UNIVERSIDADE FEDERAL DO AMAZONAS
INSTITUTO DE COMPUTAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA

LIVE SHINE - UMA FERRAMENTA PARA SUPORTE À AVALIAÇÃO DE IMPACTO DE EVENTOS CIENTÍFICOS EM COMPUTAÇÃO

Leonardo Fontes do Nascimento

Abril de 2016

Manaus - AM

UNIVERSIDADE FEDERAL DO AMAZONAS
INSTITUTO DE COMPUTAÇÃO
PROGRAMA DE PÓS-GRADUAÇÃO EM INFORMÁTICA

LIVE SHINE - UMA FERRAMENTA PARA SUPORTE À AVALIAÇÃO DE IMPACTO DE EVENTOS CIENTÍFICOS EM COMPUTAÇÃO

Leonardo Fontes do Nascimento

Dissertação de Mestrado apresentada ao Programa de Pós-Graduação em Informática, Instituto de Computação - IComp, da Universidade Federal do Amazonas, como parte dos requisitos necessários à obtenção do título de Mestre em Informática.

Orientador: Altigran Soares da Silva

Abril de 2016

Manaus - AM

Ficha Catalográfica

Ficha catalográfica elaborada automaticamente de acordo com os dados fornecidos pelo(a) autor(a).

Nascimento, Leonardo Fontes do
N244l Live Shine - Uma ferramenta para suporte à avaliação de impacto de eventos científicos em computação / Leonardo Fontes do Nascimento. 2016
86 f.: il. color; 31 cm.

Orientador: Altigran Soares da Silva
Dissertação (Mestrado em Informática) - Universidade Federal do Amazonas.

1. Live Shine . 2. Bibliotecas Digitais . 3. Coleta Colaborativa. 4. Índices de Impacto. I. Silva, Altigran Soares da II. Universidade Federal do Amazonas III. Título

LIVE SHINE - UMA FERRAMENTA PARA SUPORTE À AVALIAÇÃO DE IMPACTO DE EVENTOS CIENTÍFICOS EM COMPUTAÇÃO

Leonardo Fontes do Nascimento

DISSERTAÇÃO SUBMETIDA AO CORPO DOCENTE DO PROGRAMA DE PÓS-GRADUAÇÃO DO INSTITUTO DE COMPUTAÇÃO DA UNIVERSIDADE FEDERAL DO AMAZONAS COMO PARTE DOS REQUISITOS NECESSÁRIOS PARA A OBTENÇÃO DO GRAU DE MESTRE EM INFORMÁTICA.

Aprovado por:

Prof. Altigran Soares da Silva, Doutor

Prof. Philippe Olivier Alexandre Navaux, Doutor

Prof. David Fernandes de Oliveira, Doutor

Prof. Edleno Silva de Moura, Doutor

ABRIL DE 2016

MANAUS, AM – BRASIL

Ao meu irmão Leandro (in memoriam), pela amizade, companheirismo e por ter me ensinado que a vida é mais simples do que parece ser.

Agradecimentos

A Deus, por permitir realizar meus sonhos, por me dar forças para superar as dificuldades e não me deixar desistir.

Aos meus pais Alginéia e Francisco, pelo apoio em todos os dias da minha vida e que mesmo em dificuldades me ajudaram a concluir meus estudos.

A minha esposa Paloma, pelo incentivo nos momentos mais difíceis nesse último ano e por sempre me ajudar a alcançar os meus objetivos.

A minha irmã Letícia e minhas sobrinhas, pela felicidade e alegria que me trazem todos os dias.

Aos professores Altigran Soares e David Fernandes, pelos ensinamentos, dedicação, paciência e orientação nesses dois últimos anos.

Ao professor Tiago de Melo, pelo ajuda prestada no desenvolvimento deste trabalho.

Aos professores Fábio Santos e Edgard Silva, pelo reconhecimento e indicação ao programa de mestrado.

Ao Diego Barros, pela amizade e ajuda prestada durante o programa de mestrado.

Aos amigos de mestrado, pelos grupos de estudos e também por compartilharem momentos alegres comigo.

A todos vocês meu muito obrigado.

"O que sabemos é uma gota; o que ignoramos é um oceano." (Isaac Newton)

Resumo

Uma preocupação frequente entre os pesquisadores é que os resultados de suas pesquisas sejam publicados em veículos de impacto na comunidade científica. Geralmente, os índices de impacto são obtidos através de métricas baseadas no número de citações que seus artigos recebem. Instituições tais como o *SCImago* e *Thomson Reuters* fornecem índices de impacto precisos para os principais periódicos internacionais. Embora isso seja suficiente para a maioria das áreas, para a área de Ciência da Computação as conferências e outros eventos científicos são igualmente importantes como veículos de publicação. No entanto, atualmente não existe nenhuma solução que seja universalmente aceita para se obter índices precisos sobre conferências, pois as ferramentas mais utilizadas para esse fim apresentam divergências entre os índices gerados para uma mesma conferência e ano. Neste trabalho propomos uma ferramenta denominada Live SHINE, cujo objetivo é gerar índices de impacto de alta precisão de conferências de Ciência da Computação a partir de dados fornecidos pelo *Google Scholar*. Nossa ferramenta utiliza um método baseado em técnicas de aprendizagem de máquina que filtra automaticamente os metadados fornecidos pelo Google Scholar e considera no cálculo dos índices de impacto apenas os dados de citações de artigos que de fato pertencem a conferência. Os experimentos realizados indicam que nosso método é eficaz, alcançando uma métrica F1 média acima de 0.9 considerando 30 conferências analisadas. Além disso, desenvolvemos também uma nova estratégia distribuída e colaborativa de coleta de citações, na qual as consultas enviadas ao Google Scholar para recuperar os valores atualizados de citações de artigos são disparadas pela própria interface do usuário, evitando problemas como sobrecarga da rede, demora na atualização das citações e bloqueio frequente por parte do Google Scholar. Assim, essa estratégia faz com que a comunidade de usuários colabore para manter os dados de citações atualizados para o benefício de todos.

Abstract

A common concern among researchers is that the results of their research are published in venues of impact in the scientific community. In general, the impact indices are obtained through metrics based on the number of citations that your articles receive. Institutions such as the *SCImago* and *Thomson Reuters* provide precise impact indices for major international journals. While this is sufficient for most areas, in Computer Science conferences and other scientific events are also important as publishing venues. However, currently, there is no solution that is universally accepted to obtain accurate indices on conferences, because the tools most commonly used for this purpose have differences between the indices generated for the same conference and year. In this dissertation, we propose a tool called Live SHINE, whose goal is to generate high-precision impact indices of Computer Science Conferences from data provided by *Google Scholar*. Our tool uses a method based on machine learning techniques that automatically filters the metadata provided by Google Scholar, and considers in the calculation of impact indices only citation data from articles that truly belong to the conference. Our experiments show that our method is effective, achieving an average F1 metric above 0.9 for 30 analyzed conferences. In addition, we also developed a new distributed and collaborative strategy of collecting citations, in which the queries sent to Google Scholar to retrieve the updated values of article citations are triggered by the user interface, avoiding problems such as network overload, delay in update citations and frequent blocking by Google Scholar. Thus, this strategy makes the community of users collaborate to keep updated the data citations for the benefit of all.

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

Uma das preocupações mais presentes entre os pesquisadores das diversas áreas científicas é assegurar que os resultados de suas pesquisas sejam publicados em veículos (periódicos, simpósios, conferências e outros) de grande visibilidade e impacto na comunidade científica mundial. Em geral, os índices de impacto de um determinado veículo são obtidos através de métricas baseadas no número de vezes que os artigos desse veículo são citados por outros artigos. Exemplos bem conhecidos de tais métricas baseadas em citações são o *Impact Factor (IF)* [Garfield, 1972] e o *H-Index* [Hirsch, 2005, Braun et al., 2006].

Para a grande maioria das áreas do conhecimento, é possível obter os índices de impacto de seus principais veículos através de instituições tais como o *SCImago*[1] e o *Thomson-Reuters*[2]. Esse último, por exemplo, estima o impacto dos veículos através da métrica *IF*. Anualmente essas métricas são calculadas para os periódicos indexados na sua base de dados *Web of Science* e depois são publicados no *Journal Citations Reports (JCR)* da *Thomson-Reuters*. Dado seu enorme prestígio perante a comunidade científica mundial, o *JCR* é comumente aceito e utilizado por pesquisadores de várias áreas para avaliar a qualidade da produção intelectual desses periódicos.

No caso de publicações em periódicos, as instituições acima mencionadas, bem como outras instituições similares, monitoram continuamente as citações de artigos publicados em um conjunto fechado de veículos e periodicamente reportam os índices de impacto desse conjunto, assim como ocorre no *JCR*. No entanto, no caso de artigos publicados em anais de conferências[3], atualmente não existem instituições semelhantes ou ferramen-

[1] http://www.scimagojr.com

[2] http://thomsonreuters.com

[3] O termo "conferência" é utilizado aqui para referir-se também a outros eventos científicos tais como

tas que são universalmente aceitas, apesar de existirem algumas tentativas por diferentes indivíduos e instituições em todo o mundo.

Ferramentas como o *Publish or Perish (PoP)* [Harzing, 2007] e o *SHINE*[4]; máquinas de busca acadêmicas como o *Google Scholar (GS)*[5], o *Microsoft Academic Search (MAS)*[6] e *CiteSeerX*[7]; e bibliotecas digitais como a *ACM Digital Library*[8] e *IEEE Xplore*[9], que comumente são utilizadas para obter índices de impacto de conferências, apresentam problemas ou limitações que podem prejudicar a precisão de tais índices. Isso representa um enorme inconveniente para os pesquisadores da comunidade de Ciência da Computação que, ao contrário dos pesquisadores da maioria das comunidades científicas, consideram conferências igualmente importantes para se publicar artigos científicos [Zhang, 2011, Bar-Ilan, 2010, Vardi, 2009].

Como ilustrado no exemplo que apresentamos no Capítulo 3 desta dissertação, essas soluções dificilmente concordam com relação aos índices de impacto por elas calculados para uma mesma conferência/ano, e, de fato, muitas vezes apresentam uma grande discrepância nas métricas de impacto. Essa diferença pode ser causada por muitos fatores, no entanto, os dois fatores principais são: *discrepância na lista de artigos considerada para a conferência e nos dados de citações de artigos*.

Com relação ao primeiro fator, existem ferramentas e serviços que embora utilizem a mesma fonte de dados para obter as listas de artigos e dados de citações, fornecem índices de impacto bastante diferentes para uma mesma conferência e ano, e isso acontece devido as diferenças nas listas de artigos que essas ferramentas consideram no cálculo da estimativa do impacto. Esse é um problema extremamente difícil de lidar, pois não são todas as conferências que disponibilizam seus artigos publicados online, e muitas vezes quando disponibilizam, não o fazem por meio de uma única fonte Web ou biblioteca digital padrão. Esse problema pode ser verificado na ferramenta *Publish or Perish*, cujo objetivo é coletar e calcular índices de impacto a partir de metadados e dados de citações de artigos disponíveis em fontes Web, especialmente o GS. Nessa ferramenta, o usuário pode realizar uma consulta por artigos de uma determinada conferência/ano, essa consulta

workshops, simpósios, etc.

[4]http://shine.icomp.ufam.edu.br

[5]http://scholar.google.com

[6]http://academic.research.microsoft.com

[7]http://citeseerx.ist.psu.edu

[8]http://dl.acm.org

[9]http://ieeexplore.ieee.org

é repassada ao GS, e o conjunto de resposta retornado é utilizado como a lista de artigos da conferência para o cálculo dos índices de impacto. No entanto, as consultas emitidas por essa ferramenta trazem muitos registros de artigos que não pertencem de fato a conferência/ano desejada [Jacsó, 2009], e desse modo considera artigos de outros veículos para o cálculo, o que muitas vezes resulta em índices de impacto imprecisos.

Além disso, para realizar o cálculo dos índices com alta precisão também é necessário obter os dados de citações dos artigos atualizados. Contudo, manter atualizado os dados de citações de todos os artigos de uma conferência também é uma tarefa extremamente difícil, pois isso requer um monitoramento constante das citações feitas em um número desconhecido de artigos, visto que a cada ano esses artigos podem receber novas citações e novos artigos podem ser publicados. Esse problema pode ser notado na ferramenta SHINE, cujo objetivo é fornecer um mecanismo verificável de medição de impacto das principais conferências de Ciência da Computação. Nessa ferramenta, as listas de artigos corretas das conferências são obtidas a partir de bibliotecas digitais específicas da área e os dados de citações dos artigos são obtidos a partir de consultas realizadas no *Google Scholar*. No entanto, por não manter um monitoramento constante dos artigos e seus dados de citações, essa ferramenta se encontra desatualizada e fornece atualmente índices de impacto subestimados que não representam a realidade das conferências.

Para resolver os problemas descritos, neste trabalho propomos uma ferramenta denominada *Live SHINE*, cujo objetivo é gerar índices de impacto de alta precisão de conferências de Ciência da Computação a partir de dados fornecidos pelo *Google Scholar (GS)*. Nossa solução, disponibilizada em formato de extensão para o *Navegador Web Google Chrome*[10], funciona sobre o GS de maneira a auxiliar usuários que utilizam essa máquina de busca acadêmica para obter os índices de impacto de uma determinada conferência. Ao ativar nossa extensão, o usuário pode facilmente selecionar as conferências sobre as quais deseja obter os índices de impacto. Uma vez que uma conferência for selecionada, bem como a edição ou período desejado, nossa ferramenta repassa ao GS uma consulta previamente estabelecida que retorna a maior quantidade de registros referentes a artigos que de fato pertence a esta conferência, evitando assim que o usuário tenha que formular manualmente a melhor consulta para a conferência desejada. No entanto, essa consulta não garante um conjunto de resposta totalmente preciso, pois são retornados também re-

[10]https://www.google.com.br/chrome/browser/desktop/

gistros de artigos referentes a outras conferências no conjunto de resposta. Durante nossos testes isso aconteceu com frequência e acreditamos que a razão disso é o fato do GS não ter sido concebido inicialmente para este fim, embora outras ferramentas e estudos como os realizados por [Harzing, 2014] indiquem que o GS nos últimos anos se tornou uma fonte promissora de dados de artigos e citações para a análise de impacto e pesquisas bibliométricas. Deste modo, a fim de obter a lista de artigos mais correta possível para ser considerada no cálculo do impacto, nossa ferramenta filtra automaticamente os registros fornecidos pela GS utilizando um método baseado em Aprendizagem de Máquina desenvolvido neste trabalho. Esse método, chamado *SHINER*[11], é capaz de selecionar automaticamente, no conjunto de resposta do GS, os registros que de fato referenciam artigos pertencentes a conferência escolhida e assim obter as listas de artigos corretas dessa conferência a partir do GS.

Conforme será apresentado no Capítulo 5, os resultados obtidos nos experimentos indicaram que nosso método é realmente eficaz, alcançando um valor de métrica F1 acima da média 0.9 e acima de 0.8 para 26 das 30 conferências analisadas. Seu principal objetivo é garantir que o cálculo dos índices de impacto seja realizado com base na lista de artigos mais correta possível em termos de precisão e revocação, ou seja, uma lista exclusiva que contém idealmente todos os artigos somente da conferência/ano desejada.

Além disso, conforme será apresentado no Capítulo 6, para utilizar dados de citações dos artigos atualizados nos cálculos dos índices de impacto fornecidos pela ferramenta Live SHINE, desenvolvemos também uma nova estratégia distribuída e colaborativa de coleta de citações, na qual as consultas enviadas ao GS para recuperar os valores atualizados de citações de artigos são disparadas pela própria interface do usuário, evitando problemas como sobrecarga da rede, demora na atualização das citações e bloqueio frequente por parte do GS. Assim, essa estratégia faz com que a comunidade de usuários colabore para manter os dados de citações utilizados no cálculos dos índices de impacto sempre atualizados, de modo que todos sejam beneficiados.

Deste modo, acreditamos que o *Live SHINE* pode ajudar os pesquisadores da comunidade de Ciência da Computação a lidar com os problemas encontrados em obter índices de impacto de qualidade sobre eventos científicos, e assim possam assegurar que seus artigos sejam publicados em conferências de grande impacto avaliadas por uma solução que

[11] A **S**imple **H-IN**dex **E**stimator **R**elayer.

entrega maior precisão nos índices calculados, tal como hoje ocorre com os periódicos.

Esta dissertação está estruturada da seguinte forma. No Capítulo 2, são apresentados os trabalhos relacionados e soluções mais utilizadas atualmente para estimar os índices de impacto de conferências de Ciência da Computação. No Capítulo 3, descrevemos uma visão geral do problema abordado e a nossa solução proposta para este problema. No Capítulo 4, apresentamos nosso método de filtragem automático denominado SHINER desenvolvido no trabalho. No Capítulo 5, os experimentos realizados e os resultados obtidos são descritos. No Capítulo 6, apresentamos a coleta colaborativa de metadados, inclusive dados de citações realizada pela interface da ferramenta. No Capítulo 7, apresentamos com mais detalhes a nossa ferramenta proposta denominada Live SHINE, e por fim, no Capítulo 8 discutimos as conclusões e o direcionamento para os trabalhos futuros.

Capítulo 2

Conceitos Básicos e Trabalhos Relacionados

Este capítulo introduz conceitos básicos necessários para melhor compreensão da solução proposta, bem como uma revisão da literatura relacionada à área de abrangência deste trabalho. Na discussão seguinte, são descritos as métricas de impacto utilizadas para avaliar conferências. Em seguida, são apresentadas ferramentas bastante utilizadas pelos pesquisadores da área para obter os índices de impacto de conferências de Ciência da Computação.

2.1 Métricas de Avaliação de Impacto

Avaliar o impacto de conferências científicas é um desafio cada vez mais importante para pesquisadores e bibliotecários, especialmente nas áreas de Ciência da Computação e Engenharia Elétrica [Zhuang et al., 2007, Martins et al., 2009, Vardi, 2009]. A forma mais comum para avaliar esse impacto é através de métricas baseadas em citação [Laender et al., 2008, Zhuang et al., 2007, Chiu & Fu, 2010, Bornmann & Daniel, 2005, Bornmann & Daniel, 2007], como o *Impact Factor (IF)* da Thomson [Garfield, 1972], *G-Index* [Egghe, 2006] e *H-Index* [Hirsch, 2005]. No entanto, o uso dessas métricas requer o acesso a lista de artigos de cada veículo de interesse, bem como os dados de citação de cada um desses artigos.

Nas próximas seções são apresentadas algumas das métricas de estimativa de impacto mais utilizadas para avaliar periódicos e conferências das diversas áreas.

2.1.1 Impact Factor

O *Impact Factor (IF)* [Garfield, 1972] é uma métrica proposta pelo *Institute for Scientific Information (ISI)*, hoje denominado *Thomson Reuters*, que visa estabelecer uma média de citação por artigo publicado para avaliar a importância de um periódico. O *IF* de um periódico em determinado ano é definido como o número médio de citações recebidos durante esse ano pelos artigos publicados nesse periódico nos dois anos anteriores. Por exemplo, para calcular o *IF* dos periódicos para o ano de 2016, divide-se o número de vezes que os artigos publicados em 2014 e 2015 são citados em 2016 pelo número total de artigos publicados em 2014 e 2015. Essa é a métrica mais popular utilizada para medir a qualidade de periódicos, sendo largamente aplicada e aceita mundialmente pelos pesquisadores de diversas comunidades científicas, embora estudos como os apresentados por [Harzing & Van Der Wal, 2009] apontem alguns problemas e falhas importantes que prejudicam a precisão dessa métrica.

Vale ressaltar que anualmente o instituto *Thomson Reuters* calcula o *IF* de um conjunto fechado de periódicos que estão indexados na sua base de dados *Web of Science* e publica no *Journal Citations Reports (JCR)*. Infelizmente, ainda não existem ferramentas ou serviços que forneçam valores de *IF* para conferências.

2.1.2 H-Index

O *h-index* é uma métrica introduzida originalmente por [Hirsch, 2005] para capturar a qualidade da produção científica de um indivíduo ao longo dos anos. Devido sua grande aceitação entre os pesquisadores, seu uso tem sido estendido para avaliar conferências e periódicos. De maneira análoga a definição original, uma conferência tem *h-index h* se existem pelo menos *h* artigos publicados nessa conferência que tenham recebido pelo menos *h* citações. Deste modo, o *h-index* fornece uma forma de avaliação da produção científica que combina quantidade (número de artigos) e qualidade (impacto, ou citações desses artigos) [Harzing & Van Der Wal, 2009], ou seja, uma conferência tem *h-index* alto se seus artigos publicados também tem alta qualidade e maiores chances de serem citados por outros artigos.

Estudos recentes, como os apresentados por [Harzing & Van Der Wal, 2009, Harzing, 2013, Harzing, 2014, Harzing, 2010, Harzing, 2008], [Braun et al., 2006] e [Bornmann et al., 2008], apontam que o uso dessa métrica, aliado ao uso do *Google*

Scholar (GS) como fonte de dados de artigos e de citações, se tornou nos últimos anos uma alternativa promissora e muitas vezes superior ao *IF* na tarefa de avaliar o impacto de periódicos e conferências.

Importante ressaltar que as ferramentas mais populares como o *Publish or Perish (PoP)* e o *SHINE* utilizam o *h-index* e outras métricas derivadas dele para estimar o impacto de conferências de Ciência da Computação e outras áreas. Além disso, o *Google Scholar* disponibiliza um serviço chamado *Google Metrics* que fornece o valor do *h-index*, limitado a um período de 5 anos, para vários veículos de publicação, tanto conferências quanto periódicos, para várias áreas do conhecimento.

