SHINE: A Simple HIndex Estimator

David Fernandes de Oliveira, david@icomp.ufam.edu.br, Instituto de Computação/UFAM.
Altigran Soares da Silva, alti@icomp.ufam.edu.br, Instituto de Computação/UFAM.
Henrique Cavalcante, hpc@icomp.ufam.edu.br, Instituto de Computação/UFAM.

Neste artigo apresentamos o projeto SHINE, que tem por finalidade disponibilizar referenciais sobre a qualidade e o impacto de conferências para a comunidade de Computação. Os referenciais de impacto são estimados através da métrica índice-H, que é baseada no número de citações das publicações dos veículos. Adotando a metodologia apresentada neste artigo, o projeto SHINE resultou em uma ferramenta online capaz de calcular o índice-H de 1.788 conferências, tornando-se uma importante fonte de informações sobre a qualidade desses veículos para a comunidade de Computação.

Uma das preocupações mais presentes entre os pesquisadores das diversas áreas científicas é publicar os resultados de seus trabalhos em veículos (periódicos, simpósios, conferências, etc) de grande visibilidade e impacto na comunidade científica mundial. Em geral, os índices de impacto de determinado veículo são obtidos através de métricas baseadas no número de vezes que as publicações desse veículo são citadas por outras publicações. Exemplos de tais métricas são o Fator de Impacto (FI), criado por Eugene Garfield em 1955, e o índice-H.

Para a grande maioria das áreas, é possível obter os índices de impacto de seus principais veículos através de institutos tais como o *SCImago* (www.scimagojr.com) e o *Thomson Reuters* (thomsonreuters.com). Esse último, por exemplo, estima o impacto dos veículos através da métrica FI. Desde 1972, os FIs são calculados anualmente para os periódicos indexados na base de dados *Web of Science* e depois publicados no *Journal of Citation Reportes* (JCR) da *Thomson Reuters*. Dado seu enorme prestígio perante a comunidade científica mundial, o JCR é utilizado por vários comitês de área da CAPES para avaliar a qualidade da produção intelectual dos programas de pós-graduação brasileiros nestas áreas.

O *SCImago* e o *Thomson Reuters*, bem como outros institutos similares, reportam apenas índices de impacto de periódicos, o que é suficiente para a grande maioria das áreas científicas. No entanto, a área de Ciência da Computação se diferencia da maioria das demais por considerar que as conferências são veículos igualmente válidos para a publicação de trabalhos científicos. Muitas vezes, na área de Computação, os artigos em periódicos são apenas versões mais aprofundadas de artigos já publicados em conferências. Não havendo um instituto ou ferramenta para reportar referenciais de qualidade sobre as conferências, sobretudo no caso de conferências nacionais, o Instituto de Computação da UFAM, em parceria com a SBC, criou o projeto SHINE (**shine.icomp.ufam.edu.br**), que objetiva estimar e apresentar índices de impacto das principais conferências da área de Ciência da Computação através da métrica índice-H.

O projeto SHINE faz parte de uma iniciativa chamada Perfil-CC SBC, que foi inspirada na iniciativa Perfil-CC realizada em 2007 pela PUC-Rio, UFMG, COPPE/UFRJ, UNICAMP e ICMC-USP. O objetivo da Perfil-CC de 2007 foi avaliar as principais conferências nacionais e internacionais da área de acordo com sua relevância, prestígio e impacto no cenário mundial. A iniciativa atual, que foi conduzida por um conjunto de pesquisadores do Instituto de Computação (IComp) da UFAM, possui os seguintes objetivos: 1) levantar o conjunto de veículos de publicação de interesse para a Comunidade Brasileira de Computação; e 2) disponibilizar um mecanismo verificável de medição de impacto destes veículos.

Atualmente, o SHINE apresenta referenciais de qualidade de 1.788 conferências, calculados a partir do número de citações de cerca de 800.000 artigos. Dada a confiabilidade do método adotado para se calcular essas estimativas de impacto, a ferramenta foi usada pela SBC durante sua interlocução com a CAPES para qualificar as conferências da área através do QUALIS. Nesse artigo apresentamos alguns detalhes sobre a metodologia adotada para o desenvolvimento dessa ferramenta, que atualmente pode ser considerada como um importante referencial sobre qualidade das conferências para a comunidade de Computação.

