi possível verificar que os autores mais prolíficos na área publicam eqüitativamente em conferências e periódicos. Ademais, o número de citações por publicação tem crescido mais rapidamente para artigos publicados em conferências. Em números absolutos, os periódicos apresentam maior número de artigos muito citados, porém este fato é motivado por sua tradição, e em alguns momentos também devido ao jogo do fator de impacto, discutido na revisão bibliográfica. Foi também constatada uma mudança de comportamento dos autores a favor das conferências, tornando-as mais visíveis e até mesmo mais confiáveis, mas esta mudança não aparenta ter abalado a tradição dos periódicos.

Esta tradição dos periódicos, que os mantém recebendo citações de forma crescente com o decorrer do tempo, ainda não pôde ser constatado em conferências, pois sua visibilidade crescente consta apenas da última década. No entanto, as tendências temporais indicam que ambos veículos apresentam influência semelhante, e a tendência é de conferências se tornarem mais importantes. Sua importância, portanto, mostrou

não ser limitada como em outras áreas, mas tão grande quanto à de periódicos da área em estudo.

Este trabalho tem como foco apenas a análise do topo de uma distribuição que se assemelha a uma distribuição logarítmica, o que torna este estudo válido para os principais autores. Como trabalho futuro, existe o objetivo de fazer uma análise mais aprofundada da cauda dessa lista, constituindo-se de autores com média e baixa quantidade de citações em artigos veiculados em periódicos e conferências. Esta coleta é muito extensa e demanda tempo tanto para preparação da coleta (separação de autores homônimos) quanto para a coleta em si, com suas atuais limitações. Unindo estas duas análises é possível obter com precisão a influência relativa a periódicos e conferências. Além disso, é interessante manter uma análise do tempo de vida de publicações de conferências, dado que a grande visibilidade de conferências na área é ainda recente, e ainda há pouco a se analisar por este ponto de vista. Relacionados à coleta de dados, existe a necessidade de se estabelecer uma maneira de vencer a barreira da quantidade de artigos sem perder na qualidade dos mesmos, problema este que vem sido amplamente abordado na área de Bibliotecas Digitais. Finalmente, com uma base de dados mais homogênea e sem omissão de dados de coleta (como nomes de alguns autores), poderia ser refeita a análise desconsiderando-se auto-citações, problema não abordado nesse estudo devido à falta de todos os nomes dos autores para cada artigo, fato este que tornaria a análise tendenciosa ao tirarmos auto-citações apenas dos autores cujos nomes estavam relacionados e não removendo as auto-citações de autores que tiveram seus nomes omitidos.

Referências Bibliográficas

[1] ACM. The acm digital library. disponível através do endereço *Web* http://portal.acm.org/dl.cfm, 10 2006.

[2] ACM. The acm digital library faqs - frequently asked questions. disponível através do endereço *Web* http://portal.acm.org/faq_dl.cfm#1, 10 2006.

[3] Ricardo A. Baeza-Yates and Berthier A. Ribeiro-Neto. *Modern Information Retrieval*. ACM Press / Addison-Wesley, 1999.

[4] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual web search engine. *Proceedings of the Seventh World Wide Web Conference/ Computer Networks*, 30(1-7):107–117, 1998.

[5] S. Chakrabarti. *Mining the Web: Discovering Knowledge from Hypertext Data*. Morgan Kaufmann, 2003.

[6] William W. Cohen, Pradeep Ravikumar, and Stephen E. Fienberg. A comparison of string distance metrics for name-matching tasks. In Subbarao Kambhampati and Craig A. Knoblock, editors, *IIWeb*, pages 73–78, 2003.

[7] Editorial. Not-so-deep impact. *Nature*, 435:1003–1004, 2005.

[8] The PLoS Editors. The impact factor game. It is time to find a better way to assess the scientific literature. *PLoS Medicine*, 3(6):e291, 2006.

[9] F. Hecht, BK Hecht, and AA Sandberg. The journal"impact factor": a misnamed, misleading, misused measure. *Cancer Genet Cytogenet*, 104(2):77–81, 1998.

[10] S. Lawrence. Online or Invisible? *Nature*, 411:521, 2001.

[11] Michael Ley. Dblp - computer science bibliography. disponível através do endereço *Web* http://www.informatik.uni-trier.de/ ley/db/, 10 2006.

[12] MEJ Newman. The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences*, 98(2):404–409, 2001.

[13] B. Rousseau and R. Rousseau. LOTKA: A program to fit a power law distribution to observed frequency data. *Cybermetrics*, 4(4), 2000.

[14] P.O. Seglen. Why the impact factor of journals should not be used for evaluating research, 1997.

[15] Lee Giles Steve Lawrence, Kurt Ballacker. Estimated impact of publication venues in computer science. disponível através do endereço *Web* http://citeseer.ist.psu.edu/oai.html, 05 2003.

[16] Lee Giles Steve Lawrence, Kurt Ballacker. About citeseer. disponível através do endereço *Web* http://citeseer.ist.psu.edu/citeseer.html, 10 2006.

[17] Lee Giles Steve Lawrence, Kurt Ballacker. Citeseer oai records. disponível através do endereço *Web* http://citeseer.ist.psu.edu/impact.html, 11 2006.

[18] Lee Giles Steve Lawrence, Kurt Ballacker. Citeseer.ist - scientific literature digital library. disponível através do endereço *Web* http://citeseer.ist.psu.edu/, 10 2006.