2.1.3 G-Index

O *g-index* é uma métrica proposta por [Egghe, 2006] como uma forma de melhoria do *h-index* de [Hirsch, 2005] para medir o desempenho de citação global de um conjunto de artigos. Proposto originalmente para avaliar o impacto da produção científica de pesquisadores, seu uso também tem sido estendido para estimar o impacto de conferências e periódicos. De maneira análoga a sua definição original, se o conjunto de artigos de uma conferência está classificado em ordem decrescente de número de citações que eles receberam, o *g-index* é o (único) maior número g, tal que os g artigos melhor classificados receberam (juntos) no mínimo g^2 citações. Isso também significa que os $g+1$ artigos melhor classificados têm menos que $(g+1)^2$ citações. Deste modo, assim como o *h-index*, *o g-index* representa uma relação entre artigos publicados e o nível de citações que eles recebem, no entanto, essa métrica é mais sensível a artigos altamente citados [Rosenstreich & Wooliscroft, 2009].

Para exemplificar o funcionamento dessa métrica, assim com apresentar a diferença entre o *h-index* e o *g-index*, podemos analisar os artigos publicados em uma conferência fictícia X em um determinado ano Y por meio da Tabela 2.1. Nesta tabela, TC representa o número total de citações para cada artigo classificado na posição r, e $\sum TC$ representa o total acumulado de citações dos artigos em cada posição da classificação. Os dados destacados mostram um *h-index* $h = 11$ e um *g-index* $g = 18$ para esta conferência. Ou seja, $h = 11$ pois 11 é a posição de classificação em que todos os artigos acima têm pelo menos 11 citações (o que significa também que os artigos a partir da posição de classificação 12, não tem mais do que 11 citações). E temos $g = 18$ porque 18 é a posição

de classificação mais alta, tal qual os 18 artigos melhor classificados têm pelo menos $18^2 = 324$ citações (neste ponto $333 > 324$); enquanto que na posição de classificação 19 temos $343 < 19^2 = 361$.

TC	r	$\sum TC$	r^2
45	1	45	1
40	2	85	4
35	3	120	9
34	4	154	16
19	5	173	25
16	6	189	36
15	7	204	49
14	8	218	64
14	9	232	81
14	10	246	100
13	**11**	259	121
11	12	270	144
11	13	281	169
11	14	292	196
11	15	303	225
10	16	313	256
10	17	323	289
10	**18**	**333**	**324**
10	19	343	361
9	20	352	400

Tabela 2.1: Classificação dos artigos de uma conferência fictícia X para um ano Y de acordo com os números de citações recebidos.

[Rosenstreich & Wooliscroft, 2009] apresentam uma abordagem para estimar o impacto de periódicos científicos através do uso desta métrica combinado ao uso do motor de busca acadêmico *Google Scholar (GS)*. Para eles o GS foi escolhido como forma de obter uma cobertura mais ampla de contagem de citações do que os bancos de dados conhecidos como o *Web of Science* da *Thomson*, no entanto, o seu uso requer o filtro dos registros retornados para remover dados duplicados e errados. Além disso, para eles a utilização do *g-index* como forma de classificar periódicos mostrou-se inovadora e útil na análise de citações, permitindo uma avaliação mais robusta do impacto de periódicos.

Vale destacar que a ferramenta *Publish or Perish (PoP)* ao realizar a avaliação de periódicos e conferências, gera o *h-index* e o *g-index* para a lista de artigos retornados na consulta por um veículo.

2.1.4 Eigenfactor

As métricas *Eigenfactor* são fornecidas por um algoritmo proposto por [Bergstrom et al., 2008] como uma forma mais sofisticada de avaliar periódicos utilizando dados de citações. Essas métricas visam explorar informações úteis que a rede de citações dos artigos podem oferecer. Por exemplo, a partir dessa rede é possível verificar a origem das citações, e consequentemente é possível explorar o fato de que citações vindas de periódicos de grande prestígio valem mais do que citações vindas de periódicos de prestígio inferior. Desta forma, a ideia por trás dessas métricas é que podemos utilizar o poder computacional para extrair a riqueza de informação inerente a estrutura de redes de citações.

A abordagem utilizada para o levantamento dessas métricas é semelhante a abordagem utiliza no algoritmo *PageRank* [Page et al., 1999] do *Google* para retornar resultados de pesquisas. Quando o algoritmo *PageRank* classifica páginas na Web, considera não apenas a quantidade de *hiperlinks* que uma página recebe, mas também leva em conta de onde vêm esses *hiperlinks*. Assim, o algoritmo *Eigenfactor* faz algo semelhante, contudo, ao invés de classificar páginas na Web, classifica periódicos, e ao invés de utilizar *hiperlinks*, utiliza as citações dos artigos. A Figura 2.1 apresenta a principal diferença entre a maioria dos algoritmos de cálculo de impacto e o *Eigenfactor*, que ao invés de utilizar apenas os dados de citações locais, utiliza toda a estrutura da rede de citações para avaliar a importância de um determinado periódico *x*.

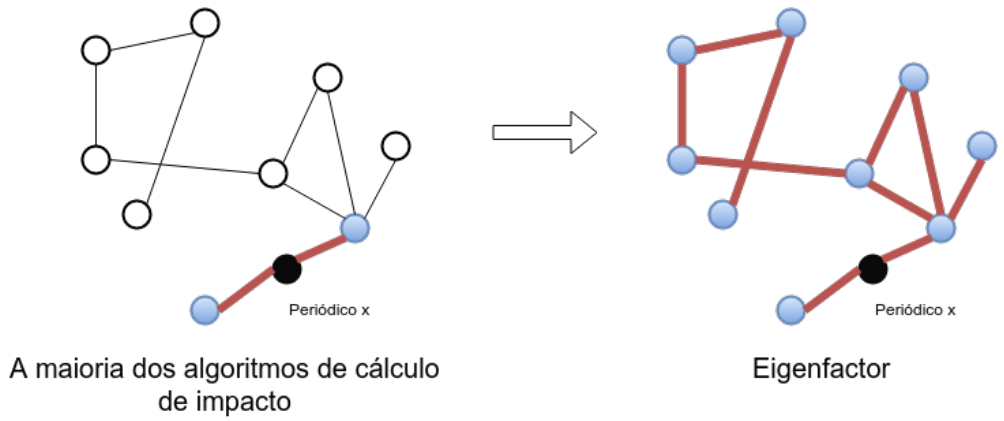


Figura 2.1: Principal diferença entre a maioria dos algoritmos de cálculo de impacto e o *Eigenfactor*.

Esse algoritmo foi aplicado em vários conjuntos de dados bibliométricos a partir mui-

tas fontes. No website oficial[1] desse algoritmo são apresentados os resultados da sua aplicação para os dados do *Journal Citation Reports (JCR)* da *Thomson Reuters*. Para cada um dos mais de 7000 periódicos listados no JCR foram levantadas duas principais métricas: *o Eigenfactor Score* e o *Article Influence Score*[2].

O *Eigenfactor Score* é uma métrica que representa a importância do periódico para a comunidade científica, se um periódico dobra em tamanho enquanto a qualidade de seus artigos permanece constante, é esperado que seu *Eigenfactor Score* passe a ser o dobro. Assim, um grande periódico que publica milhares de artigos por ano, vai ter um *Eigenfactor Score* altíssimo simplesmente baseado em seu tamanho. Por exemplo, se um periódico tem *Eigenfactor Score* de 1.0, significa que ele tem 1% da influência total de todas as publicações indexadas. Em 2013, o periódico *Nature* teve o maior *Eigenfactor Score*, com um valor de 1.603.

O *Article Influence Score* é uma métrica da influência média, por artigo, de todos os artigos de um periódico, e como tal pode ser comparada com o *Impact Factor* da *Thomson Reuters*. Essa métrica é normalizada com a média de artigos na base de dados do JCR tendo um *Article Influence Score* de 1.0. Assim se um periódico tem *Article Influence Score* de 3.0, significa que seus artigos são em média 3 vezes mais influentes do que a média de artigos na base de dados do JCR. Em 2006, o periódico com maior *Article Influence Score* foi o *Annual Reviews of Immunology*, com um valor de 27.5. Isso significa que a média de artigos nesse periódico teve 27 vezes a influência da média de artigos no JCR.

Para [Bergstrom et al., 2008] classificar periódicos é um dos muitos usos para os dados de citações. Além de fornecer as métricas *Eigenfactor*, para ele é possível fazer uma análise do cenário acadêmico e a forma como esse cenário evolui. Infelizmente, esse algoritmo ainda não foi aplicado para avaliar conferências, e portanto não temos acesso a essas métricas para eventos científicos. No entanto, se no futuro esse algoritmo for utilizado para avaliar o impacto de conferências, essas métricas podem vir a se tornar uma ferramenta valiosa para a área de Ciência da Computação.

[1] http://www.eigenfactor.org

[2] http://www.eigenfactor.org/methods.htm

2.2 Ferramentas Para Obter Métricas de Impacto

Um grande passo no sentido de facilitar o acesso a informações de impacto de conferências é o advento das máquinas de busca acadêmicas, como o Google Scholar (GS). Graças a acordos com editoras de veículos científicos, fornecedores de base de dados e sociedades acadêmicas, a quantidade de publicações indexadas pelo GS tem aumentado consideravelmente desde que foi lançado em 2004 [Harzing, 2014, Harzing, 2013]. Experimentos apresentados por [Harzing, 2014] mostram que o GS aumentou a sua cobertura e estabilidade, e hoje é muito mais adequado para fins de pesquisas de avaliação e pesquisas bibliométricas do que era no passado.

Atualmente, existem várias ferramentas e serviços que são criados para coletar dados do GS e apresentá-los com algum tipo de processamento, como por exemplo, calcular indicadores cienciométricos para conferências e artigos. Dentre eles, destacamos alguns mais utilizados como o *Scholarmeter*[3], o *My Citation* do próprio *Google*, o *SHINE*, e o *Publish or Perish (PoP)* [Harzing, 2007]. Nas próximas subseções essas soluções serão descritas, bem como a máquina de busca acadêmica *CiteSeerX* [Giles et al., 1998] que embora não utilize metadados do GS, também é bastante utilizada pela comunidade científica de computação para obter índices de impacto de conferências.

2.2.1 Scholarometer

O Scholarometer é uma ferramenta social desenvolvida para facilitar a análise de citações e ajudar a avaliar o impacto dos autores dos artigos [Kaur et al., 2014]. Os serviços oferecidos por essa ferramenta permitem que os pesquisadores possam computar várias métricas de impacto baseadas em citação utilizando uma abordagem baseada em *crowdsourcing*.

Para [Kaur et al., 2014], a abordagem baseada em *crowdsourcing* torna a ferramenta mais poderosa, uma vez que pesquisadores podem colaborar entre si ao marcar ou desmarcar artigos de autores de várias áreas como pertencentes ou não a eles. O modelo tem a vantagem de que, quando combinado com os dados de citação, pode permitir a coleta de dados estatísticos necessários para o cálculo de métricas de impacto interdisciplinares.

Deste modo em março de 2014, as consultas emitidas ao *Scholarmeter* resultaram em

[3]http://scholarometer.indiana.edu

uma base de dados de citação de cerca de 39000 autores de 2.8 milhões de artigos em 2400 áreas da ciência [Kaur et al., 2014]. No entanto, essa ferramenta requer desambiguação manual dos metadados dos artigos, que é feita pelos usuários da ferramenta, o que torna seu uso inviável quando o objetivo é manter os dados em grande escala para o cálculo preciso das métricas de impacto de conferências. A Figura 2.2 apresenta a interface do Scholarmeter para uma consulta ambígua, solicitando o *feedback* do usuário para determinar se todos os artigos apresentados são de fato pertencentes a esse autor.

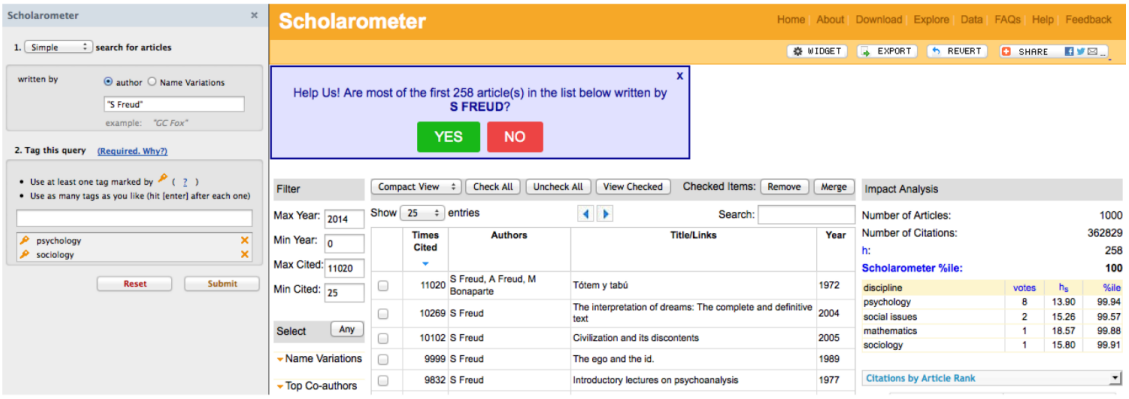


Figura 2.2: Estimativa do Scholarmeter para uma consulta ambígua.

O *Live SHINE* por sua vez, utiliza a abordagem *crowndsourcing* apenas na coleta dos metadados dos artigos e dados de citação a partir do *Google Scholar (GS)*, e não na filtragem dos artigos. Nossa ferramenta conta com o *SHINER*, um classificador automático que é capaz de separar os metadados dos artigos que pertencem ou não a conferência desejada, ou seja, não é preciso realizar a desambiguação dos metadados manualmente, tornando o cálculo do *h-index* mais preciso e menos vulnerável a ruídos provocados pela interação do usuário no processo de filtragem.

2.2.2 My Citations

Sete anos após o lançamento do *Google Scholar* em 2004, ele foi reforçado com um novo módulo, o *Google Scholar Author Citation Tracker (GSACT)*, atualmente um pequeno subconjunto da base de dados do *Google Scholar (GS)*. Ele permite que os pesquisadores criem e editem os seus perfis científicos com dados pessoais (nome, afiliação, disciplinas de interesse), uma bibliografia (automaticamente criada a partir de uma lista limpa de resultados do GS), e alguns indicadores bibliométricos, tais como o *h-index*, contagem de citações, e o i10-index (o número de publicações do autor que receberam pelo menos 10

citações) [Jacsó, 2012].

Esse novo serviço, também chamado de *My Citations*[4], fornece essas métricas para toda a carreira acadêmica dos autores e para o período correspondente aos últimos 5 anos. O módulo também oferece serviços com opções essenciais para o pesquisador, tais como a ordenação das listas de resultados dos artigos por seu ano de publicação, título e citações recebidas [Jacsó, 2012].

No entanto, o *My Citation*, apesar de ser um serviço do próprio GS, carece de registros de citações de cerca de 75% dos pesquisadores do GS, o que torna difícil a indicação do seu uso [Houzanme, 2012]. Segundo [Houzanme, 2012], uma possível explicação é que esses pesquisadores ainda não se inscreveram no módulo *My Citation*. Acredita-se que eles não perceberam grandes benefícios pessoais, embora como podemos verificar na Figura 2.3, sua participação o ajudaria a verificar os benefícios gerados a partir do acesso de suas publicações e perfil no *My Citation*, como por exemplo, acompanhar a evolução das citações de seus próprios artigos.

O *Live SHINE* não tem por objetivo mostrar métricas para autores e não requer um cadastro de usuário para se obter metadados e dados de citações de artigos, para então gerar índices de impacto de conferências. Nossa ferramenta utiliza puramente o motor de busca do GS para obter a lista de artigos e os dados de citações. Além disso, nossa ferramenta não limita o cálculo de *h-index* para o período dos últimos 5 anos, o usuário pode determinar a janela de tempo para o qual deseja obter os índices de impacto da conferência.

2.2.3 SHINE

A ferramenta denominada *Simple H-Index Estimation (SHINE)* é parte de uma iniciativa conduzida pela Sociedade Brasileira de Computação (SBC) em parceria com o Instituto de Computação (IComp) da Universidade Federal do Amazonas (UFAM)[5]. Seu objetivo é disponibilizar um mecanismo verificável de medição de impacto das principais conferências de Ciência da Computação através da métrica *h-index*.

Atualmente o SHINE mantém cerca de 1800 conferências e 800 mil artigos cadastrados em sua base de dados. Nesta ferramenta, as listas de artigos de cada conferência

[4]https://scholar.google.com.br/citations
[5]http://shine.icomp.ufam.edu.br/about.php

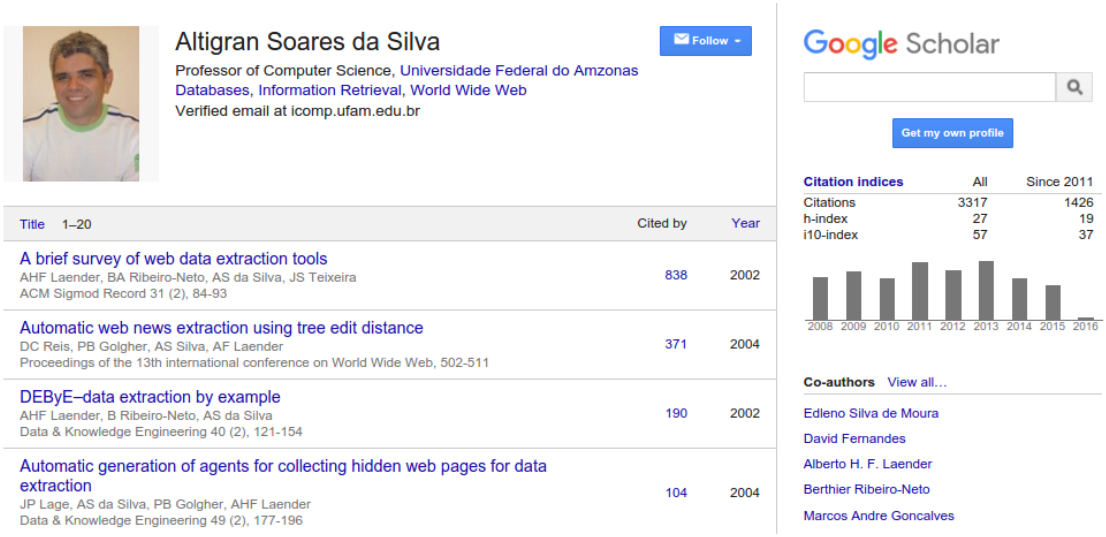


Figura 2.3: Exemplo de perfil do My Citations.

é coletada manualmente a partir de várias fontes online e bibliotecas digitais tais como ACM Digital Library, IEEE Xplore e DBLP[6], enquanto que os dados de citação de cada artigo é obtido a partir de consultas com os nomes dos artigos emitidas ao motor de busca do *Googgle Scholar (GS)*.

Os responsáveis pelo projeto SHINE tentaram manter sua base de dados sempre atualizada. No entanto, coletar manualmente a lista de artigos das conferências é ineficiente e inviável, visto que todos os anos acontecem novas edições das conferências, bem como os artigos podem receber novas citações. Por conta disso, a ferramenta precisa manter um constante monitoramento das conferências a partir de muitas fontes online distintas, para então coletar os dados de citação dos novos e antigos artigos no GS, bem como atualizar as citações dos artigos já conhecidos, e isso representa um trabalho manual extremamente custoso e desafiador.

Dessa forma, a ferramenta hoje se encontra desatualizada e gera índices de impacto que não representam o real impacto das conferências. O SHINE apresenta estimativas de impacto precisas para os anos de 2000 à 2012. No entanto, para os anos posteriores a 2012, a ferramenta apresenta índices imprecisos e subestimados, exigindo constantemente a necessidade de complementar e atualizar as informações fornecidas pela ferramenta [Alves et al., 2013, Lima et al., 2013, Vasilescu et al., 2013]. Podemos ver por meio da Figura 2.4 uma consulta pelo *h-index* de uma conferência que tem edições e artigos publicados entre 2013 e 2015 e que a ferramenta não apresenta listas de artigos para

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

este período.

Ao contrário do que acontece nessa ferramenta, a abordagem que adotamos no *Live SHINE* tenta obter automaticamente as listas de artigos bem como seus respectivos metadados e dados de citação para cada conferência. O processo de filtragem automática dos metadados é feito pelo método *SHINER*, que garante que o *h-index* seja calculado com a lista de artigos mais correta possível.

Figura 2.4: Estimativa do SHINE desatualizada para o período entre 2013 e 2015.

2.2.4 Publish or Perish

O *Publish or Perish (PoP)* é um programa desenvolvido por [Harzing, 2007, Harzing, 2010] capaz de calcular índices de impacto e outros indicadores cienciométricos a partir de metadados de artigos coletados a partir do *Google Scholar (GS)* [Harzing, 2010].

Este programa é capaz de gerar métricas tais como *h-index* e *g-index*, médias de citações por artigo, citações por autor, artigos por autor e citações por ano, além de algumas variações de *h-index* como o individual, a média anual de crescimento e outros [Harzing, 2007].

O PoP fornece uma interface onde o usuário pode digitar o nome de uma determinada conferência, que é então passada ao motor de busca do GS como uma consulta. Em seguida, o programa apresenta uma lista de artigos e estatísticas de citação com base no conjunto de resposta retornado pelo GS para a consulta submetida.

No entanto, uma consulta que contém o nome e/ou sigla de uma determinada conferência pode retornar artigos de muitos outros veículos. [Jacsó, 2009] apresenta um exemplo onde a metade dos 1000 artigos retornados pelo GS (o número máximo que o GS apresenta) não pertence a conferência desejada, de modo que o número de publicações e indicadores relativos ficaram altamente inflados. O GS não oferece uma opção para os usuários corrigirem e filtrarem esses resultados imprecisos. Na Figura 2.5, podemos observar que a interface do PoP permite que os usuários removam esses registros errados, no entanto, este procedimento é completamente manual e, portanto, não é escalável.