O trabalho das comissões especiais da SBC

A primeira etapa do projeto Perfil-CC SBC foi selecionar o conjunto de eventos mais relevantes para os pesquisadores brasileiros da área de Computação. Para ajudar nessa tarefa, o projeto contou com a expertise das Comissões Especiais (CE) da SBC. Cada CE é formada por um conjunto de pesquisadores de uma determinada subárea da Computação, fato que facilitou a distribuição de esforços e garantiu grande qualidade na seleção de veículos. Para facilitar o trabalho das CEs, foi disponibilizada uma ferramenta online para que os membros de cada CE pudessem escolher os veículos a partir da lista de veículos do Perfil-CC de 2007, bem como a partir de outras fontes tais como DBLP, ACM-DL, IEEE Explorer, BDBComp, JEMS, etc. Durante esta fase do projeto, também foi possível selecionar veículos não presentes nestas fontes.

O resultado desse trabalho foi uma base contendo 1.788 veículos cadastrados, o que representa um crescimento de 69% em relação à iniciativa de 2007. Esse crescimento se deve, obviamente, à fundamental participação das Comissões Especais da SBC.

Coleta dos artigos e de suas citações

Conforme dito anteriormente, para se calcular o índice-H de um determinado veículo é necessário obter o número de citações de todos os seus artigos. Esse é um problema extremamente difícil pois para se obter este dado com precisão seria necessário monitorar continuamente um número desconhecido de veículos onde potencialmente pudesse haver um citação para estes artigos. Uma forma viável, ainda que aproximada de se obter tais dados de citação, é através do Google Scholar (GS), que mantém um base de informações sobre publicações de inúmeros tipos de veículos, incluindo conferências. A Figura 1 mostra um fragmento da página de respostas do GS obtida ao submetermos a consulta "*computing block importance for searching on web sites*". Observe que os artigos retornados pela máquina de busca possuem informações sobre o número de vezes que tais artigos foram citados por outras publicações.

Figura 1: Fragmento de uma página de respostas do Google Scholar, obtida ao submetermos a consulta "computing block importance for searching on web sites"

Um estudo com 317 periódicos reportado pelo Documento de Área da CAPES (2009) demonstrou que existe uma forte correlação entre os valores de citações extraídos do Google Scholar e os valores de citações usados pela *SCImago* para estimar seus índices de impacto. Esse estudo levou o Comitê da Área de Computação da CAPES a sugerir o uso das citações do Scholar para se estimar os índices de impacto das conferências da área de Computação. A ferramenta *Publish or Perish* também usa as citações do Scholar para este mesmo fim.

Uma forma trivial de se obter o número de citações de uma parte dos artigos de uma conferência é formular uma consulta baseada no nome dessa conferência e submetê-la ao Google Scholar. Através desse procedimento, o Scholar retornará uma listagem de artigos da conferência que se deseja coletar. No entanto, esse procedimento não garante que todos os artigos da conferência sejam retornados pelo Scholar, visto que os nomes das conferências estão sujeitos a muitas variações e ambiguidades. Por exemplo, listamos abaixo os nomes encontrados em várias citações para um mesmo artigo do SBES 2003:

- Proc. of Brazilian Symposium on Software Engineering (SBES'03),
- Proc. of Brazilian Symp. on Soft. Eng.,
- Proc. Brazilian Symp. on Software Engineering,
- Proc of Simpósio Brasileiro de Engenharia de Software

Desta forma, para se adotar a técnica descrita, seria necessário antecipar todos os possíveis nomes de todas as conferências da base. Além disso, ao submetermos uma consulta com o nome da conferência, o Scholar pode retornar publicações de veículos com nomes similares ao da conferência desejada.

Para contornar esse problema, a ferramenta SHINE calcula o índice-H de um veículo através da agregação das citações dos artigos deste veículo. Para entendermos melhor, seja V uma conferência e seja P_i a lista de artigos publicados em V no ano i. Cada artigo p_{ij} em P_i tem um número $C(p_{ij})$ de citações obtidas no Scholar. O $C(p_{ij})$, que é utilizado para o cálculo do índice-H de V, é obtido enviando uma consulta ao GS contendo o título, o ano, e o primeiro autor de p_{ij}.

Julgamos que esse processo melhora o anterior, pois a consulta ao Google Scholar é feita através de informações do artigo ao invés de pelo nome da conferência, que é muito mais sujeito a mudanças e a ambiguidades. Para cada artigo de uma dada conferência, o coletor do SHINE formula uma consulta ao Google Scholar contendo informações sobre o título, o ano, e o primeiro autor do artigo. O coletor então identifica, na página de respostas do GS, a entrada que corresponde exatamente ao paper que se deseja encontrar, para aí então coletar as informações sobre as citações desse artigo.

De onde vieram a lista dos artigos? DBLP, IEEE, BDBComp, JEMS, etc. Além dessas, as Comissões Especiais da SBC também indicaram outras fontes.

