ABORDAGENS PARA AVALIAÇÃO AUTOMÁTICA DA QUALIDADE DE CONFERÊNCIAS CIENTÍFICAS: UM ESTUDO DE CASO EM CIÊNCIA DA COMPUTAÇÃO

WAISTER SILVA MARTINS

ABORDAGENS PARA AVALIAÇÃO AUTOMÁTICA DA QUALIDADE DE CONFERÊNCIAS CIENTÍFICAS: UM ESTUDO DE CASO EM CIÊNCIA DA COMPUTAÇÃO

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

Orientador: Marcos André Gonçalves

Belo Horizonte - MG

26 de janeiro de 2010

Martins, Waister Silva
M379a Abordagens para avaliação automática da qualidade de conferências científicas: um estudo de caso em Ciência da Computação / Waister Silva Martins. — Belo Horizonte - MG, 2010
viii, 63 f. ; 29cm

Dissertação (mestrado) — Universidade Federal de Minas Gerais
Orientador: Marcos André Gonçalves

1. Aprendizado de Máquina. 2. Bibliometria.
3. Bibliotecas Digitais. 4. Classificação de Conferências.
I. Título.

CDU 519.6*73(043)

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

FOLHA DE APROVAÇÃO

Abordagens para Avaliação Automática de Conferências Científicas:
Um Estudo de Caso em Ciência da Computação

WAISTER SILVA MARTINS

Dissertação defendida e aprovada pela banca examinadora constituída pelos Senhores:

PROF. MARCOS ANDRÉ GONÇALVES - Orientador
Departamento de Ciência da Computação - UFMG

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

PROF. RICARDO DE OLIVEIRA ANIDO
Instituto de Computação - UNICAMP

PROFA. GISELE LOBO PAPPA
Departamento de Ciência da Computação - UFMG

Belo Horizonte, 27 de março de 2009.

Resumo

Avaliar a qualidade de conferências científicas é um importante serviço que pode ser provido por bibliotecas digitais e sistemas similares, principalmente para campos do conhecimento dinâmicos como a Ciência da Computação e a Engenharia Elétrica. Entretanto, a maioria das abordagens existentes está voltada para a avaliação de periódicos. Neste trabalho, propomos duas abordagens para avaliar automaticamente a qualidade de conferências. Na primeira abordagem, realizamos uma análise das deficiências das métricas baseadas em citações bibliográficas usadas para periódicos e propomos um conjunto de novas métricas especialmente projetadas para capturar aspectos intrínsecos e importantes relacionados a conferências, tais como longevidade, popularidade, prestígio e periodicidade. Para demonstrar a efetividade das métricas propostas, conduzimos dois conjuntos de experimentos. No primeiro, nossas métricas foram contrastadas com um gabarito produzido por um grande número de especialistas. Então, utilizamos nossas métricas para classificar essas conferências em níveis de qualidade pré-estabelecidos, também de acordo com o gabarito. Nossas métricas obtiveram ganhos de até 8,4% na comparação de similaridade e 7,8% na acurácia quando comparadas com as métricas tradicionais para classificação de periódicos.

Na segunda abordagem, identificamos um grande número de características (por exemplo, citações, tradição, taxas de submissão e aceitação, reputação dos membros do comitê de programa) que podem ser usadas como critérios para avaliar a qualidade de conferências científicas e estudamos como essas características podem ser automaticamente combinadas através de técnicas de aprendizado de máquina para executar essa tarefa efetivamente. Entre nossos principais resultados, podemos citar: (1) a separação de conferências de alta qualidade de conferências de média e baixa qualidade pode ser executada efetivamente, mas separar os dois últimos tipos é uma tarefa muito díficil e (2) as características baseadas em citações seguidas pelas associadas com a tradição da conferência são as mais importantes para essa tarefa.

Em suma, as principais contribuições desta dissertação são: (*i*) estudar a eficácia, para avaliação de conferências, de métricas baseadas em citações bibliográficas projetadas para periódicos; (*ii*) apresentar um conjunto de novas métricas baseadas em citações bibliográficas projetadas especificamente para avaliação de conferências e que capturam aspectos importantes que não são considerados pelas métricas existentes (para periódicos); (*iii*) apresentar e detalhar um conjunto de características que podem ser utilizadas como indicadores de qualidade para conferências científicas; (*iv*) estudar como essas características podem ser

combinadas através de técnicas de aprendizado de máquina para automática e efetivamente classificar conferências de acordo com sua a qualidade; e (v) apresentar uma análise detalhada das dificuldades inerentes ao problema de classificação de conferências de acordo com a sua qualidade.