Para resolver esse problema, nossa ferramenta proposta *Live SHINE* utiliza um método previamente treinado que identifica de maneira completamente automática, os artigos que de fato pertencem a conferência desejada, e então considera apenas esses artigos para o cálculo dos índices de impacto. Por se tratar de uma abordagem automática, nossa ferramenta é capaz de entregar valores de *h-index* mais precisos com um processo de filtragem e coleta colaborativa de metadados que pode ser facilmente escalável.

2.2.5 CiteSeerX

O *CiteSeerX* [Giles et al., 1998, Caragea et al., 2006] é uma biblioteca digital e motor de busca, desenvolvida originalmente por [Giles et al., 1998], que fornece indexação de citações e acesso livre para artigos técnicos e científicos. Ele é focado principalmente na

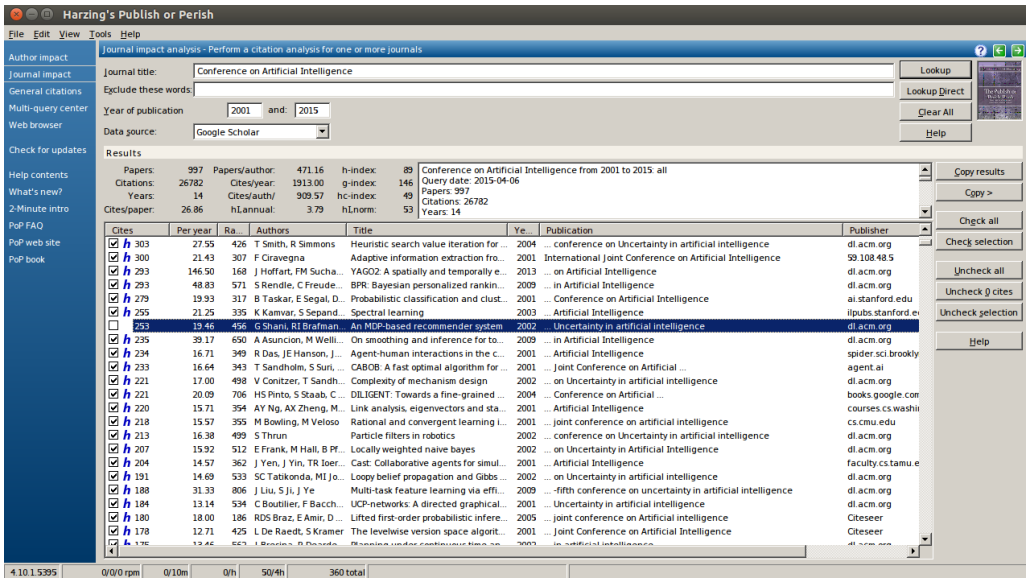


Figura 2.5: Estimativa do Publish or Perish com filtragem manual do usuário.

literatura das áreas de Ciência da Computação e de informação, e seu objetivo geral[7] é melhorar a disseminação da literatura científica além de proporcionar melhorias na funcionalidade, usabilidade, disponibilidade, custo, abrangência, eficácia e pontualidade no acesso ao conhecimento científico e acadêmico.

Esse motor de busca rastreia continuamente documentos em formato *PDF* e *PostScript* que estão disponíveis publicamente na Web, e indexa o texto completo dos artigos, juntamente com seus metadados e dados de citação. No entanto, arquiteturas de rastreamento tradicionais não podem ser facilmente escaladas e nem podem acompanhar o crescimento rápido, em grande escala, e heterogêneo de grandes bases de dados acadêmicas.

Devido aos desafios desta tarefa, em 2014 o *CiteSeerX* teve uma cobertura de apenas cerca de 10% dos documentos indexados pelos motores de busca acadêmicos *Google Scholar (GS)* e *Miscrosoft Academic Search (MAS)* [Wu et al., 2014]. Neste trabalho, nós evitamos esses desafios e resolvemos o problema da cobertura de artigos ao utilizar a base de dados do GS, enviando consultas ao seu motor de busca acadêmico para recuperar os metadados e dados de citação dos artigos.

[7] http://csxstatic.ist.psu.edu/about

Capítulo 3

Visão Geral

Este capítulo apresenta com mais detalhes os principais problemas enfrentados por pesquisadores na tarefa de estimar o impacto de conferências de Ciência da Computação. Em seguida, apresenta uma visão geral de como propomos superar tais problemas.

3.1 Descrição do Problema

A forma mais comum de estimar os índices de impacto de conferências, é através de métricas baseadas em citação [Laender et al., 2008, Zhuang et al., 2007, Chiu & Fu, 2010, Bornmann & Daniel, 2005, Bornmann & Daniel, 2007], como por exemplo o *Impact Factor (IF)* [Garfield, 1972] e o *H-Index* [Hirsch, 2005]. No entanto, para que essas métricas sejam geradas com a precisão adequada que essa tarefa exige, dois desafios devem ser superados: (*i*) obter as listas de artigos corretas das conferências e (*ii*) obter dados de citações desses artigos atualizados. Por não terem resolvido adequadamente esses desafios, as ferramentas mais utilizadas para se obter índices de impacto de conferências de Ciência da Computação apresentam problemas e limitações, e por isso geralmente não concordam entre si com relação aos índices obtidos para uma mesma conferência/ano, o que diminui a confiança dos pesquisadores nessas ferramentas.

A seguir, esses dois problemas são descritos com mais detalhes.

3.1.1 Listas de Artigos Incorretas

Considere um cenário em que um pesquisador precisa estimar o impacto de uma determinada conferência. Uma vez que este impacto é muitas vezes estimado pelo número de

citações dos artigos publicados na conferência, o pesquisador opta por usar uma máquina de busca acadêmica para recuperar a lista de registro de metadados dos artigos desta conferência específica. O pesquisador submete uma consulta para esta máquina de busca, de modo que cada artigo que pertence à conferência alvo e está disponível no seu índice aparece no conjunto de resposta. Esta tarefa é muitas vezes realizada por meio de ferramentas como o *Publish or Perish (PoP)* [Harzing, 2007].

No entanto, as consultas submetidas pelo pesquisador podem conter termos que também correspondem a registros de outras conferências. De modo que é frequente o caso em que alguns registros retornados como resposta não se referem a artigos da conferência-alvo. Neste caso, o usuário teria que percorrer os resultados retornados e selecionar somente os que se referem de fato à conferência. Muitas vezes, isso pode exigir que o usuário pesquise manualmente em uma lista de milhares de respostas para encontrar apenas os poucos registros que são de fato relevantes para o cálculo dos índices de impacto, o que representa um trabalho extremamente custoso. Esse trabalho passa a ser ainda maior nos casos onde esta tarefa necessita ser executada para muitas conferências.

A Figura 3.1 ilustra mais precisamente o problema. Considere que R_C é o conjunto de registros correspondentes a todos os artigos de uma determinada conferência C disponível no índice de uma máquina de busca acadêmica. Observe que podem haver casos em que o índice não inclui todos os artigos da conferência C. Além disso, considere que R_Q é um conjunto de registros retornados de uma consulta Q, submetida a esta máquina de busca e cuja intenção é recuperar os registros correspondentes a artigos de C. Neste trabalho, abordamos o problema de encontrar $R'_C = R_C \cap R_Q$, isto é, queremos encontrar, entre os registros de R_Q, somente aqueles que realmente pertencem ao conjunto R_C de registros de artigos da conferência C, sem conhecer antecipadamente o conjunto R_C.

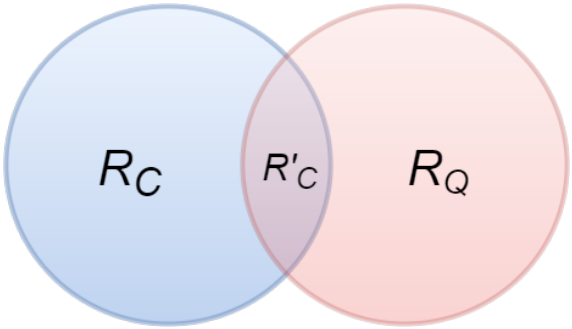


Figura 3.1: Visão geral do problema.

Para exemplificar esse problema, considere os artigos publicados na *International So-*

ciety for Music Information Retrieval Conference (ISMIR) em 2010. São obtidos para esse ano da conferência valores de *h-index* discrepantes através de diferentes ferramentas. Para este exemplo, o *Publish or Perish (PoP)* calculou um valor de *h-index* 55 quando a consulta foi feita mediante o preenchimento do campo de pesquisa genérica *("all the words")*, e calculou um valor de *h-index* 7, quando o campo específico *("publication")* foi utilizado. O *Google Scholar Citations (GSC)* por sua vez calculou um valor de *h-index* 10, enquanto o SHINE relatou um valor de *h-index* 4 para essa mesma conferência e ano .

Notamos que todas as quatro alternativas relataram valores de *h-index* discrepantes, apesar do fato de que todas elas utilizam os mesmos dados de citações fornecidos pelo *Google Scholar* para cada artigo. Observamos que a diferença de valores de *h-index* encontrados no exemplo acima é devido o fato de que cada ferramenta considera um conjunto diferente de artigos como pertencentes a conferência ISMIR em 2010. Na realidade, com base nas informações disponíveis no site da edição de 2010 da conferência ISMIR[1], verificamos para cada artigo o número de citações fornecidas pelo GS, e calculamos manualmente o valor de *h-index* como sendo 22.

Para ilustrar o quão desafiador é obter a lista correta de artigos de uma determinada conferência/ano, realizamos a mesma consulta em diferentes máquinas de busca acadêmicas bem conhecidas, a fim de obter a lista de artigos que aparecem na edição de 2010 do ISMIR, e obtivemos resultados que estão abaixo do esperado. A Tabela 3.1 mostra um resumo desses resultados. Como referência, essa tabela utiliza a lista correta dos 114 artigos disponíveis no site oficial da conferência para 2010.

	Resultados	**Corretos**	**Incorretos**	**Precisão**	**Revocação**
CSX	5	4	1	0.80	0.03
MAS	1	1	0	1.00	0.01
GS-D	1000	97	911	0.10	0.85
GS-A	20	11	9	0.55	0.09

Tabela 3.1: Análise da lista de artigos obtidos para a conferência ISMIR 2010 a partir de diferentes máquinas de busca.

Na Tabela 3.1, cada linha mostra os resultados obtidos usando as seguintes ferramentas/serviços: *CiteSeerX (CSX)*, *Microsoft Academic Search (MAS)*, *Google Scholar - Interface de Busca Padrão (GS-D)* e *Google Scholar - Interface de Busca Avançada (GS-A)*. Em todos os casos, construímos a consulta com o nome completo da conferên-

[1] http://ismir2010.ismir.net/

cia e o ano desejado, no caso 2010. Consideramos que tais consultas seriam tipicamente especificadas por usuários comuns. No caso do GS-A, o usuário pode preencher alguns campos de forma a melhorar a precisão da consulta. Neste caso, preenchemos o campo de pesquisa *"Exibir artigos publicados em"* com o nome completo da conferência e o campo *"Exibir artigos com data entre"* com o ano 2010.

Como pode ser notado, o CSX e o MAS alcançaram boa precisão, no entanto, em contra partida sua revocação foi muito baixa, tornando seus resultados quase inúteis. Além disso, vale a pena destacar que o MAS não atualiza a sua base de dados desde 2013 [Noorden, 2014].

Para o *Google Scholar (GS)*, a consulta utilizando o GS-D gerou 1000 resultados acessíveis ao usuário, embora o site reporte que em sua base existe um número muito maior de resultados, mais precisamente 10600, que correspondem a consulta submetida. Essa limitação é imposta pela máquina de busca. Neste caso, o GS atingiu 0.85 de revocação, ou seja, a maioria dos artigos da conferência para esse ano foi retornada. No entanto, apenas 97 dos 1000 resultados eram de fato do ISMIR 2010, isto é, alcançou um valor de precisão de apenas 0.10. Já a consulta utilizando o GS-A, gerou apenas 20 resultados, e apenas 11 deles estavam de fato corretos, o que resultou em valores baixos de precisão (0.55) e revocação (0.09).

Os resultados obtidos através do GS-D, que consideramos a interface de consulta mais utilizada pelos pesquisadores, são detalhados na Figura 3.2, onde mostramos a porcentagem de registros positivos (+), ou seja, aqueles que verdadeiramente correspondem a artigos do ISMIR 2010, e registros negativos (-) em cada intervalo do *ranking* de resultados fornecido pelo GS. Como pode ser observado, a maioria dos 100 registros melhor classificados é realmente positiva. No entanto, uma porção significativa deles é negativa. Assim, cabe ao usuário identificar esses registros negativos. Além disso, há muitos registros positivos espalhados em muitas das posições mais baixas do ranking, misturados com outros registros negativos. Como a tarefa é a obtenção de todos os artigos, mais uma vez, cabe ao usuário navegar pelo ranking para selecionar os registros positivos.

Notamos que os resultados obtidos foram muito longe do esperado, apesar do fato de que esta conferência em particular tem o seu conjunto completo de artigos disponibilizados *online*. Os resultados são similares ou piores para conferências que não publicam seus metadados *online*. Na verdade, realizamos testes similares com outras conferências

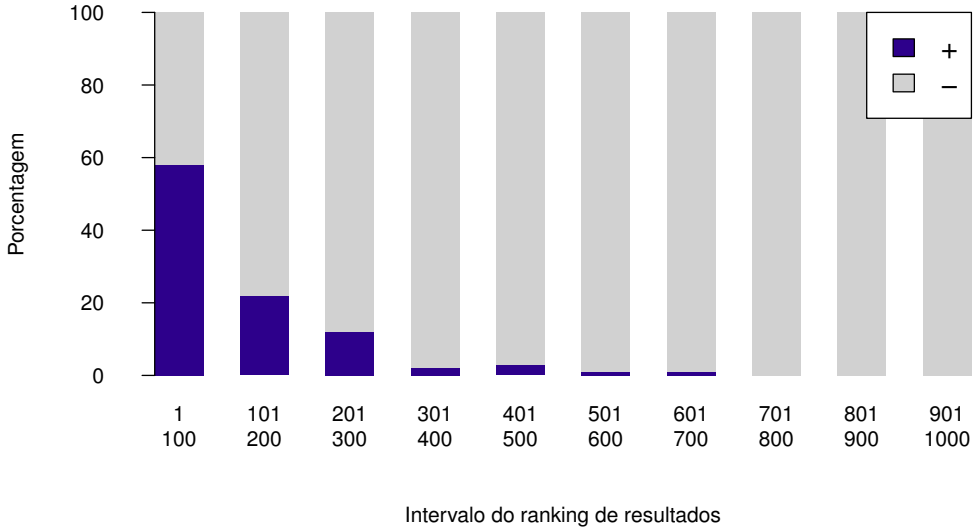


Figura 3.2: Resultados do GS-D para ISMIR 2010

da área de Ciência da Computação, e obtivemos resultados semelhantes para muitas delas. Estes resultados serão melhor discutidos em nossos experimentos apresentados no Capítulo 5.

As observações acima podem ajudar a entender o porque de as ferramentas atuais mais utilizadas que são baseadas em máquinas de busca acadêmicas, como o *Publish or Perish (PoP)*, *Scholarometer* e o serviço *My Citation* do próprio do *Google Scholar (GS)*, não fornecerem uma solução adequada para estimar índices de impacto de conferências, como discutido por [Houzanme, 2012] e [Di Iorio et al., 2015].

É muito importante ressaltar que o GS é uma máquina de busca acadêmica cujo objetivo é obter como respostas os documentos que correspondem (em geral) à consulta emitida por um usuário. Ou seja, o GS não foi concebido para atacar os desafios específicos que motivam este trabalho, e não se espera que ele por si só supere esses desafios. Nossa proposta neste trabalho consiste em aproveitar este recurso *online* inestimável para resolver adequadamente a nossa tarefa de avaliar o impacto de conferências.

Uma alternativa interessante para obter o conjunto completo e correto de artigos com dados de citações para estimar o impacto seria utilizando bibliotecas digitais de editoras como a *ACM Digital Library* e *IEEE Explore*. No entanto, tais bibliotecas digitais, muitas vezes incluem um conjunto restrito de conferências em suas bases de dados. Por exemplo, nenhuma destas duas bibliotecas digitais incluem a conferência ISMIR do nosso exemplo.

Além disso, como no caso do JCR, estas bibliotecas também monitoram citações em uma coleção fechada de veículos, o que pode prejudicar o cálculo correto das métricas de impacto. Deste modo, utilizar bibliotecas digitais para a obtenção do conjunto completo de metadados dos artigos e dados de citações de uma determinada conferência, em geral, não é uma alternativa viável.

Uma outra alternativa seria os *websites* de conferências específicas, que mantêm as listas corretas de artigos de cada uma das suas edições, como no caso da conferência ISMIR. No entanto, isso exige monitorar inúmeros *websites* e extrair seus conteúdos, que pode resultar em uma carga considerável de desenvolvimento específico. Além disso, esses *websites* muitas vezes não possuem informações sobre citações dos artigos.

3.1.2 Dados de Citações Desatualizados

Manter atualizado os dados de citações de todos os artigos de uma conferência é um problema extremamente difícil de lidar, pois isso requer um rastreamento constante de um número desconhecido de conferências, visto que a cada ano os artigos já conhecidos podem receber novas citações bem como novos artigos podem ser publicados em novas edições dessa conferência. Além disso, todos os anos podem surgir novas conferências, algumas podem deixar de existir e outras podem se juntar a conferências maiores, o que aumenta ainda mais o desafio de manter sempre atualizado esse rastreamento de artigos e os dados de citações.

Outro grande problema é fato de que muitas conferências não disponibilizam as listas de artigos publicados na Web, no entanto, quando disponibilizam, esse processo não é feito através de em um único local padrão, ou em uma única biblioteca digital oficial, muitas dessas conferências costumam publicar seus artigos nos seus próprios *websites*. Deste modo, para se manter os dados dos artigos e citações sempre atualizados, é necessário buscar essas informações em um número desconhecido de fontes espalhadas pela Web. Por isso, para superar esses desafios, as ferramentas mais utilizadas para estimar o impacto de conferências como o *Publish or Perish (PoP)* e o *SHINE* tomam proveito das informações fornecidas pelo motor de busca acadêmico *Google Scholar (GS)*, de modo a transferir a função de rastreamento dos artigos e a contagem de citações para essa ferramenta.

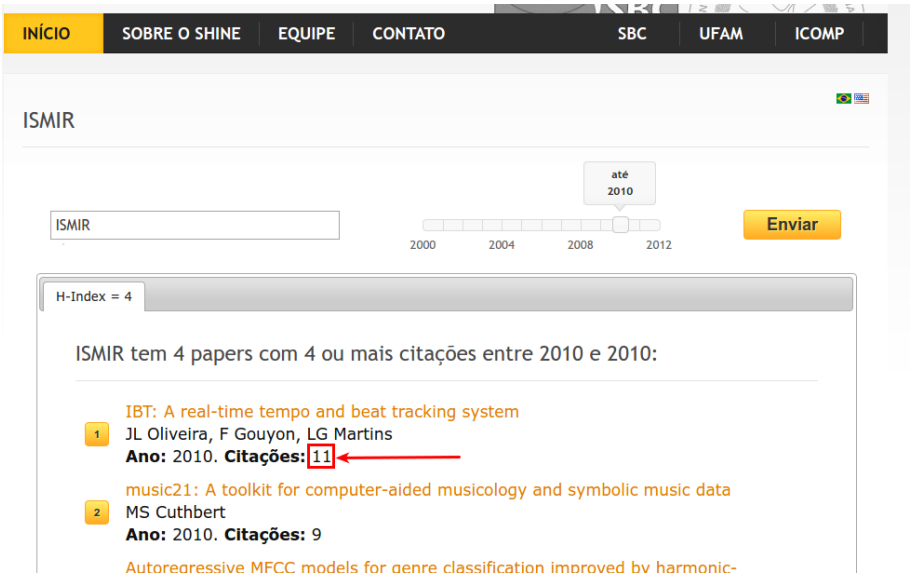
Atualmente, as máquinas de busca acadêmicas como o *Google Scholar (GS)*, facilitam

o acesso a informações de artigos e citações. Essas ferramentas são capazes de rastrear na Web uma quantidade significativa de informações a respeito de artigos e veículos científicos, e além disso, são capazes de realizar a contagem automática das citações desses artigos. Por isso, essa ferramenta é muito utilizada como fonte de informação para gerar índices de impacto de conferências. Estudos como os apresentados por [Harzing, 2014] mostram que o GS nos últimos anos aumentou consideravelmente sua cobertura, e deste modo, se tornou uma fonte adequada e promissora para pesquisas de avaliação de conferências e pesquisas bibliométricas.