O índice-H

O índice de impacto de um veículo é uma estimativa do interesse da comunidade científica pelas publicações desse veículo. Conforme já dito, a maioria das métricas para se estimar esse interesse são fórmulas baseadas no número de citações das publicações. Dentre as métricas existentes, destacamos as já citadas: fator de impacto (FI) e índice-H. O índice-H foi originalmente proposto por J. Hirsch, e seu propósito original foi quantificar o impacto das publicações de cientistas individuais. No SHINE, assim como no SCimago, o índice-H é usado para calcular o impacto de veículos, e não de pesquisadores.

O índice-H de um evento é calculado com base no número de artigos publicados em determinada quantidade de anos, e na quantidade de citações de cada artigo. O índice-H é definido como o número de artigos de uma conferência com número de citações maior ou igual a H. Por exemplo, se uma conferência possui índice-H igual a 20, isso significa dizer que esse veículo possui 20 artigos com 20 ou mais citações. Na próxima seção, vamos exemplificar esse cálculo usando um caso real através da ferramenta SHINE.

A Ferramenta

A Figura 2 mostra um *screenshot* da página principal da ferramenta SHINE, que pode ser acessada através do endereço **shine.icomp.ufam.edu.br**. Através dessa página, o visitante pode selecionar o veículo que deseja analisar, e o intervalo de anos que será considerado durante o cálculo do índice-H. Uma vez que os parâmetros foram setados, o visitante pode clicar no botão Enviar para que a ferramenta efetive o cálculo do índice-H.

Figura 2: A tela principal da ferramenta SHINE, onde o visitante pode informar o nome da conferência que deseia consultar, e o intervalo de anos que serão considerados durante o cálculo do índice-H.

A Figura 3 mostra o resultado obtido após selecionarmos a conferência "*SIGIR - Conference on Research & Development in Information Retrieval*", e o intervalo de anos de 2000 até 2012. Observe que o resultado indica que o SIGIR possui índice-H igual a 113 para o intervalo de anos selecionado. A página de resultados apresenta a listagem de artigos publicados na conferência durante esse intervalo de anos, ordenados de acordo com o número de citações dos artigos (em ordem decrescente). Por questões de espaço, a figura em questão apresenta apenas os três primeiros artigos dessa listagem.

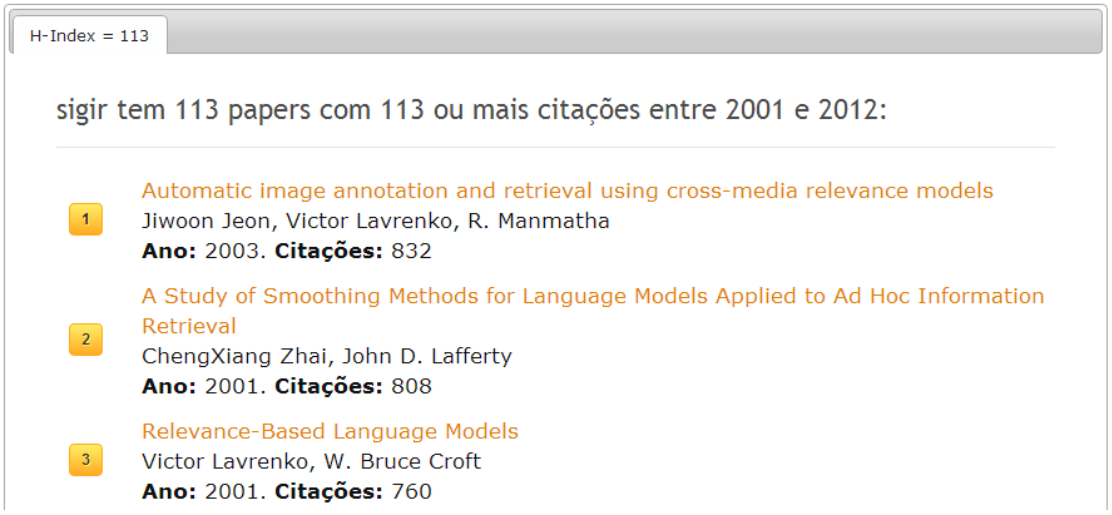

Figura 3: Parte da página de resultados obtida ao selecionarmos a conferência SIGIR, e o intervalo de 2000 a 2012. Observe que o resultado indica que o SIGIR possui índice-H igual a 122 nesse intervalo de anos.

Acessando a ferramenta SHINE, podemos verificar que os primeiros 113 artigos da listagem acima possuem 113 ou mais citações, enquanto que todos os demais possuem menos de 113 citações. É justamente esse fato que confere ao SIGIR um índice-H igual a 113 para os anos considerados.

Atualmente, a ferramenta SHINE possui 1.778 veículos e 754.832 artigos cadastrados, uma média de 424 artigos por conferência. O número total de citações da base é de 7.356.004, representando uma média de 9 citações por artigo. O gráfico da esquerda da Figura 4 mostra a distribuição de citações por artigo da base do SHINE. Em relação aos índices de impacto, o índice-H médio dos veículos da base é igual a 18. O gráfico da direita da Figura abaixo mostra a distribuição dos valores de índice-H encontrados. Muitos pesquisadores têm reportado que os valores de índice-H encontrados são muito coerentes com a realidade das conferências de suas áreas, fato que confere grande confiabilidade aos referenciais de impacto divulgados pela ferramenta SHINE.

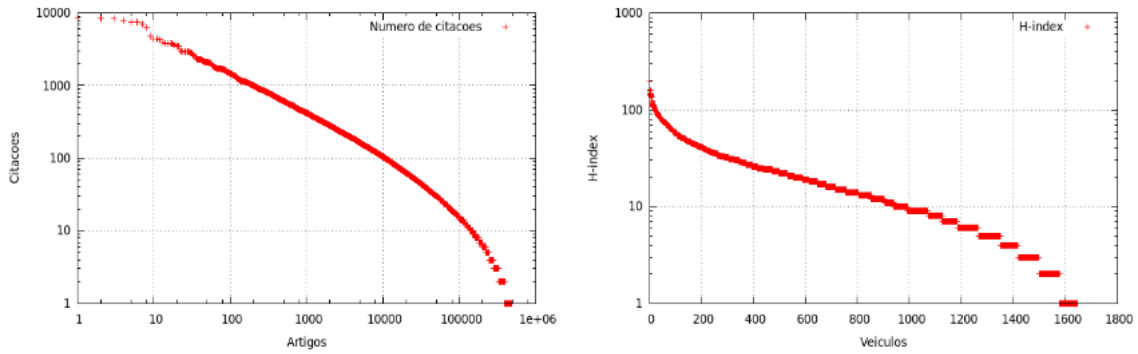

Figura 4: O gráfico da esquerda mostra a distribuição de citações por artigos da base de dados do SHINE, enquanto que o gráfico da direita mostra a distribuição dos valores de H-index dos veículos da base.

Embaixo do Capô

O SHINE também tem sido usado como um laboratório para as tecnologias que estão em contínuo desenvolvimento no Grupo de Bancos de Dados e Recuperação de Informação do Instituto de Computação da UFAM. Diversos resultados recentes na área de Extração de Dados, Recuperação de Informação, Coletores de Páginas, Integração e Limpeza de Dados, etc., além da expertise de professores e alunos (graduação, mestrado e doutorado) do grupo têm sido, empregados no desenvolvimento de vários dos módulos do SHINE.

Concluindo

A ferramenta SHINE constitui uma importante conquista para a comunidade de Computação, por ser a primeira fonte verificável de informações sobre a qualidade das conferências da área. O desenvolvimento da ferramenta contou com a participação das Comissões Especiais da SBC para selecionar as conferências de interesse da comunidade de Computação, e adotou o Google Scholar como fonte de citações dos artigos. Dada a confiabilidade de seus resultados, a ferramenta SHINE foi adotada como base para a classificação realizada pela Comissão de Avaliação do QUALIS de Computação de 2012.

Sobre os Autores

David F. de Oliveira possui graduação e mestrado em Ciência da Computação pela Universidade Federal do Amazonas, e doutorado em Ciência da Computação pela Universidade Federal de Minas Gerais (2010). Atualmente é professor adjunto do Instituto de Computação da Universidade Federal do Amazonas. Suas áreas de interesse são recuperação de informação, mineração de dados na Web e bancos de dados.

Altigran S. da Silva é professor associado do Instituto de Computação da UFAM onde atua como pesquisador, professor e orientador na graduação e na pós. Concluiu seu doutorado em Ciência da Computação pela UFMG em 2002. Seus interesses de pesquisa envolvem Gerência de Dados, Recuperação de Informação e Mineração de Dados na Web. É atualmente o Coordenador Adjunto da área de Computação na CAPES e desde 2005 é membro da diretoria da SBC.

Henrique Cavalcanti concluiu o ensino-medio na Fundação Nokia de Ensino (2010). Atualmente, é graduando de Ciência da Computação pelo Instituto de Computação da Universidade Federal do Amazonas. Tem experiência na área de Banco de Dados e Recuperação de Informação.