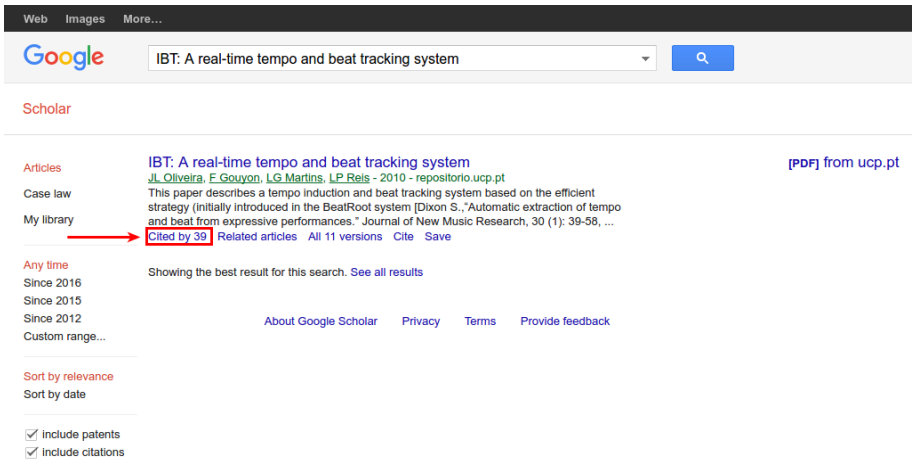
No entanto, utilizar as informações fornecidas pelo GS para estimar o impacto de conferências não é garantia de precisão nos índices que são gerados. Como discutido anteriormente, as respostas obtidas por consultas a esse motor de busca acadêmico podem conter informações erradas e precisam ser filtradas. Além disso, o número de citações dos artigos varia constantemente todos os anos, e exige que esses dados sejam constantemente verificados junto ao GS. Esse problema pode ser notado através da ferramenta SHINE, que embora tenha utilizado dados de citações fornecidos pelo GS a partir de consultas emitidas ao motor de busca com os títulos dos artigos, desde 2012 não realiza novas consultas a fim de atualizar sua base de dados, e hoje apresenta dados de citações de artigos inferiores aos apresentados atualmente pelo GS.

Para exemplificar, a Figura 3.3a mostra que o artigo *"IBT: A real-time tempo and beat tracking system"*, da conferência *International Society for Music Information Retrieval Conference (ISMIR)* em 2010, possui 11 citações segundo a ferramenta SHINE, enquanto que a Figura 3.3b mostra que o mesmo artigo contém na verdade 39 citações, conforme consulta ao GS feita em março de 2016. Essa diferença de dados de citações faz com que o SHINE entregue aos pesquisadores índices de impacto subestimados e sem precisão para as conferências cadastradas em sua base de dados.

Utilizar uma lista incorreta de artigos, bem como dados de citações desatualizados para estimar o impacto de uma determinada conferência são problemas graves que prejudicam a precisão dos índices gerados. Como vimos neste capítulo, as ferramentas mais utilizadas para estimar o impacto de conferências não conseguem resolver de maneira satisfatória esses problemas, e geram índices de impacto discrepantes para uma mesma conferência/ano. Neste trabalho, procuramos desenvolver uma solução automatizada para estes problemas como uma forma de auxiliar os pesquisadores da comunidade de Ciência



(a) Ferramenta SHINE apresentando 11 citações para o artigo "IBT: A real-time tempo and beat tracking system".



(b) Google Scholar apresentando, em março de 2016, 39 citações para o mesmo artigo, contrariando a informação da Ferramenta SHINE.

Figura 3.3: Diferença de citações entre o SHINE e o Google Scholar para um mesmo artigo.

da Computação na realização de tais tarefas.

3.2 Proposta

Neste trabalho propomos uma ferramenta denominada *Live SHINE*, cujo objetivo é gerar índices de impacto de alta precisão de conferências da área de Ciência da Computação. Nossa solução, concebida em forma de extensão para o *Navegador Web Google Chrome*[2], funciona sobre o site do *Google Scholar* a fim de aproveitar os dados de artigos e citações

[2]https://www.google.com.br/chrome/browser/desktop/

fornecidos por essa máquina de busca acadêmica. Seu uso busca auxiliar os pesquisadores a lidar de maneira satisfatória com os desafios envolvidos na tarefa de estimar o impacto de conferências. Ao contrário das ferramentas mais utilizadas para esse fim, nossa proposta filtra automaticamente as informações obtidas a partir do GS, e considera no cálculo das métricas apenas os dados de artigos e citações que de fato pertencem a uma determinada conferência/ano. Os experimentos apresentados no Capítulo 5 mostram que o *Live SHINE* é capaz de entregar aos pesquisadores índices de impacto precisos sobre conferências de Ciência da Computação.

Todavia, utilizar o GS como fonte de dados de artigos e citações não é garantia de precisão nas métricas de impacto geradas. Vimos que as ferramentas baseadas neste motor de busca acadêmico dificilmente concordam com os índices calculados para uma mesma conferência/ano. Isto acontece principalmente devido o fato do GS não ter sido concebido originalmente para atender essa tarefa, e por isso apresenta dois grandes desafios que devem ser superados. Em nossa proposta, tentamos resolver de maneira satisfatória e automática esses desafios que consideramos os mais importantes para gerar índices de impacto de conferências com precisão a partir do GS. Esses desafios estão listados abaixo:

> **Desafio 1:** Gerar as métricas de impacto com base nas listas corretas de artigos publicados para uma determinada conferência/ano. Como discutido anteriormente, essas informações estão espalhadas pela Web e exige a busca contínua dessas informações em um número desconhecido de fontes Web. Embora o *Google Scholar* tenha se tornado uma fonte promissora de dados de artigos e citações de conferências, com uma cobertura que cresce todos os anos, uma consulta emitida ao seu motor de busca com o objetivo de trazer registros de artigos de uma determinada conferência/ano não garante uma alta precisão no conjunto de resultados retornado, pois o GS retorna também registros referentes a artigos de outros veículos relacionados. Para tentar resolver esse problema, o *Live SHINE* realiza um filtro automático dos metadados dos registros de artigos retornados nas páginas de resposta do GS. Para realizar tal tarefa, nossa proposta utiliza um método denominado *SHINER*, desenvolvido com base em técnicas de aprendizagem de máquina. Esse método é capaz de navegar e selecionar automaticamente os dados de registros que realmente pertencem a conferência/ano alvo dentro do conjunto de resposta retornado pelo GS, e assim considera no cálculo a lista de artigos mais correta possível para gerar

índices de impacto precisos, ou seja, com base apenas nos dados de artigos que de fato pertencem a conferência/ano.

Desafio 2: Gerar as métricas de impacto com base em dados de citações de artigos sempre atualizados. Todos os anos, esse cenário de citações de artigos muda constantemente, pois os artigos podem receber novas citações, novos artigos podem surgir, novas conferências podem surgir ou deixar de existir. O GS lida bem com o monitoramento dos artigos e das conferências, por isso consegue entregar uma contagem de citações que está sempre sendo atualizada. No entanto, essa atualização contínua exige que os dados de citações sejam constantemente verificados junto a esse motor de busca acadêmico, antes que sejam utilizados nos cálculos dos índices de impacto, pois por mais que a lista de artigos esteja correta, o uso de dados de citações desatualizados pode gerar métricas subestimadas. Para tentar resolver esse problema, a ferramenta *Live SHINE*, por meio de sua interface de consulta, realiza uma coleta colaborativa dos dados de citações dos artigos no momento que o pesquisador utiliza a ferramenta para obter as métricas de impacto de uma conferência/ano, e em seguida armazena essas informações. Toda vez que a ferramenta é utilizada, esses dados de citações são verificados e atualizados caso seja necessário. Deste modo, o *Live SHINE* realiza a estimativa do impacto das conferências apenas com dados de citações sempre atualizados.

Nos próximos capítulos, serão apresentados mais detalhes dessas estrategias desenvolvidas em nossa proposta para tentar superar tais desafios.

Capítulo 4

SHINER - Filtragem de Metadados de Artigos

Este capítulo apresenta uma descrição detalhada sobre o SHINER, o método proposto neste trabalho para filtrar automaticamente os metadados fornecidos por um motor de busca acadêmico. A seguir, é apresentada uma visão geral do funcionamento do método, os algoritmos para construção e aplicação dos classificadores, e por fim as estratégias utilizadas nas consultas emitidas ao *Google Scholar*.

4.1 Visão Geral

O *SHINER - **S**imple **H-IN**dex **E**stimator **R**elayer* é o nosso método desenvolvido com base em técnicas de aprendizagem de máquina, capaz de filtrar automaticamente os metadados obtidos a partir de uma máquina de busca. Seu objetivo é auxiliar os pesquisadores na tarefa de obter as listas corretas de artigos das conferências da área de Ciência da Computação, e desse modo possibilitar uma estimativa mais precisa do impacto dessas conferências. Neste trabalho, o SHINER é empregado como parte da solução Live SHINE que será descrita no Capítulo 7.

A Figura 4.1 apresenta uma visão geral do funcionamento do SHINER. Dado uma conferência e um ano de interesse, primeiramente o SHINER submete uma consulta a uma máquina de busca acadêmica a fim de alcançar uma alta revocação dos artigos de uma conferência/ano nas primeiras páginas. Uma vez que a consulta é submetida, é feita uma navegação através das páginas de resultados da máquina de busca para coletar todos

os resultados retornadas para essa consulta. A saída de uma máquina de busca acadêmica para uma determinada consulta é um conjunto de *snippets* que contêm dados vindos de vários tipos de documentos, tais como: artigos de diferentes veículos, relatórios técnicos, teses, livros, etc. Depois de coletada, essa saída é automaticamente filtrada através de um classificador cujo objetivo é identificar quais *snippets* referenciam artigos da conferência alvo. Ao identificar esses *snippets*, o classificador gera a lista de artigos mais correta possível em termos de precisão, sem perder a revocação alcançada pela máquina de busca acadêmica.

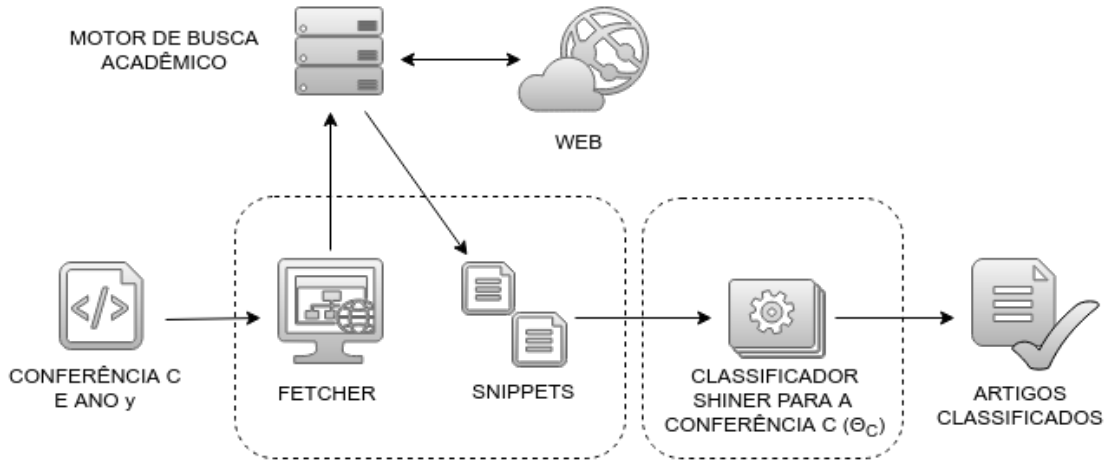


Figura 4.1: Visão geral do método SHINER.

Como a Figura 4.1 mostra, o método SHINER emprega um classificador para cada conferência de interesse. Esse método utiliza uma abordagem de aprendizagem de máquina para filtrar os resultados retornados pela máquina de busca acadêmica. Em termos formais, cada um desses classificadores pode ser definido como segue:

Definição 1. *Seja R_Q o conjunto de snippets retornados por um determinado motor de busca acadêmico para uma consulta Q, que foi emitida para recuperar os registros correspondentes de uma determinada conferência C e ano y. Um classificador SHINER para a conferência C é uma função binária $\Theta_C(y) : R_Q \rightarrow \{+,-\}$, que recebe como parâmetro um ano y, e atribui um valor de positivo* (+) *ou negativo* (-) *para cada snippet de R_Q. Se o valor atribuído é positivo* (+)*, então o snippet se refere a um artigo publicado na conferência C e ano y. Se o valor atribuído é negativo* (-)*, então o snippet não se refere a um artigo publicado nessa conferência/ano.*

A seguir apresentamos em detalhes os passos para a construção e o uso dos classifica-

dores Θ_C.

4.2 Construção dos Classificadores

O Algoritmo 1 apresenta o algoritmo que desenvolvemos para construção de um classificador Θ_C para uma determinada conferência C. Esse algoritmo recebe como parâmetros a conferência alvo C, um ano y, e um conjunto de dados de treino $papers[C,y]$ contendo a lista completa de artigos publicados em C no ano y. Bibliotecas Digitais, como *DBLP*, *IEEE*, *ACM*, ou até mesmo o *website* oficial de uma edição da conferência, constituem boas fontes para se obter a lista de artigos para compor os dados de treino $papers[C,y]$. Utilizando técnicas de aprendizagem de máquina, o Algoritmo 1 utiliza a lista de artigos passado como parâmetro para construir um classificador que é capaz de filtrar os *snippets* de qualquer edição da conferência. A seguir esse algoritmo é descrito em detalhes.

Algoritmo 1 Algoritmo para construir um classificador Θ_C.

Input: $papers[C,y], C, y$
Output: a classifier Θ_C for the conference C

```
1:  Q ← buildQuery(C, y)
2:  R_Q ← fetchResults(Q)
3:  foreach (s ∈ R_Q) do
4:      if (belongsTo(s, papers[C, y])) then
5:          s.label ← +
6:      else
7:          s.label ← -
8:      end if
9:      s.bow ← getBoW(s)
10:     s.bow ← s.bow − {y}
11:     s.bow ← nGram(s.bow)
12:     s.bow ← fSelection(s.bow)
13: end foreach
14: Θ_C ← buildClassifier(R_Q)
15: return Θ_C
```

O algoritmo inicia com a submissão de uma consulta Q para a máquina de busca acadêmica especificando a conferência C e ano y. Esse passo é feito na Linha 1 do Algoritmo 1 através da função $buildQuery()$. A consulta Q deve ser capaz de recuperar uma alta porcentagem dos artigos dessa conferência e ano entre as primeiras páginas de

resultados, embora muitos documentos não relevantes também possam aparecer ao longo das demais páginas. Na Seção 4.4 são descritas as estratégias utilizadas para a seleção da consulta que é empregada nesse passo.

O próximo passo é submeter a consulta Q à máquina de busca acadêmica, incluindo todos os resultados da consulta no conjunto R_Q. Esse passo é feito na Linha 2 do algoritmo através da função $fetchResults()$, que recebe a consulta Q como o único parâmetro. Observe que todos os resultados da consulta devem ser salvos em R_Q, de modo que a função $fetchResults()$ deve ser capaz de navegar através das páginas de resultados da máquina de busca acadêmica para acessar e salvar cada *snippet* recuperado.

Entre as Linhas 3 e 7 do algoritmo, o conjunto $papers[C, y]$ é utilizado para identificar quais *snippets* $s \in R_Q$ de fato referenciam artigos publicados na conferência C e ano y. Quando um *snippet* referencia um artigo no conjunto $papers[C, y]$, o valor de $s.label$ é definido como positivo (+) (Linha 5). Por outro lado, quando o *snippet* não referencia um artigo em $papers[C, y]$, o valor de $s.label$ é definido como negativo (-) (Linha 7). Essa verificação é realizada pela função $belongsTo()$ (Linha 4), que utiliza a comparação simples de *string* considerando o título normalizado de cada artigo publicado. Observe que, quando o algoritmo identifica um determinado artigo de $papers[C, y]$ em R_Q, o número de citações do artigo pode ser coletado e salvo em s.

Na Linha 9, cada *snippet* s é convertido em um *bag-of-words (BoW)* e então essa nova representação é salva em $s.bow$. O modelo *bag-of-words*, comumente utilizado em técnicas de aprendizagem de máquina, representa a frequência de cada palavra usada como uma característica a ser usada nas fases de treino e classificação.

Muitas vezes, os *snippets* retornados pelos motores de busca acadêmicos irão conter a *string* do ano y passado como um parâmetro para o Algoritmo 1. No entanto, foram realizados alguns testes iniciais e foi notado que manter a *string* y como uma característica do processo de classificação pode levar a resultados imprecisos. Isso ocorre porque, embora o Algoritmo 1 utilize a lista de artigos de um ano específico y como dados de treino, sua saída é um classificador que deve ser capaz de identificar os artigos de qualquer ano da conferência alvo. Assim, manter o ano y como uma característica pode afetar negativamente o processo de classificação de outras edições da conferência, e por isso na Linha 10 essa *string* é eliminada de $s.bow$.

Embora a remoção de *stopwords* e o uso de *stemming* sejam normalmente aplicados

em problemas de classificação de texto [Pant et al., 2004], alguns experimentos iniciais mostraram que a utilização dessas técnicas pode reduzir a precisão dos classificadores Θ_C. Por exemplo, observamos que alguns *stopwords* desempenham um papel importante na forma que os *snippets* são representados. Por isso, o Algoritmo 1 não aplica essas técnicas na construção dos classificadores Θ_C.

Muitas vezes, considerar palavras de forma isolada pode não ser adequado como característica [Wang et al., 2007]. Isso foi observado em versões iniciais do SHINER quando utilizavam apenas *bag-of-words (BoW)* como representação do vetor de característica dos *snippets*. Assim, afim de descobrir padrões de frases que ocorrem frequentemente nos *snippets* de uma determinada conferência, aplicamos a técnica probabilística de categorização de texto *n-grams* [Damashek, 1995], e passamos a utilizar essas frases ao invés de somente palavras como características importantes para o processo de classificação. Foram testados diferentes tamanhos de *n-grams* e foram obtidos melhores resultados de F1 nas listas de artigos geradas pelo SHINER utilizando a técnica *3-grams*. Observamos também que com a aplicação da técnica *3-grams*, o endereço Web (URL) dos documentos recuperados pelas máquinas de busca acadêmicas passou a ser identificado como característica importante dos *snippets*. Esse passo é feito na Linha 11 do Algoritmo 1 pela função $nGram()$.

Uma questão importante em problemas de classificação de texto é a alta dimensionalidade do espaço de características. Reduzir a dimensionalidade do espaço utilizando técnicas de seleção de características pode ser útil para melhorar o desempenho de classificação e, em muitos casos, para alcançar melhores resultados de classificação reduzindo *overfitting* [Sebastiani, 2002, Caropreso et al., 2001]. Foram avaliados três métodos para redução de dimensionalidade: *document frequency*, *information gain* e *gain ratio* [Baeza-Yates & Ribeiro-Neto, 2013]. Conforme será descrito no Capítulo 5, os melhores resultados foram alcançados durante a utilização do *information gain*, e essa técnica foi utilizada para selecionar apenas as 100 características mais relevantes para os processos de treino e classificação. Esse passo é realizado na Linha 12 do Algoritmo 1 pela função $fSelection()$.

Observe que, entre as Linhas 9 e 12, são realizadas algumas etapas de pré-processamento para transformar cada *snippet* $s \in R_Q$ em uma representação mais adequada para o processo de treinamento. Uma vez concluída essa fase, são utilizadas téc-

nicas de aprendizagem de máquina para gerar o classificador Θ_C do conjunto R_Q. Foi realizada uma análise comparativa para selecionar o método de aprendizagem de máquina que melhor se adapta ao problema que tratamos aqui. Foram considerados os seguintes métodos: *Naive Bayes*, *Naive Bayes Multinomial*, *Decision Tree* and *Support Vector Machine (SVM)*. Como resultado dessa análise, verificamos que o método SVM foi o mais eficaz entre todas as alternativas, e por isso optamos por usar esse método em nossos experimentos. Utilizamos a implementação SVM do pacote Weka [Witten & Frank, 2005], disponível gratuitamente.

Muitas vezes, a maioria dos *snippets* $s \in R_Q$ não referencia artigos publicados na conferência C e ano y ($s.label =$ -). Isso poderia ser uma problema, já que com exemplos de treinamento desequilibrados, o SVM pode dar alta precisão mas baixa revocação na classe minoritária [Sun et al., 2009]. No entanto, nossos experimentos iniciais demonstraram que os classificadores Θ_C não sofrem com esse problema, pois a relação de desequilíbrio tende a se repetir em todas as edições da mesma conferência.

4.3 Aplicação dos Classificadores

Depois que um classificador Θ_C é construído, ele pode ser utilizado para identificar quais respostas de um motor de busca acadêmico referenciam artigos de uma determinada conferência C. O Algoritmo 2 apresenta o algoritmo desenvolvido para aplicar um classificador Θ_C. Esse método recebe como parâmetros: a conferência alvo C, um ano y e o classificador Θ_C da conferência, e no fim esse algoritmo retorna um conjunto $papers[C,y]$ contendo a lista de artigos publicados em C no ano y. A seguir esse algoritmo é descrito em detalhes.

O algoritmo começa através da inicialização de um conjunto vazio $papers[C,y]$. O objetivo desse conjunto é armazenar os *snippets* que referenciam artigos publicados em C no ano y recuperados até um determinado momento pelo algoritmo. Depois disso, o algoritmo segue um conjunto de passos semelhantes aos do Algoritmo 1. Por exemplo, nas Linhas 2 e 3 também são utilizadas as funções $buildQuery()$ e $fetchResults()$ para recuperar um conjunto de *snippets* da máquina de busca acadêmica, e a partir da Linha 5 até 8 são realizadas as mesmas etapas de pré-processamento para transformar cada *snippet* s em uma representação mais útil para o processo de classificação.

Algoritmo 2 Algoritmo para aplicar um classificador Θ_C.

Input: Θ_C, C, y

Output: a array $papers[C,y]$ containing of list of papers published at C in year y

1: $papers[C,y] \leftarrow \emptyset$
2: $Q \leftarrow buildQuery(C,y)$
3: $R_Q \leftarrow fetchResults(Q)$
4: **foreach** $(s \in R_Q)$ **do**
5: $\quad s.bow \leftarrow getBoW(s)$
6: $\quad s.bow \leftarrow s.bow - \{y\}$
7: $\quad s.bow \leftarrow nGram(s.bow)$
8: $\quad s.bow \leftarrow fSelection(s.bow)$
9: $\quad$ **if** $(\Theta_C(s,y) = +)$ **then**
10: $\quad\quad s.citations \leftarrow getCitations(s)$
11: $\quad\quad push(papers[C,y], s)$
12: $\quad$ **end if**
13: **end foreach**
14: $deduplic(papers[C,y])$
15: **return** $papers[C,y]$

Na Linha 9, o classificador Θ_C é finalmente utilizado para verificar se um determinado *snippet* s pertence a conferência alvo C e ano y. Em um caso positivo, o número de citações do artigo é armazenado em $s.citations$ (Linha 7), e o artigo é então adicionado ao conjunto $papers[C,y]$, ou seja, esse artigo é adicionado a lista de artigos mais correta possível da conferência C e ano y.

Às vezes um mesmo artigo pode ser referenciado por mais de um *snippet* de R_Q e consequentemente esse artigo é repetido em $papers[C,y]$. Por isso, na Linha 14 é realizado um processo de deduplicação nesse conjunto para eliminar artigos duplicados.

4.4 Estratégias Para Construção da Consulta

O objetivo da função $buildQuery()$ é criar a consulta que será utilizada para recuperar os artigos de uma determinada conferência. Desse modo, essa função desempenha um papel fundamental no Algoritmos 1 e 2, uma vez que a revocação dos resultados retornados pela máquina de busca acadêmica representa um limite superior sobre a revocação que o método SHINER pode alcançar.

A função $buildQuery()$ inicia através da escolha da melhor estratégia para a constru-

ção da consulta de uma determinada conferência. Essa escolha é feita apenas na fase de treinamento (quando a função é chamada pelo Algoritmo 1), e a mesma estratégia será repetida durante a fase de classificação. Como os *snippets* são altamente dependentes da consulta, o uso da mesma estratégia para a consulta à maquina de busca assegura a conservação das propriedades dos *snippets* na fase de treinamento e na fase de classificação, o que facilita a filtragem desses *snippets*.

A função $buildQuery()$ recebe como parâmetros a conferência alvo C e o ano y. Ambos são utilizados para construir a consulta a ser submetida à maquina de busca. Foram testadas várias estratégias para a construção das consultas submetidas ao *Google Scholar*, por meio de sua interface de busca padrão (GS-D) bem como por meio de sua interface de busca avançada (GS-A). Abaixo são listadas as quatro melhores estratégias encontradas (classificadas por ordem decrescente da sua capacidade de recuperar conjuntos de respostas com alta revocação):

1. $Name_A$ - Consulta submetida preenchendo os campos de pesquisa "Exibir artigos publicados em" e "Exibir artigos com data entre" da GS-A com o nome completo da conferência C e o ano y, respectivamente.

2. $Acronym_D$ - Consulta submetida preenchendo o campo de pesquisa da GS-D com a sigla de C e o o campo "Exibir artigos com data entre" da GS-A com o ano y.

3. $Name_D$ - Consulta submetida preenchendo o campo de pesquisa da GS-D com o nome completo de C e o campo "Exibir artigos com data entre" da GS-A com o ano y.

4. $Acronym_A$ - Consulta submetida preenchendo os campos de pesquisa "Exibir artigos publicados em" e "Exibir artigos com data entre" da GS-A com a sigla de C e o ano y, respectivamente.

A melhor estratégia varia de conferência para conferência, mas as estratégias $Name_A$ e $Acronym_D$ retornaram os melhores resultados, em termos de revocação, na maioria das vezes. Para selecionar a estratégia que melhor se adapta a uma determinada conferência, adotamos a seguinte estratégia:

> *Seguindo a ordem da lista, selecionamos uma estratégia e testamos a sua revocação entre os* α *primeiros resultados recuperados pelo motor de busca . Se a revocação*

encontrada é menor que um determinado limiar β*, selecionamos a próxima estratégia da lista. O valor de* α *é calculado com base no número total de respostas retornados pelo motor de busca para a consulta, e em nossos experimentos adotamos 10% para esse limite. Para o limiar* β*, adotamos 20% em nossos experimentos. Assim, em nossos experimentos, se uma determinada estratégia não alcança 20% de revocação entre os 10% primeiros resultados do Google Scholar, selecionamos a próxima estratégia.*

Capítulo 5

SHINER - Resultados Experimentais

Este capítulo apresenta a descrição de um conjunto de experimentos realizados para avaliar o método SHINER. A seguir, primeiramente é descrita a configuração experimental utilizada, e em seguida são apresentados os resultados obtidos e lições aprendidas.

5.1 Configuração

Os experimentos foram realizados utilizando um conjunto de 30 conferências selecionadas aleatoriamente a partir do índice *Qualis* mantido pela CAPES para a área de Ciência da Computação. Esse índice classifica conferências em sete níveis, e por definição, o número de conferências nesses níveis segue uma distribuição normal. Nos experimentos, a seleção aleatória foi segmentada em 3 grupos que representam respectivamente conferências de alto, médio e baixo impacto. Essa divisão foi organizada da seguinte forma: Grupo A (conferências de níveis *Qualis* A1, A2 e B1); Grupo B (conferências de níveis *Qualis* B2 e B3); e Grupo C (conferências de níveis *Qualis* B4 e B5). Além disso, nessa configuração também foi mantida uma distribuição normal em todos os três grupos.

Essa divisão em grupos foi realizada com o intuito de verificar se existia uma grande diferença nos valores de precisão e revocação das listas de artigos geradas em cada grupo de conferências, visto que as conferências de alto impacto, que possuem os artigos mais citados entre os 3 grupos, poderiam ter um maior volume de informações indexadas nas máquinas de busca acadêmicas, e consequentemente poderiam ter suas listas de artigos geradas com valores de precisão e revocação superiores.

A lista das conferências selecionadas em cada grupo é apresentada na Tabela 5.1,

assim como as informações sobre os dados experimentais relacionados, considerando os anos 2009, 2010 e 2011. Nessa tabela, cada linha corresponde a uma conferência X_i, identificada por seu grupo $X \in \{A,B,C\}$ e um índice i dentro do grupo. Nos experimentos foram consideradas apenas conferências que tiveram edições nesses três anos.

A coluna "Artigos reais" da Tabela 5.1 corresponde ao número correto de artigos que aparecem em cada conferência em um determinado ano. As listas corretas de artigos em cada ano foram obtidas da biblioteca digital *DBLP* ou dos *websites* oficiais das edições das conferências. A coluna "*Snippets* recuperados" representa o número de *snippets* retornados pela máquina de busca acadêmica como resposta a uma consulta submetida. A máquina de busca acadêmica utilizada nos experimentos foi o *Google Scholar (GS)*, que foi escolhido porque é reconhecido por ter a maior cobertura de artigos científicos disponível na *Web* [Silva et al., 2009, Bar-Ilan, 2008, Meho & Yang, 2007]. As estratégias descritas na Seção 4.4 foram utilizadas para a construção de cada consulta submetida ao GS. Para a maioria das conferências foi utilizada a estratégia $Name_A$. Para aquelas marcadas com um o símbolo † na Tabela 5.1, a estratégia $Acronym_D$ ofereceu uma revocação melhor e por isso foi utilizada nessas conferências. Observe que o número máximo de *snippets* recuperados é 1000, que é o tamanho limite do conjunto de respostas retornado pelo GS.

Finalmente, a coluna "Artigos únicos" apresenta o número de artigos únicos cujos *snippets* foram recuperados pela máquina de busca acadêmica. Por exemplo, para a conferência A_1 em 2009, apenas 339 dos 381 artigos reais da lista oficial da conferência nesse ano foram encontrados entre os 1000 *snippets* recuperados pela consulta submetida, isso considerando apenas um *snippet* dos vários *snippets* que referenciam um mesmo artigo, o que acontece com certa frequência no conjunto de resposta.

Conforme descrito no Capítulo 4, os experimentos foram realizados da seguinte forma: primeiramente foi treinado um classificador para cada conferência X_i utilizando os *snippets* buscados de 2009 para cada conferência. Assim, um classificador distinto Θ_{X_i} foi gerado para cada conferência X_i. Em seguida, esse classificador Θ_{X_i} foi aplicado fornecendo como entrada os *snippets* recuperados de 2010 e 2011 para X_i. Finalmente, a saída do classificador foi avaliada comparando-a com a lista de artigos reais obtida para X_i no respectivo ano.

	Artigos reais			***Snippets* recuperados**			**Artigos únicos**		
C	**2009**	**2010**	**2011**	**2009**	**2010**	**2011**	**2009**	**2010**	**2011**
A_1†	381	463	442	1000	1000	1000	339	362	339
A_2†	377	379	422	1000	1000	1000	329	333	261
A_3	124	139	143	119	146	138	116	139	136
A_4	79	82	84	79	73	88	53	56	62
A_5	288	328	275	299	341	272	273	313	266
A_6	317	350	326	347	366	329	294	333	306
A_7†	124	114	135	887	948	1000	115	104	125
A_8	112	66	102	110	70	105	102	64	98
A_9	83	95	90	82	87	126	78	79	83
A_{10}	207	218	237	211	214	235	202	209	225
B_1	32	46	91	32	51	98	29	45	88
B_2	589	557	624	521	647	520	483	447	503
B_3†	74	74	118	412	588	933	70	61	95
B_4	83	80	75	92	79	86	80	77	69
B_5	337	288	285	350	293	281	313	256	263
B_6	28	22	20	30	22	24	23	21	16
B_7	58	87	88	62	88	86	58	87	86
B_8	61	58	80	64	68	78	61	57	76
B_9†	34	42	36	181	174	254	33	35	30
B_{10}	74	60	41	81	64	42	67	58	39
B_{11}	71	101	76	76	102	76	69	100	75
B_{12}	76	54	71	78	54	69	69	50	60
B_{13}	99	90	103	107	91	105	95	87	102
C_1	342	207	163	320	197	159	299	188	148
C_2	492	695	534	519	773	577	467	688	505
C_3	141	208	145	277	210	145	140	200	141
C_4†	128	140	122	210	285	278	126	136	120
C_5	374	129	100	422	130	98	368	128	98
C_6	38	46	49	39	43	45	36	37	40
C_7	78	137	80	78	139	79	78	135	78

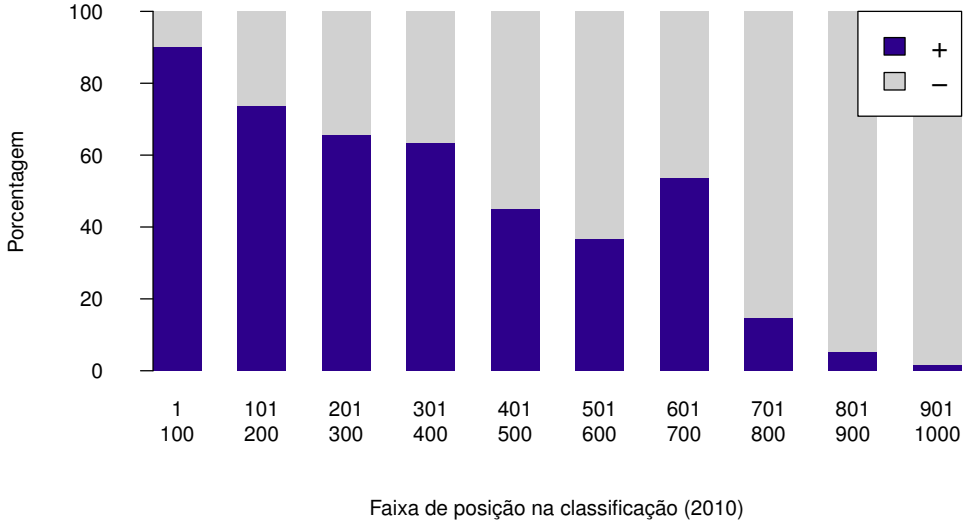
Tabela 5.1: Resumo do conjunto de dados utilizado nos experimentos.

5.1.1 Classificação do Google Scholar

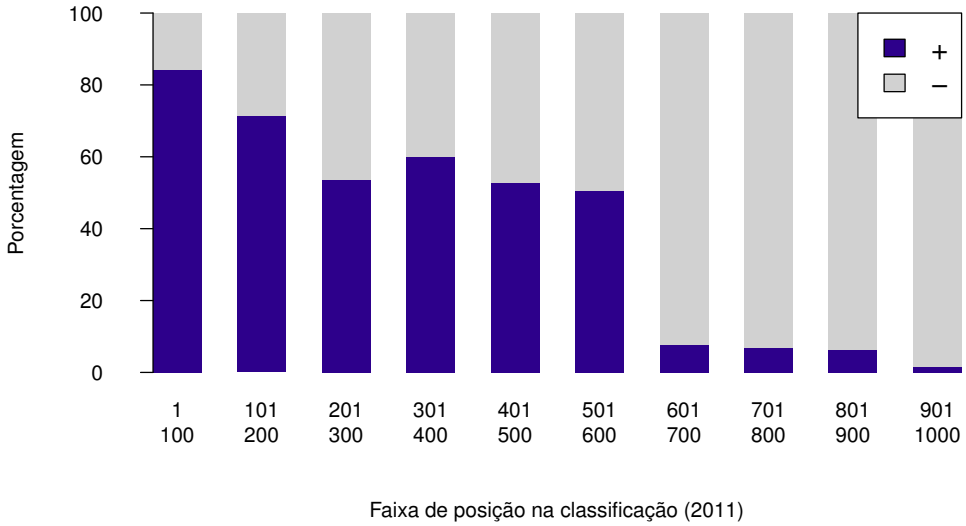
Para confirmar os problemas enfrentados pelos usuários ao tentar utilizar o GS para recuperar a lista de artigos de uma edição de uma determinada conferência, como acontece em ferramentas como o *Publish or Perish* [Harzing, 2007], a Figura 5.1 apresenta dois gráficos que descrevem o comportamento da classificação do GS para as 30 conferências testadas. Para gerar esses gráficos, para cada conferência X_i, foi submetida a consulta correspondente q_i utilizando as estratégias da Seção 4.4. Em seguida, para cada intervalo $[p_{k+1}, p_{k+100}](k{=}0,100,\ldots,900)$ da classificação retornada pelo GS, foi verificada a fração de *snippets* positivos (+), ou seja, *snippets* que realmente correspondem a artigos da conferência, e *snippets* negativos (-), ou seja, *snippets* que não representam artigos da conferência alvo. Cada barra nos gráficos da Figura 5.1 corresponde à média de Pos e Neg considerando todos as 30 conferências em 2010 e 2011.

Como pode ser observado, em ambos os anos, a maioria dos 100 primeiros itens de resposta retornados pelo GS é realmente positivo, embora uma porção significativa deles seja negativo. Assim, cabe ao usuário detectar esses *snippets* negativos entre os positivos. Além disso, também há muitos *snippets* positivos espalhados em posições muito mais baixas na classificação, misturados com os *snippets* negativos. Desse modo, se a tarefa é obter todos os artigos que aparecem em uma conferência/ano, mais uma vez, cabe ao usuário navegar até o último resultado enquanto tenta detectar os poucos *snippets* positivos entre os negativos espalhados pela classificação do GS. Por sua vez, o método SHINER produz uma lista de artigos muito mais próxima da correta, o que é essencial para permitir uma avaliação mais precisa das conferências, uma vez que permite utilizar apenas artigos que de fato pertencem à conferência alvo.

Os gráficos da Figura 5.1 indicam o quão desafiador é a tarefa de obter a lista correta dos artigos de uma conferência, até mesmo para uma máquina de busca acadêmica como o *Google Scholar (GS)*, que tem uma grande cobertura de documentos científicos.



(a) Classificação gerada pelo *Google Scholar* para todas as conferências para o ano de 2010.



(b) Classificação gerada pelo *Google Scholar* para todas as conferências para o ano de 2011.

Figura 5.1: Classificação gerada pelo *Google Scholar* de todas as conferências testadas para os anos 2010 (a) e 2011 (b).

5.2 Resultados

5.2.1 Precisão

A Figura 5.2 apresenta a precisão alcançada pelos classificadores SHINER para os anos de 2010 e 2011 considerando todas as conferências contidas em cada grupo. A precisão para cada conferência X_i é calculada pelo número total de artigos únicos retornados por

Θ_{X_i} que realmente pertencem a conferência X_i, dividido pelo número total de de artigos que o classificador selecionou como pertencentes a conferência X_i.

Na grande maioria dos casos, os valores de precisão ficaram acima ou em torno de 0.9, independentemente do grupo. Além disso, todas as médias, por grupo, por ano e geral, alcançaram valores em torno de 0.9. Isso sugere, que no geral, não houve a diferença esperada na precisão obtida nas listas de artigos geradas em cada grupo de acordo com o impacto de suas conferências, bem como sugere que os classificadores Θ_{X_i} são bastante eficazes no processo de filtragem dos metadados fornecidos pelo *Google Scholar*.

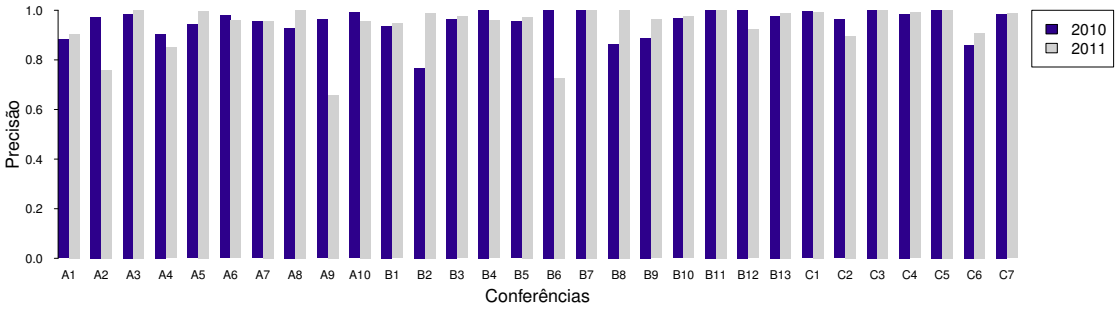


Figura 5.2: Precisão alcançada pelo classificador para as conferências.

5.2.2 Revocação

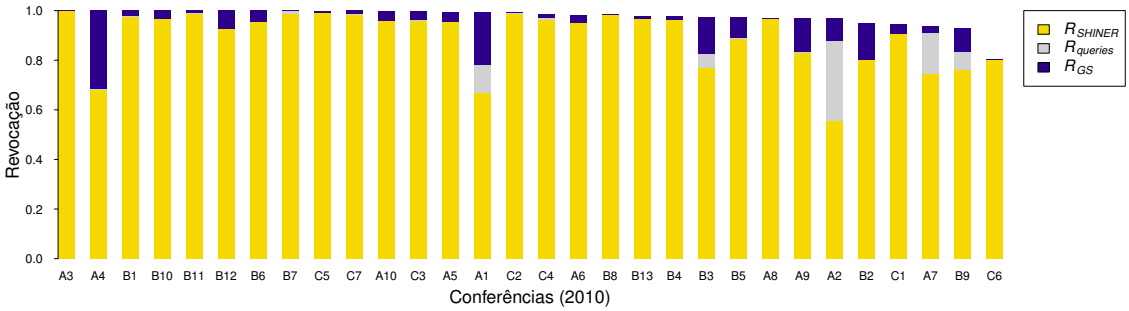
A Figura 5.3 apresenta os resultados em relação aos valores de revocação obtidos em 2010 (5.3a) e 2011 (5.3b), para as conferências testadas. Cada barra mostra três valores de revocação distintos que são descritos a seguir. Os valores $R_{queries}$ correspondem a revocação considerando todos os *snippets* retornados pelo *Google Scholar* para a conferência/ano correspondente quando uma consulta é submetida. Os valores R_{SHINER} correspondem a revocação considerando apenas os *snippets* dados como saída pelo classificador, tendo como entrada todos os *snippets* retornados pela consulta. E finalmente, os valores R_{GS} correspondem à cobertura efetiva do *Google Scholar* para cada conferência/ano. Esse último foi calculado tomando o título de cada artigo da lista de artigos que realmente aparecem na conferência/ano, submetendo esse título como uma consulta, e pesquisando esse artigo particular na resposta retornada. Chamamos esse valor de *revocação absoluta*. Observe que o conjunto de referência é sempre o conjunto de artigos reais que aparecem na edição da conferência correspondente. Assim, por definição, $R_{SHINER} \leq R_{queries} \leq R_{GS}$. Na Figura 5.3, as barras são ordenadas em ordem decrescente dos valores de R_{GS}.

Para exemplificar, considere a conferência A_1 em 2010. Submetendo os títulos de cada artigo de A_1 em 2010 como consultas ao GS, identificamos que o *Scholar* possui 460 dos 463 artigos reais dessa conferência (assim, R_{GS}=0.99). Ao utilizar a melhor consulta definida para essa conferência, foram coletados 362 *snippets* positivos distintos, resultando $R_{queries}$=0.81. Quando esses 362 *snippets* somados aos *snippets* retornados pela consulta são dados ao classificador, nesse casso o classificador Θ_{A_1}, 309 deles são identificados como positivos, levando a R_{SHINER}=0.67.

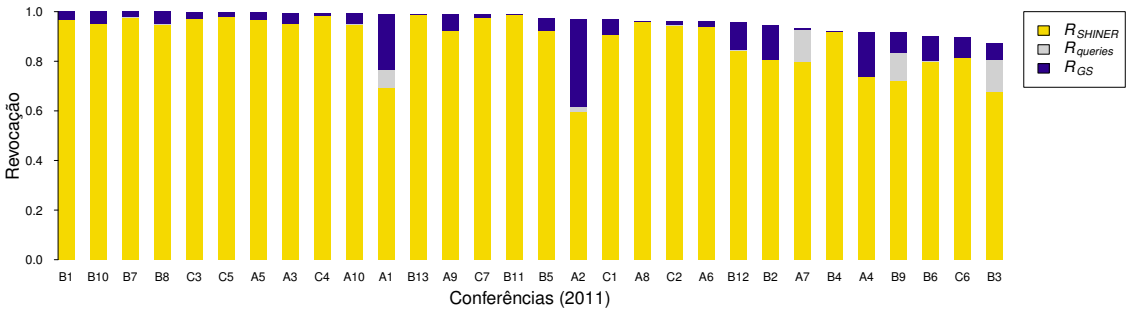
A Figura 5.3 mostra que, na maioria dos casos, os classificadores foram capazes de detectar a maioria dos *snippets* correspondentes a artigos reais da conferência/ano recuperados pela consulta. Na verdade, a perda de revocação do R_{SHINER} em comparação com $R_{queries}$ foi menos de 0.05 em média. Outra inferência interessante que pode ser tirada a partir da Figura 5.3 é que, as nossas estratégias para a submissão das consultas para cada conferência/ano são bastante eficazes, uma vez que em ambos os anos, em média, a perda de revocação de $R_{queries}$ com relação a R_{GS} foi de apenas 0.06.

Em alguns casos, os valores de $R_{queries}$ estão abaixo da média. Por exemplo, para a conferência A_4 em 2010, o R_{GS} é 1, porque todos os 82 artigos reais podem ser encontrados ao realizar consultas pelo título de cada artigo. No entanto, apenas 73 artigos foram encontrados quando a consulta utiliza o nome da conferência. Casos como esse, ocorrem quando os metadados (isto é, os acrônimos, título da conferência, etc) nos *snippets* de alguns dos artigos está escrito de forma diferente do original ou está simplesmente incorreto. Assim, esses *snippets* não podem ser recuperados durante a utilização das conferências como uma consulta.

Em outros casos vemos R_{SHINER} muito menor que $R_{queries}$, como na conferência A_2. Nesse e em outros casos semelhantes, verificamos que alguns exemplos negativos são muito similares a exemplos positivos, por exemplo, os *snippets* referentes a artigos que foram publicados em conferências com nomes muito semelhantes, e assim, na fase de treinamento as características relativas a esses exemplos positivos acabaram sendo removidas. Observe na Figura 5.3, que essa anomalia não é generalizada entre o conjunto de conferências testadas.



(a) Edição de 2010 das conferências.



(b) Edição de 2011 das conferências.

Figura 5.3: Revocação alcançada para as conferências em 2010 (a) e 2011 (b).

5.3 Papel da Seleção de Características

Como discutido na Seção 4.2, um ponto interessante a ser observado no processo de geração dos classificadores SHINER é o papel das técnicas de seleção de características que foram adotadas. A Figura 5.4 apresenta uma comparação dos resultados de F1 obtidos pelos classificadores gerados com a seleção de características (FS) e sem seleção de características (Sem FS). Cada barra no gráfico representa o valor médio de F1 para cada conferência em ambos os anos 2010 e 2011. Os classificadores que utilizam seleção de características alcançaram melhores resultados em 22 das 30 conferências testadas. Na maioria dos casos, a diferença em termos de F1 é bastante pequena, menos de 0.05 em média. No entanto, em alguns casos, por exemplo, para a conferência $C3$, o uso de seleção de características foi fundamental para os resultados alcançados.

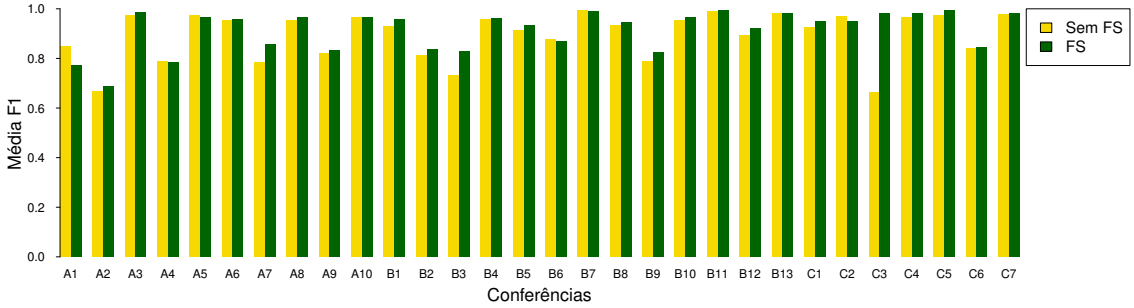


Figura 5.4: Média F1 para 2010 e 2011 alcançada pelos classificadores gerados com e sem seleção de características.

Capítulo 6

Coleta Colaborativa de Metadados de Artigos

Este capítulo apresenta uma discussão sobre a coleta colaborativa de metadados de artigos realizada pela ferramenta Live SHINE. Adicionalmente, são apresentadas algumas limitações que o *Google Scholar* impõe sobre o *crawling* de seu conteúdo, bem como nossa proposta para contornar tais limitações.

6.1 Coleta de Dados de Citações

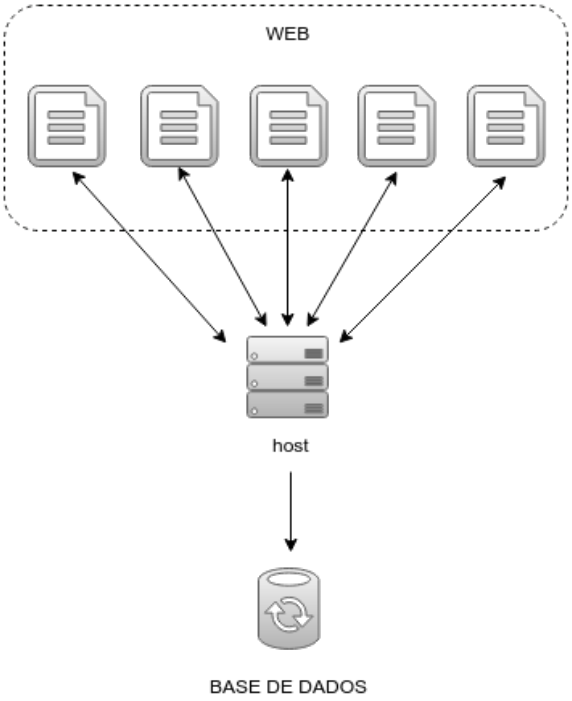
Conforme discutido no Capítulo 3, para estimar índices de impacto de conferências com alta precisão é necessário obter dados de citações atualizados dos artigos publicados em tais veículos. No entanto, manter esses dados de citações sempre atualizados requer um monitoramento constante de um número desconhecido de publicações, porque, em teoria, qualquer artigo publicado em qualquer veículo poder trazer uma citação a um artigo de uma conferência de interesse. Isso representa um enorme desafio a ser superado. Por conta disso, as ferramentas mais utilizadas para se obter índices de impacto de conferências, como por exemplo o Publish or Perish (PoP) e o SHINE, utilizam os dados fornecidos pela máquina de busca Google Scholar (GS), que nos últimos anos aumentou consideravelmente sua cobertura e se tornou uma fonte promissora para pesquisas de avaliação de conferências e pesquisas bibliométricas [Harzing, 2014].

Para se obter os dados de citações a partir do GS, é necessário realizar consultas a essa máquina de busca, coletar as páginas de respostas e extrair os metadados dos artigos. Tal

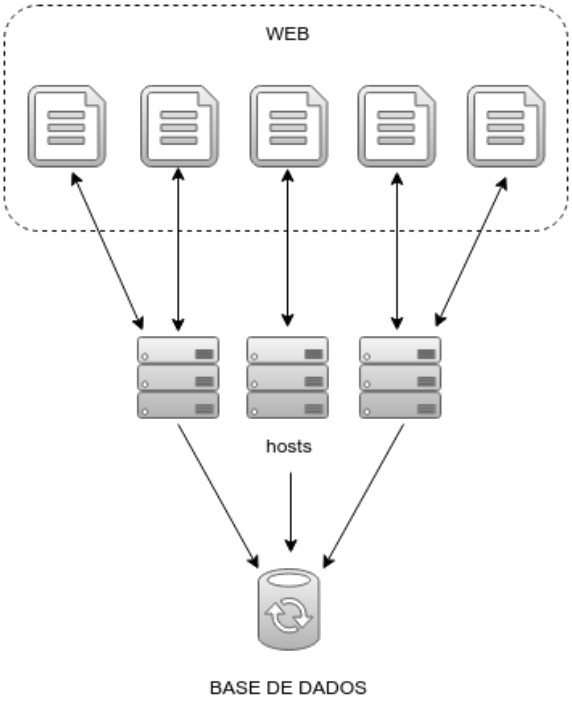
tarefa é realizada a partir da utilização de um *Web Crawler*, também chamado de robô, coletor ou simplesmente crawler. De maneira geral, um crawler [Olston & Najork, 2010] é um sistema criado para coletar grandes conjuntos de páginas Web. Esse sistema é um dos principais componentes das máquinas de buscas, que visam coletar páginas na Web para indexá-las e permitir que os usuários realizem consultas sobre os índices gerados. Além dos sistemas de busca, o uso de crawlers tem se estendido para vários outros fins. Por exemplo, na área de mineração de dados, os crawlers têm sido utilizados para coletar, analisar e monitorar páginas e documentos disponíveis na Web [Olston & Najork, 2010].

Muitas vezes, os crawlers são implementados através de uma arquitetura centralizada. Como ilustra a Figura 6.1a, nesse tipo de arquitetura, a coleta das páginas Web é realizada a partir de um único *host* central, que pode realizar requisições a várias páginas Web diferentes ao mesmo tempo, para em seguida armazená-las em uma base de dados. Essa arquitetura é a mais utilizada pois, além de ser mais simples, não requer um alto custo de implementação e operação. No entanto, como toda a coleta é feita a partir de um único *host*, essa arquitetura pode sofrer com algumas limitações. Para exemplificar, considere o caso da ferramenta SHINE descrita na Capítulo 2. Atualmente essa ferramenta possui uma base com cerca de 800 mil artigos, distribuídos em mais de 1800 veículos cadastrados. Para obter os dados de citações desses artigos, os autores dessa ferramenta desenvolveram um crawler centralizado capaz de fazer consultas ao GS usando os títulos de cada artigo da base, e então buscar na página de respostas os metadados do artigo desejado. Desta forma, buscando os metadados de um único artigo por vez, o crawler do SHINE precisa fazer 800 mil consultas ao GS para atualizar toda a sua base. Se consideramos que o crawler realiza uma consulta a cada 30 segundos para não sobrecarregar o servidor do GS e nem a banda de internet local, seriam necessários cerca de 278 dias ininterruptos de coleta para concluir a atualização. Assim, a medida que o número de artigos armazenados na base de dados aumenta com as novas edições das conferências, esse processo de coleta a partir de um único *host* central se torna cada vez mais inviável. Por conta disso, desde 2012 a base de dados do SHINE não é atualizada.

Uma alternativa à abordagem centralizada é realizar a coleta dos dados de citações do GS a partir de um crawler distribuído. Como ilustra a Figura 6.1b, nessa arquitetura, a coleta das páginas Web é feita por um conjunto de *hosts* espalhados ao redor do mundo, que podem realizar múltiplas requisições ao mesmo tempo sem sobrecarregar a banda de



(a) Arquitetura de um crawler centralizado.



(b) Arquitetura de um crawler distribuído.

Figura 6.1: Diferença entre a arquitetura de um crawler centralizado (a) e distribuído (b).

internet. Quanto mais *hosts* disponíveis, mais rápido a coleta é finalizada, pois cada *host* pode coletar uma ou várias páginas Web ao mesmo tempo. No entanto, essa arquitetura pode se tornar inviável, uma vez que seria necessário um alto custo para adquirir e manter vários *hosts* trabalhando em conjunto ao redor do mundo.

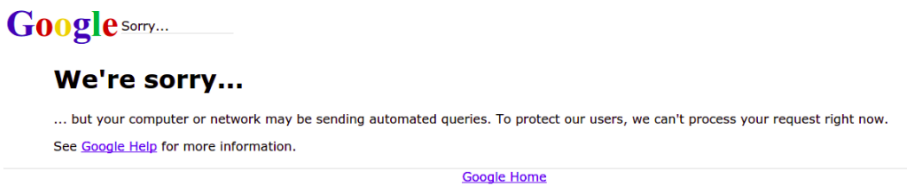
6.2 Limitações do *Google Scholar*

Outra dificuldade associada ao uso do Google Scholar para coleta de citações, é que essa máquina de busca acadêmica limita a quantidade de requisições que um único *host* pode submeter dentro de um intervalo de tempo. Quando o GS identifica que *hosts* está realizando consultas frequentes e automáticas à sua máquina de busca, imediatamente ele deixa de atender as consultas submetidas e passa a retornar páginas de notificações ao usuário. A Figura 6.2 apresenta duas páginas distintas que podem ser retornadas pelo GS, cada uma sendo responsável por notificar o usuário de um tipo de bloqueio específico.

A primeira página, apresentada na Figura 6.2a, é uma página de verificação através da qual o usuário é notificado de que seu computador gerou um padrão de requisições incomum ao Google Scholar. Para resolver esse problema, o GS solicita que o usuário resolva um *captcha* para comprovar que as consultas submetidas estão sendo realizadas por um ser humano, e não por um sistema automatizado. Caso o usuário resolva corretamente esse *captcha*, as próximas requisições serão atendidas normalmente pelo GS.

A segunda página, apresentada na Figura 6.2b, mostra uma notificação de bloqueio bem mais restritivo do que o da primeira página. Quando este bloqueio é aplicado, as consultas submetidas pelo usuário (ou crawler) não serão atendidas durante um determinado período de tempo, que pode variar entre algumas horas e alguns poucos dias. Importante ressaltar esse segundo tipo de bloqueio só é apresentado quando o *captcha* da primeira página foi resolvido corretamente pelo menos uma vez, e ainda assim o GS identificou que o computador bloqueado estava disparando requisições automáticas. Após o período de bloqueio, as consultas voltam a ser atendidas normalmente pelo GS.

(a) Página de notificação com *captcha*.



(b) Página de notificação com bloqueio temporário.

Figura 6.2: Páginas de notificação do *Google Scholar (GS)*.

6.3 Solução baseada em Coleta Colaborativa

A fim de superar tais limitações, ao invés de utilizar um sistema centralizado ou distribuído de crawling, nossa solução oferece um mecanismo em que os próprios pesquisadores (usuários) podem colaborar entre si para coletar os metadados fornecidos pelo GS, incluindo os dados de citações. Como ilustra a Figura 6.3, nosso sistema de coleta utiliza conceitos de uma nova arquitetura de crawling denominada *Crowd Crawling* [Ding et al., 2013], que chamamos aqui de *Coleta Colaborativa*, que aproveita os *hosts* de vários grupos de usuários que trabalham em conjunto para atualizar os dados de citações de uma base central. Assim, essa arquitetura pode garantir um baixo custo em sua implementação e manutenção, bem como pode garantir maior escalabilidade e

cobertura na coleta dos metadados dos artigos do GS, e ao mesmo, satisfazer a política do GS, pois as consultas são disparadas sempre pelo o usuário, ao invés de por um crawler tradicional.

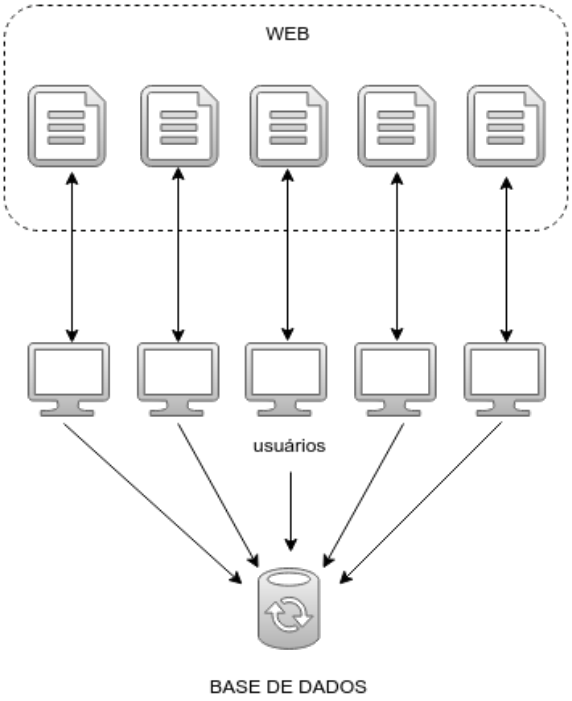


Figura 6.3: Arquitetura de um crowd crawler.

Para viabilizar esta estratégia de coleta colaborativa, a interface de usuário do ambiente Live SHINE é implementada na forma de uma extensão (*plugin*) que é instalado nos navegadores dos usuários (pesquisadores). Ao utilizar esta extensão para obter os índices de impacto de uma dada conferência, o Live SHINE verifica se os dados de citações dos artigos da conferência estão desatualizados em nossa base central. Em caso positivo, a extensão acessa o GS para coletar os dados de citações desses artigos, e envia os dados coletados para o servidor da base central. Assim, toda vez que um usuário utiliza nossa extensão, ele automaticamente colabora com a coleta e atualização dos metadados utilizados no cálculo dos índices de impacto.

Observe que o uso do GS por nossa arquitetura de coleta colaborativa não difere muito do uso regular dos usuários dessa máquina de busca acadêmica. A coleta dos metadados de artigos é feita a partir da seleção de uma conferência feita pelo próprio usuário, sendo portanto mais aderente às políticas de uso do GS do que as demais abordagens de crawling apresentadas na Seção 6.1. Note também que essa abordagem de acesso aos dados do GS é muito similar à abordagem utilizada pela ferramenta Publish or Perish, sendo que a

única diferença entre as duas abordagens é que os dados coletados pelo Live SHINE são enviados para uma base central, ficando disponíveis para futuros acessos à ferramenta.

Outra característica importante da abordagem coleta colaborativa é que os usuários poderão escolher as conferências que desejam atualizar, bastando para isso consultar os índices de impacto da conferência desejada na interface da ferramenta. Essa abordagem difere muito da abordagem centralizada da aplicação SHINE, onde um algoritmo central deveria decidir que conferências seriam ser atualizadas em um determinado intervalo de tempo. Observe também que outro efeito dessa característica do Live SHINE é que as conferências mais importantes e populares tenderão a estar sempre atualizadas na base central do sistema.

Para minimizar os dados trocados entre os navegadores dos usuários e o servidor da base central, a ferramenta Live SHINE faz uma pré-filtragem dos dados coletados do Google Scholar, deixando apenas os dados interesse. Por exemplo, a ferramenta elimina todo o conteúdo HTML, CSS e JavaScript retornado pelo GS, mantendo apenas o conteúdo textual dos snippets presentes nas páginas de resposta. Desta forma, a utilização da banda do servidor da base central é muito menor do que na abordagem centralizada, discutida na Seção 6.1.

Mais detalhes sobre a ferramenta Live SHINE e como a coleta colaborativa é implementada são apresentados no Capítulo 7.

Capítulo 7

A Ferramenta Live SHINE

Neste capítulo são apresentados os detalhes da ferramenta Live SHINE, proposta neste trabalho, cujo objetivo é gerar índices de impacto de alta precisão de conferências de Ciência da Computação a partir de dados fornecidos pelo *Google Scholar (GS)*. A seguir, descrevemos a arquitetura e seus principais módulos, a coleta colaborativa de metadados de artigos, e por fim, a extensão e suas interfaces de consulta e avaliação.

7.1 Arquitetura

A Figura 7.1 apresenta a arquitetura geral da ferramenta. Ao ativar a extensão Live SHINE no navegador Web, o usuário pode facilmente selecionar uma determinada conferência *C* e ano *y* (ou intervalo) sobre o qual deseja obter os índices de impacto. Uma vez que esses dados são selecionados e o usuário submete a consulta, o Live SHINE então inicia duas *threads* que são executadas em paralelo.

A primeira *thread* é responsável por estimar os índices de impacto da conferência e entregá-los imediatamente ao usuário. Quando o Live SHINE recebe a conferência *C* e ano *y* selecionados, repassa esses dados ao módulo "Estimador de Impacto", que utiliza a lista de artigos da conferência *C* disponível no cache de metadados para o cálculo dos índices de impacto. Uma vez que essa lista de artigos é recuperada, ela é ordenada em ordem não ascendente de citações dos artigos, e então é utilizada no cálculo dos índices de impacto. Finalmente, assim que os índices são calculados, o Live SHINE apresenta para o usuário a lista de artigos e seus respectivos metadados, bem como os índices de impacto calculados. Nesse momento essa *thread* é encerrada.

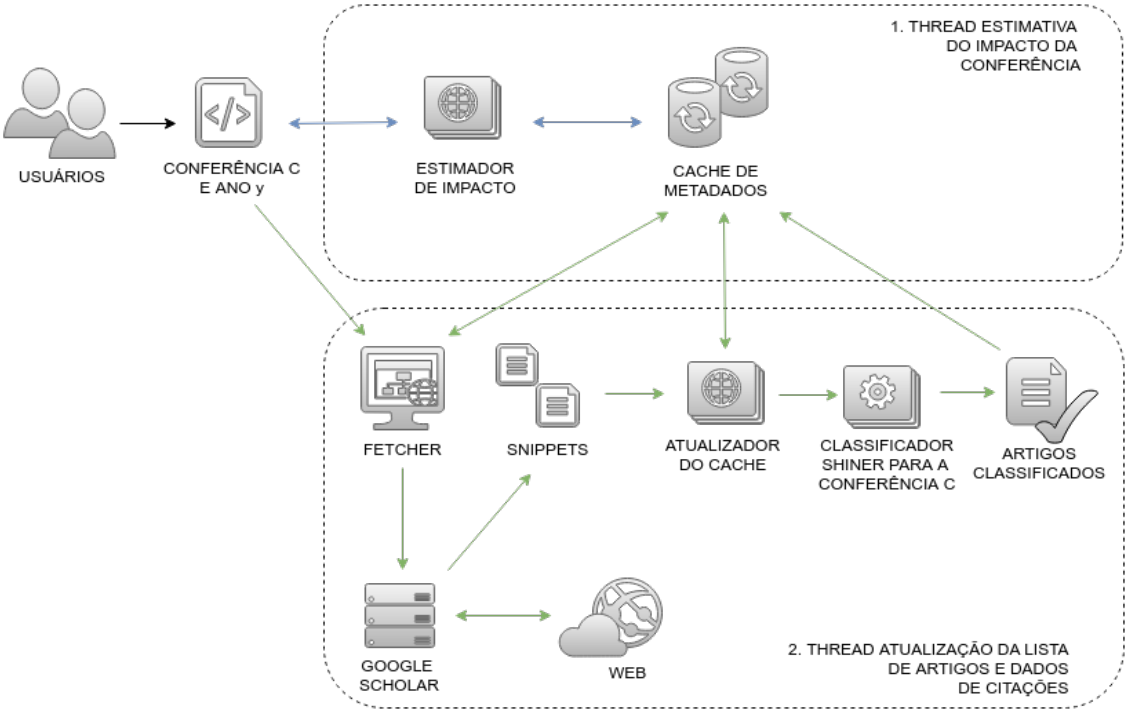


Figura 7.1: Arquitetura Geral do Live SHINE.

A segunda *thread* é responsável por atualizar as listas de artigos armazenadas no cache de metadados. Quando o Live SHINE recebe a conferência C e ano y selecionados, esses dados são repassados ao módulo "Fetcher", que verifica no cache de metadados a necessidade de atualizar a lista de artigos da conferência C e ano y. Essa lista precisará ser atualizada se uma proporção superior a um limiar γ de seus artigos não têm seus dados de citações coletados há mais de M meses. γ e M são parâmetros de configuração da ferramenta. Caso seja necessária a atualização dessa lista de artigos, esse módulo realiza consultas ao GS de acordo com as estratégias apresentadas no Capítulo 4 e coleta os *snippets* retornados no conjunto de resposta para a conferência C. Caso contrário, esse módulo identifica a lista da conferência e ano mais desatualizada no cache do Live SHINE e realiza os mesmos passos anteriores para essa nova lista. Esse processo de atualização é apresentado na Seção 7.1.1.

Depois de coletados, esses *snippets* são passados para o módulo "Atualizador do Cache", que verifica quais *snippets* recuperados pelo "Fetcher" referenciam artigos armazenados no cache de metadados, e atualiza os dados de citações desses artigos. Em alguns casos, *snippets* que não referenciam artigos no cache podem ser encontrados no conjunto de resposta do GS. Dois casos podem ocorrer. Primeiro, o GS recuperou erroneamente um *snippet* de um artigo que não pertence a conferência C e ano y. Segundo, o *snippet*

referencia um artigo da conferência C, mas não havia sido recuperado antes. Assim, os *snippets* que não referenciam artigos do cache de metadados, ou seja, os novos *snippets*, são repassados para um "Classificador SHINER", previamente treinado para a conferência C de acordo com o algoritmo apresentado no Capítulo 4. Esse classificador rotula esses *snippets* como pertencentes ou não a conferência C, e armazena os *snippets* classificados no cache de metadados junto aos seus respectivos artigos. Desta forma, na próxima vez que esses *snippets* aparecerem no conjunto de resposta do GS, já estarão armazenados no cache de metadados e não precisarão ser classificados novamente. Essa *thread* é executada em segundo plano pelo Live SHINE e é muito importante para a ferramenta, pois é através dela que implementamos a estratégia de coleta colaborativa de metadados descrita na Seção 7.2.

Vale ressaltar que devido o fato dessas duas *threads* serem executadas ao mesmo tempo pelo Live SHINE, os novos metadados coletados pela segunda *thread* não influenciam no cálculo dos índices de impacto apresentados pela primeira *thread*.

Nas próximas seções os principais módulos da ferramenta são descritos com mais detalhes.

7.1.1 Módulo Fetcher

O módulo *Fetcher* é responsável por coletar do *Google Scholar (GS)* metadados de artigos atualizados da conferência C selecionada pelo usuário, caso tais metadados estejam desatualizados no cache. Inicialmente, esse módulo realiza uma consulta ao cache de metadados para recuperar a lista de artigos da conferência C e ano y.

Em seguida, o modulo contabiliza a quantidade de artigos desatualizados há mais de M meses dessa lista. Caso a proporção de artigos desatualizados em relação ao tamanho da lista seja maior que um limiar γ, o módulo *Fetcher* cria uma consulta de acordo com as estratégias apresentadas no Capítulo 4 para a conferência C, e, logo depois submete essa consulta ao GS e coleta os *snippets* retornados a fim de atualizar os metadados dos artigos da lista (incluindo as citações). No entanto, caso a proporção de artigos desatualizados seja menor do que o limiar γ, o módulo assume que não é necessário fazer a coleta para a conferência C e ano y, e então realiza uma outra consulta ao cache a fim de recuperar a conferência C_D e ano y_D com a maior proporção de artigos desatualizados no cache. Finalmente, o módulo realiza a consulta ao GS para essa nova conferência C_D e ano y_D a

fim de atualizar os metadados dos artigos da sua respectiva lista.

O valores do limiar γ e do período M adotados em nossos experimentos foi respectivamente de 30% e 2 meses. Escolhemos esse valor para γ porque dentre todas as conferências testadas, nenhuma conferência/ano depois de atualizada permaneceu com uma proporção de artigos desatualizados maior que esse valor. E escolhemos esse valor para o período M porque consideramos que é um curto período de tempo para haver grandes mudanças nas citações dos artigos. No entanto, com a evolução e o crescimento contínuo do cache de metadados, no futuro pode surgir a necessidade de ajustar esses valores, o que poderá ser facilmente configurado ferramenta.

O Live SHINE também permite obter os índices de impacto de uma conferência C para um intervalo de até 10 anos. Durante nossos experimentos, observamos que esse tipo de consulta no GS tende a oferecer uma revocação muito inferior em relação a uma consulta por um ano específico y. Isso acontece principalmente porque o GS impõe um limite de 1000 respostas para qualquer consulta submetida, impossibilitando que seu algoritmo de classificação apresente todos os artigos de todas as edições dentro desse intervalo. Assim, para esse tipo de consulta, o módulo *Fetcher* verifica no cache o ano com a maior proporção de artigos desatualizados dentro desse intervalo, e em seguida, realiza a consulta ao GS a fim de atualizar apenas a lista de artigos desse ano selecionado. Observe que, no Live SHINE, se uma consulta para um conferência C com um intervalo de 10 anos for realizada 10 vezes, então todas as listas de artigos das 10 edições da conferência C presentes no cache serão atualizadas.

7.1.2 Módulo Atualizador do Cache

Esse módulo é responsável por atualizar o cache de metadados com base no conjunto de *snippets* do GS recuperados pelo módulo *Fetcher*. Inicialmente, esse módulo realiza uma consulta ao cache para recuperar a lista de artigos da conferência C e ano y. Caso um *snippet* recuperado pelo *Fetcher* referencie algum artigo dessa lista, seus metadados são extraídos e utilizados para atualizar o registro do artigo correspondente no cache de metadados. Caso contrário, esse *snippet* é adicionado a um conjunto de novos *snippets* que serão posteriormente passados para um Classificador SHINER (vide próxima seção). Vale ressaltar que essa lista de artigos pode conter tanto artigos de *snippets* positivos quanto de negativos, pois ela representa os *snippets* do conjunto de respostas retornado

pelo GS. Portanto, os novos *snippets* são aqueles que nunca tinham sido recuperados antes pelo módulo *Fetcher* e devem ser classificados antes de serem armazenados no cache de metadados (vide próxima seção).

Para verificar se um título de artigo de um *snippet* está contido na lista de artigos da conferência C e ano y armazenada no cache, esse módulo utiliza comparação simples de *string* normalizando os títulos antes de compará-los.

7.1.3 Módulo Classificador SHINER

Um *classificador SHINER* (apresentado em detalhes no Capítulo 4) é responsável por selecionar os *snippets* que de fato referenciam artigos da conferência C entre os novos *snippets*, ou seja, aqueles que não foram encontrados no cache de metadados pelo módulo *Atualizador do Cache*. Para cada conferência C, um *classificador SHINER* (Θ_C) foi anteriormente treinado utilizando os *snippets* coletados para o ano de 2009 da conferência C de acordo com o algoritmo proposto no Capítulo 4. Quando um Classificador (Θ_C) identifica um novo artigo da conferência C, seus metadados são extraídos e utilizados para criar um novo registro de artigo no cache de metadados do Live SHINE. Assim, na próxima vez que esse *snippet* aparecer no conjunto de resposta do GS, ele não precisará ser classificado novamente, pois já estará armazenado no cache de metadados e precisará apenas ser atualizado pelo modo *Atualizador do Cache*.

7.1.4 Estimador de Impacto

O módulo *Estimador de Impacto* é responsável por calcular os índices de impacto de todas as conferências armazenadas no cache de metadados do Live SHINE. Inicialmente, esse módulo realiza uma consulta ao cache de metadados para recuperar a lista de artigos classificados como positivos da conferência C e ano y (ou intervalo). Em seguida, ordena essa lista de artigos em ordem não ascendente de número de citações e aplica o algoritmo para calcular o índice de impacto para esses artigos. Utilizamos o algoritmo da métrica *h-index* (descrita no Capítulo 2) para gerar os índices de impacto. Escolhemos essa métrica pois, além de ser comumente utilizada, também representa o impacto de uma determinada conferência de forma quantitativa e qualitativa ao mesmo tempo. Além disso, o *h-index* é a principal métrica fornecida pelas ferramentas analisadas nesta dissertação que serviram de inspiração para o desenvolvimento do Live SHINE. Note que também é possível avaliar o

impacto de conferências utilizando outras métricas de impacto, tais como as apresentadas no Capítulo 2. A aplicação dessas métricas à ferramenta será deixada como trabalhos futuros.

Caso o usuário realize uma consulta por intervalo de anos para a conferência *C*, a lista obtida do cache de metadados irá conter todos os artigos armazenados e classificados como positivos de todas as edições dentro desse intervalo. Assim, se uma consulta foi realizada para um intervalo de 10 anos, o calculo do *h-index* será realizado com todos os artigos que pertencem a conferência dentro desse intervalo e que estão armazenados no cache de metadados.

7.1.5 Cache de Metadados

O cache de metadados é utilizado com o objetivo de possibilitar ao Live SHINE fornecer os índices de impacto de forma imediata ao usuário. Quando uma consulta é realizada ao GS, são retornados no máximo 20 *snippets* por página de resposta. Assim, se considerarmos que uma dada consulta realizada gerou um conjunto de 1000 *snippets*, seria necessário navegar por 50 páginas para obter todo o conjunto de resposta. Dessa forma, para garantir que o usuário obtenha os índices de impacto forma imediata, o Live SHINE utiliza os metadados que estão disponíveis no cache no momento que o usuário faz a consulta.

Sem o cache de metadados, toda vez que o usuário realizasse uma consulta ao Live SHINE, esperaria até que todas as páginas de respostas da consulta fossem coletadas. Vale lembrar que nossa tarefa é identificar todos os artigos possíveis contidos no índice da máquina busca, pois somente assim o impacto da conferência pode ser estimado com maior precisão. Assim, depois desse processo de coleta o usuário ainda esperaria até que todos os *snippets* fossem extraídos, filtrados, e finalmente, a lista de artigos estivesse completa para que pudesse ser utilizada na estimativa de impacto. Observe que todo esse processo resultaria em uma espera inconveniente para o usuário, uma vez que seu objetivo em utilizar a ferramenta é somente obter os índices de impacto para a conferência *C* e ano *y* selecionados na interface do Live SHINE.

Acreditamos que manter um cache com as informações das listas de artigos e dados de citações para as conferências, agiliza muito o processo de levantamento dos índices de impacto. Assim, utilizando um cache de metadados, podemos atender de forma imediata

a necessidade do usuário em obter os índices de impacto para uma conferência C e ano y, enquanto que esse cache é atualizado em segundo plano por meio da coleta colaborativa realizada no momento da consulta do usuário.

A seguir o processo de coleta e atualização é explicado com mais detalhes.

7.2 Coleta Colaborativa de Metadados

Um das características principais da nossa ferramenta é a coleta colaborativa de metadados de artigos que é realizada pela extensão Live SHINE. Quando um usuário utiliza nossa ferramenta para obter os índices de impacto de uma conferência ou mesmo para avaliar as listas de artigos, imediatante a ferramenta verifica a necessidade de coletar novos metadados para atualizar as listas armazenadas no cache da ferramenta. Basicamente, esse processo é realizado pela extensão instalada no navegador Web, que realiza em segundo plano consultas ao site do *Google Scholar (GS)* e navega por todas as páginas de respostas a fim de obter metadados de artigos de conferências de Ciência da Computação que estão presentes no índice do GS.

Quando o usuário consulta por uma conferência em nossa ferramenta, uma conferência e ano é selecionada automaticamente de acordo com sua necessidade de atualização no cache de metadados. Essa conferência e ano pode ser a mesma selecionada pelo usuário em uma das interfaces da extensão Live SHINE, ou pode ser uma outra conferência e ano que esteja mais desatualizada no cache da ferramenta. Esse processo é realizado com o intuito de manter os dados sobre artigos e conferências atualizados, possibilitando que nossa ferramenta atenda de forma imediata a necessidade dos pesquisadores em obter os índices de impacto de conferências de Ciência da Computação.

Para exemplificar, considere que um usuário tenha selecionado uma conferência C e ano y na extensão Live SHINE. Ao mesmo tempo em que a ferramenta busca no cache de metadados a lista de artigos correspondente a conferência C e ano y para calcular os respectivos índices de impacto, também analisa se essa lista de artigos precisa ter seus metadados atualizados. Caso não exista a necessidade de atualizar essa lista de artigos, outra lista correspondente a uma conferência C_D e ano y_D mais desatualizada no cache de metadados é selecionada. Em seguida, com a conferência e o ano selecionados, o Live SHINE realiza a consulta pré-estabelecida para essa conferência e ano a fim de obter

metadados atualizados. Assim, enquanto a ferramenta apresenta os índices de impacto ao usuário, em segundo plano é realizado um processo automático de navegação, extração e filtragem desses metadados presentes em todas as páginas de respostas retornadas pelo GS. Finalmente, esses dados coletados são utilizados para atualizar a lista correspondente a essa conferência e ano no cache de metadados do Live SHINE.

Denominamos esse processo de "*Coleta Colaborativa de Metadados*" pois tentamos disponibilizar uma forma dos pesquisadores colaborarem entre si através do Live SHINE. A partir desse processo, nossa ferramenta tende a utilizar sempre dados de citações de artigos atualizados nas estimativas do impacto das conferências, e assim possibilitamos também que os pesquisadores possam colaborar com a alta precisão nos índices fornecidos pelos Live SHINE.

Esse processo pode ser visto em prática nas próximas seções, onde a extensão Live SHINE é apresentada em mais detalhes.

7.3 Extensão

A ferramenta Live SHINE foi concebida em forma de extensão para o *Navegador Web Google Chrome* e funciona sobre o site do *Google Scholar (GS)*. Ela possui duas interfaces principais, das quais cada uma é voltada para um tipo específico de usuário. A primeira interface, chamada "Interface de Consulta", é a interface padrão da ferramenta, e é através dela que os pesquisadores podem obter os índices de impacto das conferências armazenadas no cache de metadados do Live SHINE. A segunda interface, chamada "Interface de Avaliação", requer um nível de acesso mais alto e é voltada a pesquisadores interessados em contribuir com o Live SHINE avaliando as listas de artigos armazenadas no cache de metadados.

A seguir essas interfaces são apresentadas com mais detalhes.

7.3.1 Interface de Consulta

Essa é a interface padrão utilizada pela maior parte dos usuários do Live SHINE. Seu objetivo é possibilitar que os pesquisadores possam obter os índices de impacto de conferências de Ciência da Computação a partir de uma consulta simples e amigável. A Figura 7.2 apresenta a Tela Inicial da interface. Em (**1**), temos um bloco azul com uma

breve instrução de como realizar uma consulta no Live SHINE, que é apresentado para qualquer usuário no momento em que a extensão é carregada sobre o site do GS. Já em (**2**), temos um bloco verde notificando que o usuário "*Leonardo Fontes do Nascimento*" teve seu *e-mail* "*Leonardo.Stenofh@gmail.com*" (que estava logado no *Google Scholar*) identificado como avaliador e caso deseje pode acessar o "*Modo Avaliador*" (apresentado na Seção 7.3.2). Esse bloco é apresentado apenas para os usuários cadastrados no Live SHINE como Pesquisador Avaliador. Em (**3**) e (**4**), temos respectivamente uma breve descrição da ferramenta e um aviso sobre a coleta colaborativa realizada no momento da consulta. E finalmente, temos os campos de pesquisa que devem ser preenchidos pelo usuário. Basicamente, em (**5**) o usuário deve preencher um campo de pesquisa com o nome ou sigla da conferência desejada, em (**6**) o ano de início e fim do intervalo sobre o qual deseja obter os índices e então em (**7**) deve clicar em enviar. Observe que caso o usuário deseje obter o impacto para um ano específico da conferência, em (**6**) ele deve selecionar o mesmo ano nos campos de ano de início e fim.

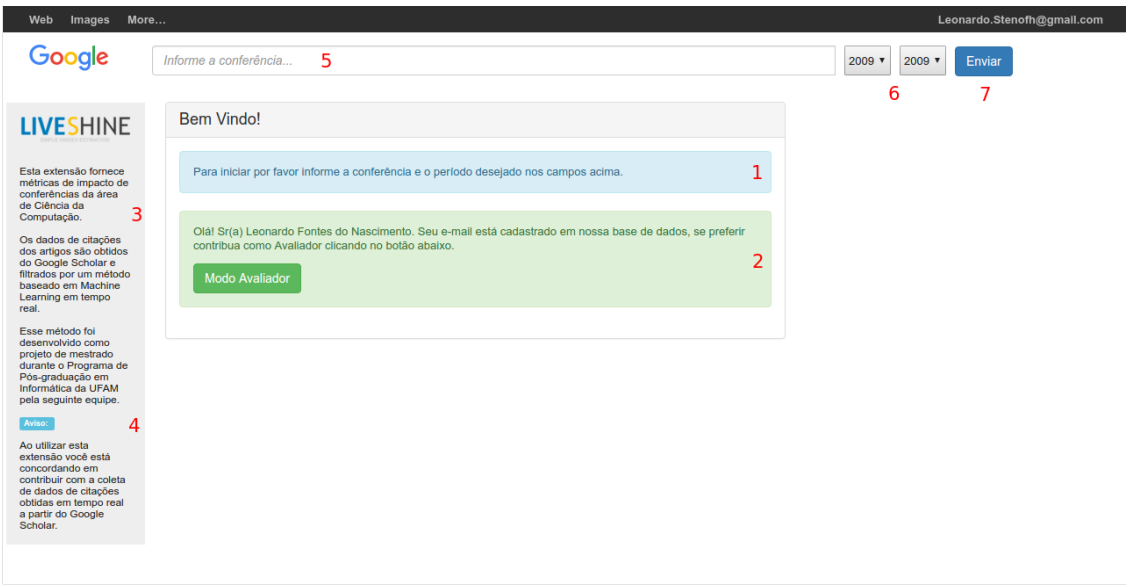


Figura 7.2: Interface de Consulta - Tela Inicial.

Depois que o usuário realiza a consulta, o Live SHINE estima os índices de impacto para a conferência e intervalo selecionados pelo usuário. A Figura 7.3 apresenta a Tela Estimativa do Impacto da Conferência. Em (**1**), temos a lista de artigos armazenada no cache de metadados para a conferência e intervalo selecionados pelo usuário. Essa lista apresenta os artigos em ordem não ascendente do número de citações. Para cada artigo são apresentados o título do artigo, autores, ano de publicação, citações e data da ultima

atualização no cache de metadados. Em (**2**), temos a estimativa da métrica *h-index* calculada para essa conferência e intervalo. Além disso, também é apresentado uma breve interpretação do valor de *h-index* retornado em relação a conferência consultada. Finalmente, em (**3**), temos a situação da coleta colaborativa de metadados dos artigos encontrados no GS para essa conferência e intervalo selecionados. Ao fim dessa coleta, essa área será utilizada para agradecer e notificar que o usuário completou sua colaboração no Live SHINE.

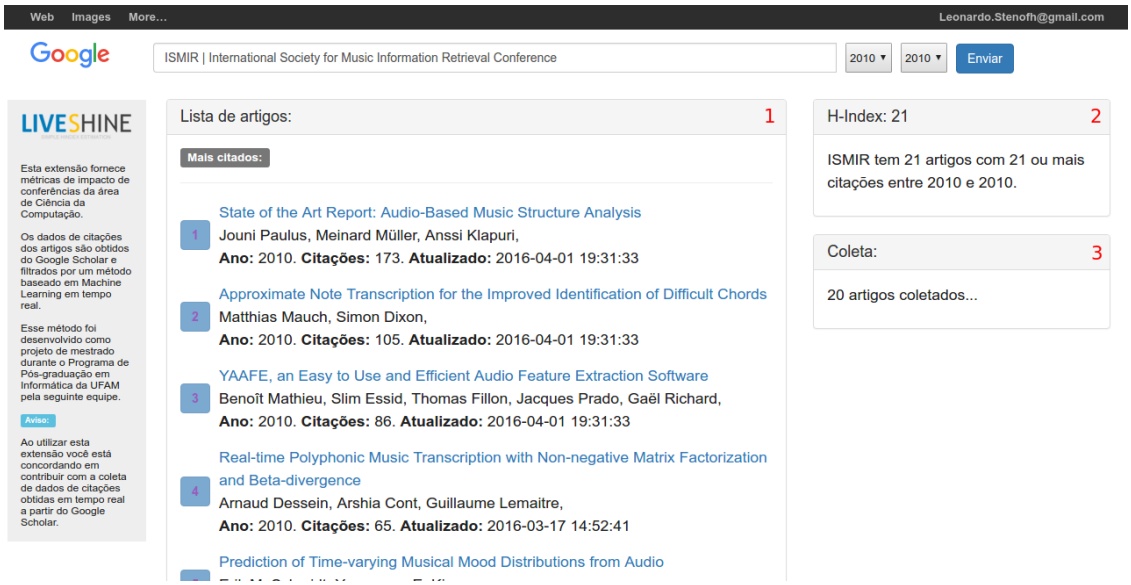


Figura 7.3: Interface de Consulta - Tela Estimativa do Impacto da Conferência.

7.3.2 Interface de Avaliação

A *Interface de Avaliação* é utilizada pelos usuários cadastrados no Live SHINE como Pesquisador Avaliador. Seu objetivo é possibilitar que os pesquisadores interessados possam contribuir avaliando a classificação feita pelo Live SHINE, bem como as listas dos artigos armazenadas no cache de metadados sobre as conferências de Ciência da Computação. Essa interface só pode ser acessada através do botão "Modo Avaliador", apresentado na interface padrão (descrita na Seção 7.3.1) para o usuário caso seu *e-mail* logado no site do *Google Scholar* esteja cadastrado em nossa base de dados como avaliador.

A Figura 7.4 apresenta a Interface de Avaliação. Semelhante a tela de consulta, o usuário deve preencher os campos de pesquisa e submeter uma consulta pelos artigos de uma determinada conferência que se encontra no cache de metadados. Basicamente, em (1) o usuário deve preencher o campo de pesquisa com o nome ou sigla da conferência,

em (2) o ano sobre o qual deseja obter a lista de artigos e então em (3) deve clicar em enviar. Observe que diferente da Interface de Consulta (descrita na Seção 7.3.1), essa interface não permite uma consulta por intervalo de anos. Fizemos dessa forma para evitar que esse tipo de usuário tenha que avaliar um número muito grande de artigos de uma mesma conferência, uma vez que isso poderia representar um grande inconveniente para os pesquisadores. Em (4), temos a lista dos artigos armazenados no cache para a conferência e ano selecionados, essa lista é dividida em dois grupos de artigos de acordo com sua classificação no Live SHINE. Em (5), temos os artigos classificados como positivos (+) e em (6) temos os artigos classificados como negativos (-) por um classificador automático SHINER treinado para essa conferência ou por um Pesquisador Avaliador. Para cada um dos grupos é possível filtrar e selecionar os artigos, bem como é possível trocar os artigos de grupos caso o Avaliador discorde de alguma classificação. Em (7), temos um resumo da lista de artigos recuperada no cache de metadados para essa conferência e ano, bem como uma legenda dos estados dos artigos. Os artigos marcados com (+) e (+?) são respectivamente artigos classificados como positivos por um avaliador ou classificador SHINER, enquanto que os artigos marcados com (-) e (-?) são respectivamente artigos classificados como negativos por um avaliador ou por um classificador SHINER. Finalmente, em (8) e (9), temos respectivamente a situação da coleta colaborativa de metadados e dados de citações dos artigos encontrados no GS para essa conferência e ano, e o botão responsável por salvar as alterações realizadas pelo avaliador.

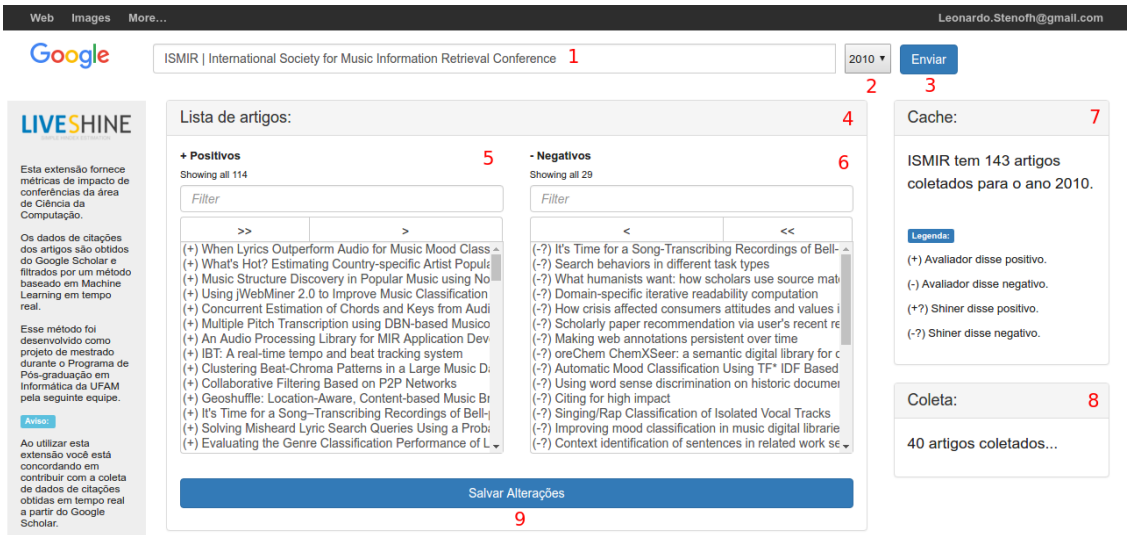


Figura 7.4: Interface de Avaliação.

Capítulo 8

Conclusão

Este capítulo apresenta as conclusões finais deste trabalho, bem como a direção principal que temos planejado para o futuro do Live SHINE.

8.1 Resultados Obtidos

Neste trabalho, apresentamos uma ferramenta capaz de fornecer índices de impacto de alta precisão de conferências científicas. Nossa motivação foi auxiliar os pesquisadores da área de Ciência da Computação, desenvolvendo uma ferramenta que entregue índices de impacto mais precisos, possibilitando que esses pesquisadores avaliem com maior confiança os eventos científicos para publicação de seus artigos. Para tanto, nossa proposta teve como objetivo superar os dois principais desafios envolvidos na tarefa de estimar o impacto de conferências com alta precisão: (i) obter as listas de artigos corretas das conferências e (ii) obter os dados de citações desses artigos atualizados.

Para superar o primeiro desafio, neste trabalho propomos o SHINER, um método baseado em técnicas de aprendizagem de máquina capaz de filtrar de maneira totalmente automática os dados fornecidos pela máquina de busca acadêmica *Google Scholar (GS)*. Assim, quando nossa ferramenta consulta o GS, obtemos as listas de artigos mais corretas possíveis das edições das conferências. Isso representa uma grande vantagem que nossa abordagem pode manter em relação a abordagem de outras ferramentas utilizadas para se obter índices de impacto de conferências, que acabam considerando listas de artigos distintas nos cálculos das métricas, resultando em índices discrepantes e imprecisos para uma mesma conferência e ano. Os experimentos apresentados no Capítulo 5 indicaram

que nosso método de filtragem foi realmente eficaz para um número representativo das conferências testadas, alcançando valores de média de precisão e revocação geral e por grupo de conferências de cerca de 0.9, o que sugere também que a superioridade nesses valores esperada em relação as listas das conferências de alto impacto não acontece.

Já para superar o segundo desafio, propomos uma coleta colaborativa de metadados e citações de artigos de conferências realizada através de nossa ferramenta Live SHINE. Nossa ferramenta foi concebida em forma de extensão para o *Navegador Web Google Chrome* e funciona sobre o site do *Google Scholar (GS)* a fim de aproveitar os dados fornecidos por essa máquina de busca acadêmica. Quando uma consulta é realizada através da extensão Live SHINE, é realizado um processo de coleta colaborativa automática para manter atualizado um sistema de cache de metadados sobre artigos e conferências. Assim, toda vez que um usuário utiliza o Live SHINE recebe imediatamente os índices de impacto da conferência desejada calculados com base em dados de citações atualizados, enquanto colabora com a atualização desse cache de metadados.

Assim podemos concluir que o Live SHINE é capaz de auxiliar os pesquisadores da comunidade de Ciência da Computação a resolver de maneira mais eficaz os problemas envolvidos na tarefa de obter índices de impacto de alta precisão sobre eventos científicos. Concluímos também que através do Live SHINE, esses pesquisadores podem assegurar que seus artigos sejam publicados em conferências de grande impacto, avaliadas por uma solução que entrega maior precisão e confiança nos índices fornecidos, tal como ocorre com os periódicos científicos.

8.2 Trabalhos Futuros

Apesar dos bons resultados alcançados com a ferramenta Live SHINE, é possível realizar trabalhos futuros em busca de aumentar ainda mais a precisão dos índices fornecidos por nossa ferramenta, bem como realizar pesquisas com o objetivo de aumentar o conjunto de conferências armazenadas no cache de metadados e melhorias na interação do usuário com as interfaces da ferramenta Live SHINE.

Primeiramente planejamos melhorar ainda mais os resultados obtidos pelo método SHINER através da introdução de técnicas de Aprendizagem de Máquina auxiliares. Por exemplo, estamos considerando a possibilidade de geração de *clusters* de artigos de confe-

rências similares e a realização de treinamento utilizando exemplos positivos e negativos desses *clusters*, de modo a melhorar os nossos níveis atuais de precisão.

Além disso, estamos considerando a possibilidade de aplicar técnicas de recuperação de informação para escalar o processo de coleta das listas de artigos e seus respectivos dados de citações. Atualmente nossa cobertura de conferências é relativamente baixa, embora o GS tenha uma alta cobertura de conferências de Ciência da Computação. Como utilizamos esse motor de busca acadêmico como fonte de dados sobre artigos e conferências, planejamos no futuro desenvolver um método que aumente consideravelmente nosso conjunto de conferências, uma vez que nossa abordagem atual demanda um tempo considerável para que uma conferência esteja presente em nosso cache de informações, e portanto, disponível para consulta do usuário.

Outro aspecto que planejamos melhorar é em relação as interfaces e recursos de interação com o usuário apresentados pela extensão Live SHINE. Atualmente temos um protótipo com interfaces simples e amigáveis ao usuário pesquisador, pois os recursos implementados foram pensados exclusivamente para atuar sobre os principais problemas abordados neste trabalho. No entanto, a fim de melhorar ainda mais a interação com o usuário, pensamos em implementar recursos adicionais para os pesquisadores. Por exemplo, pretendemos possibilitar que o Pesquisador Avaliador tenha de maneira facilitada dados complementares sobre a lista de artigos das conferências, como, informações sobre a conferência em si, o ano, local e a data do evento, os principais autores, os artigos mais citados da conferência e etc. Desse modo, acreditamos que essas informações podem ser mais úteis para o processo de avaliação das listas de artigos do que apenas os títulos dos artigos.

Finalmente, também planejamos aplicar outras métricas de impacto no Live SHINE. Apesar de inicialmente fornecemos apenas a métrica *h-index*, acreditamos que adicionando outras métricas de impacto baseadas em citações, como as apresentadas no Capítulo 2, podemos possibilitar que os pesquisadores possam avaliar as conferências de Ciência da Computação a partir de vários aspectos distintos, pois cada métrica tem uma forma diferente de avaliar as veículos científicos.

Referências Bibliográficas

[Alves et al., 2013] Alves, B. L., Benevenuto, F., & Laender, A. H. (2013). The role of research leaders on the evolution of scientific communities. In *Proceedings of the 22nd international conference on World Wide Web companion* (pp. 649–656).: International World Wide Web Conferences Steering Committee.

[Baeza-Yates & Ribeiro-Neto, 2013] Baeza-Yates, R. & Ribeiro-Neto, B. (2013). *Recuperação de Informação-: Conceitos e Tecnologia das Máquinas de Busca*. Bookman Editora.

[Bar-Ilan, 2008] Bar-Ilan, J. (2008). Which h-index? - a comparison of wos, scopus and google scholar. *Scientometrics*, 74(2), 257–271.

[Bar-Ilan, 2010] Bar-Ilan, J. (2010). Web of science with the conference proceedings citation indexes: The case of computer science. *Scientometrics*, 83(3), 809–824.

[Bergstrom et al., 2008] Bergstrom, C. T., West, J. D., & Wiseman, M. A. (2008). The eigenfactor? metrics. *The Journal of Neuroscience*, 28(45), 11433–11434.

[Bornmann & Daniel, 2005] Bornmann, L. & Daniel, H.-D. (2005). Does the h-index for ranking of scientists really work? *Scientometrics*, 65(3), 391–392.

[Bornmann & Daniel, 2007] Bornmann, L. & Daniel, H.-D. (2007). What do we know about the h index? *Journal of the American Society for Information Science and technology*, 58(9), 1381–1385.

[Bornmann et al., 2008] Bornmann, L., Mutz, R., & Daniel, H.-D. (2008). Are there better indices for evaluation purposes than the h index? a comparison of nine

different variants of the h index using data from biomedicine. *Journal of the American Society for Information Science and Technology*, 59(5), 830–837.

[Braun et al., 2006] Braun, T., Glänzel, W., & Schubert, A. (2006). A hirsch-type index for journals. *Scientometrics*, 69(1), 169–173.

[Caragea et al., 2006] Caragea, C., Wu, J., A. Ciobanu, K., Williams, J. F.-R., H.-H. Chen, Z. W., & Giles, C. L. (2006). Citeseerx: A scalable autonomous scientific digital library. In *Proceedings of the 1st International Conference on Scalable Information Systems*, InfoScale '06 New York, NY, USA: ACM.

[Caropreso et al., 2001] Caropreso, M. F., Matwin, S., & Sebastiani, F. (2001). : chapter A Learner-independent Evaluation of the Usefulness of Statistical Phrases for Automated Text Categorization, (pp. 78–102). Hershey, PA, USA: IGI Global.

[Chiu & Fu, 2010] Chiu, D. M. & Fu, T. Z. (2010). Publish or perish in the internet age: a study of publication statistics in computer networking research. *ACM SIGCOMM Computer Communication Review*, 40(1), 34–43.

[Damashek, 1995] Damashek, M. (1995). Gauging similarity with n-grams: Language-independent categorization of text. *Science*, 267(5199), 843.

[Di Iorio et al., 2015] Di Iorio, A., Giannella, R., Poggi, F., Peroni, S., & Vitali, F. (2015). Exploring scholarly papers through citations. In *Proceedings of the 2015 ACM Symposium on Document Engineering*, DocEng '15 (pp. 107–116). New York, NY, USA: ACM.

[Ding et al., 2013] Ding, C., Chen, Y., & Fu, X. (2013). Crowd crawling: towards collaborative data collection for large-scale online social networks. In *Proceedings of the first ACM conference on Online social networks* (pp. 183–188).: ACM.

[Egghe, 2006] Egghe, L. (2006). Theory and practise of the g-index. *Scientometrics*, 69(1), 131–152.

[Garfield, 1972] Garfield, E. (1972). Citation analysis as a tool in journal evaluation. *SCIENCE*, 178(4060), 471–479.

[Giles et al., 1998] Giles, C. L., Bollacker, K. D., & Lawrence, S. (1998). Citeseer: An automatic citation indexing system. In *Proceedings of the Third ACM Conference on Digital Libraries*, DL '98 (pp. 89–98). New York, NY, USA: ACM.

[Harzing, 2007] Harzing, A. (2007). Publish or perish. `http://www.harzing.com/pop.htm`.

[Harzing, 2008] Harzing, A.-W. (2008). Google scholar - a new data source for citation analysis. *University of Melbourne*.

[Harzing, 2010] Harzing, A.-W. (2010). *The publish or perish book*. Tarma Software Research Melbourne.

[Harzing, 2013] Harzing, A.-W. (2013). A preliminary test of google scholar as a source for citation data: a longitudinal study of nobel prize winners. *Scientometrics*, 94(3), 1057–1075.

[Harzing, 2014] Harzing, A.-W. (2014). A longitudinal study of google scholar coverage between 2012 and 2013. *Scientometrics*, 98(1), 565–575.

[Harzing & Van Der Wal, 2009] Harzing, A.-W. & Van Der Wal, R. (2009). A google scholar h-index for journals: An alternative metric to measure journal impact in economics and business. *Journal of the American Society for Information Science and Technology*, 60(1), 41–46.

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

[Houzanme, 2012] Houzanme, U. T. (2012). Google scholar versus google scholar: Among publish or perish, scholarometer, and my citations, which citation count tool is telling which truth? In *Science and the Internet* (pp. 223–236). DÃ¼sseldorf: DÃ¼sseldorf University Press.

[Jacsó, 2009] Jacsó, P. (2009). Calculating the h-index and other bibliometric and scientometric indicators from google scholar with the publish or perish software. *Online Information Review*, 33(6), 1189–1200.

[Jacsó, 2012] Jacsó, P. (2012). Google scholar author citation tracker: is it too little, too late? *Online Information Review*, 36(1), 126–141.

[Kaur et al., 2014] Kaur, J., JafariAsbagh, M., Radicchi, F., & Menczer, F. (2014). Scholarometer: A system for crowdsourcing scholarly impact metrics. In *Proceedings of the 2014 ACM Conference on Web Science*, WebSci '14 (pp. 285–286). New York, NY, USA: ACM.

[Laender et al., 2008] Laender, A. H. F., de Lucena, C. J. P., Maldonado, J. C., de Souza e Silva, E., & Ziviani, N. (2008). Assessing the research and education quality of the top brazilian computer science graduate programs. *SIGCSE Bull.*, 40(2), 135–145.

[Lima et al., 2013] Lima, H., Silva, T. H., Moro, M. M., Santos, R. L., Meira Jr, W., & Laender, A. H. (2013). Aggregating productivity indices for ranking researchers across multiple areas. In *Proceedings of the 13th ACM/IEEE-CS joint conference on Digital libraries* (pp. 97–106).: ACM.

[Martins et al., 2009] Martins, W. S., Gonçalves, M. A., Laender, A. H., & Pappa, G. L. (2009). Learning to assess the quality of scientific conferences: A case study in computer science. In *Proceedings of the 9th ACM/IEEE-CS Joint Conference on Digital Libraries*, JCDL '09 (pp. 193–202). New York, NY, USA: ACM.

[Meho & Yang, 2007] Meho, L. I. & Yang, K. (2007). Impact of data sources on citation counts and rankings of lis faculty: Web of science versus scopus and google scholar. *Journal of the american society for information science and technology*, 58(13), 2105–2125.

[Noorden, 2014] Noorden, R. V. (2014). The decline and fall of Microsoft Academic Search. `http://blogs.nature.com/news/2014/05/the-decline-and-fall-of-microsoft-academic-search.html`. [Online; accessed 1-March-2015].

[Olston & Najork, 2010] Olston, C. & Najork, M. (2010). Web crawling. *Foundations and Trends in Information Retrieval*, 4(3), 175–246.

[Page et al., 1999] Page, L., Brin, S., Motwani, R., & Winograd, T. (1999). The pagerank citation ranking: bringing order to the web.

[Pant et al., 2004] Pant, G., Tsioutsiouliklis, K., Johnson, J., & Giles, C. L. (2004). Panorama: extending digital libraries with topical crawlers. In *Proceedings of the 4th ACM/IEEE-CS joint conference on Digital libraries* (pp. 142–150).: ACM.

[Rosenstreich & Wooliscroft, 2009] Rosenstreich, D. & Wooliscroft, B. (2009). Measuring the impact of accounting journals using google scholar and the g-index. *The British Accounting Review*, 41(4), 227–239.

[Sebastiani, 2002] Sebastiani, F. (2002). Machine learning in automated text categorization. *ACM Comput. Surv.*, 34(1), 1–47.

[Silva et al., 2009] Silva, A. J., Gonçalves, M. A., Laender, A. H., Modesto, M. A., Cristo, M., & Ziviani, N. (2009). Finding what is missing from a digital library: A case study in the computer science field. *Information Processing & Management*, 45(3), 380–391.

[Sun et al., 2009] Sun, A., Lim, E.-P., & Liu, Y. (2009). On strategies for imbalanced text classification using svm: A comparative study. *Decision Support Systems*, 48(1), 191–201.

[Vardi, 2009] Vardi, M. Y. (2009). Conferences vs. journals in computing research. *Commun. ACM*, 52(5), 5–5.

[Vasilescu et al., 2013] Vasilescu, B., Mens, T., & Serebrenik, A. (2013). Mining software engineering conference data. In *Proceedings of the Working Conference on Mining Software Repositories (MSR 2013)*.

[Wang et al., 2007] Wang, X., McCallum, A., & Wei, X. (2007). Topical n-grams: Phrase and topic discovery, with an application to information retrieval. In *Data Mining, 2007. ICDM 2007. Seventh IEEE International Conference on* (pp. 697–702).: IEEE.

[Witten & Frank, 2005] Witten, I. H. & Frank, E. (2005). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.

[Wu et al., 2014] Wu, Z., Wu, J., Khabsa, M., Williams, K., Chen, H.-H., Huang, W., Tuarob, S., Choudhury, S. R., Ororbia, A., Mitra, P., & Giles, C. L. (2014). Towards building a scholarly big data platform: Challenges, lessons and opportunities. In *Proceedings of the 14th ACM/IEEE-CS Joint Conference on Digital Libraries*, JCDL '14 (pp. 117–126). Piscataway, NJ, USA: IEEE Press.

[Zhang, 2011] Zhang, L. (2011). Proceeding papers or journal articles? a comparative analysis on computer science versus economics, business and management. In *Artificial Intelligence, Management Science and Electronic Commerce (AIMSEC), 2011 2nd International Conference on* (pp. 1319–1322).: IEEE.

[Zhuang et al., 2007] Zhuang, Z., Elmacioglu, E., Lee, D., & Giles, C. L. (2007). Measuring conference quality by mining program committee characteristics. In *Proceedings of the 7th ACM/IEEE-CS joint conference on Digital libraries* (pp. 225–234).: ACM.