Abstract

Assessing the quality of scientific conferences is an important and useful service that can be provided by digital libraries and similar systems, mainly for dynamic fields such as Computer Science and Electric Engineering. However, the majority of the existing approaches has been proposed for measuring the quality of journals. In this MSc dissertation we propose two distinct approaches to automatically assess the quality of conferences. In the first one, we depart from a deep analysis of the deficiencies of citation-based metrics to assess the quality of journals and propose a new set of quality metrics specially designed to capture intrinsic and important aspects related to conferences such as longevity, popularity, prestige, and periodicity. To demonstrate the effectiveness of our proposed metrics, we have conducted two sets of experiments. In the first one, our metrics were used to rank a set of Computer Science conferences and the results were contrasted against a "gold standard" produced by a large group of specialists. Then, we used our metrics to classify these conferences with respect to some pre-established quality levels, also according to the gold standard. Our metrics obtained gains up to 8.4% in ranking similarity and 7.8% in classification accuracy when compared to standard journal quality metrics.

In the second approach, we characterize a large number of features (e.g., citations, tradition, submission and acceptance rates, reputation of the program committee members) that can be used as criteria to assess the quality of scientific conference and study how these features can be automatically combined using machine learning techniques to effectively perform this task. Among our several findings, we can cite that: (1) separating high quality conferences from medium and low quality ones can be performed quite effectively, but separating the last two types is a much harder task; and (2) citation features followed by those associated with the tradition of the conference are the most important ones for the task.

Thus, in summary, the major contributions of this MSc dissertation are: (*i*) a study about the relative performance of existing journal metrics in assessing the quality of scientific conferences; (*ii*) the proposal of a set of new metrics based on bibliographic citations specifically designed to evaluate the conference, which capture intrinsic and important aspects related to conferences that are not considered by existing metrics (for journals); (*iii*) the characterization of a large number of features that can be used as criteria to assess the quality of scientific conferences; (*iv*) a study of how these several features can be combined by means of machine learning techniques to automatically and effectively classify conferences; and (*v*) a deep analysis and discussion about the relative difficulty of the problem.

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

Um importante serviço que pode ser provido por bibliotecas digitais de artigos científicos e sistemas similares é informar a qualidade dos veículos de publicação cobertos nesses repositórios. A qualidade do veículo é uma informação essencial para pesquisadores, visto que ajuda na busca por trabalhos de alta qualidade que podem influenciar pesquisas em andamento ou mesmo na decisão para qual veículo um trabalho será submetido. A qualidade do veículo também pode ser utilizada para auxiliar na decisão de promoções e premiações de pesquisadores, como também no processo de decisão relativo a financiamentos e bolsas concedidas por agências de fomento.

O foco dos principais estudos que abordam o problema de avaliar a qualidade de veículos de publicação são os periódicos, visto que na maioria dos campos do conhecimento são eles os veículos de publicação de maior importância. Existem diversas métricas baseadas na análise de citações propostas na literatura para estimar a qualidade de periódicos. A análise de citações é o método mais comum da bibliometria e consiste em analisar o padrão e a freqüência das citações bibliográficas entre artigos científicos e livros. Entre as métricas existentes podemos citar, o Fator de Impacto da Thomson (Amin e Mabe, 2000), o *Y-Factor* (Bollen et al., 2006) e o *H-index* (Hirsch, 2005). O Fator de Impacto da Thomson é o método mais popular e aceito, entretanto alguns trabalhos questionam a sua utilização (Saha et al., 2003; Seglen, 1997; Yu e Wang, 2007), devido, principalmente, à dependência exclusiva da análise de citações.

Em campos em que há disseminação rápida do conhecimento, como a Ciência da Computação e a Engenharia Elétrica, as conferências[1] assumem um papel essencial na difusão dos resultados de pesquisas científicas (Patterson, 2004). Como mostrado em (Laender et al., 2008), a produção científica em Ciência da Computação é fortemente centrada em conferências com uma razão de 2,49 de artigos em conferências para cada artigo em periódico.

Além de possuírem um processo de publicação mais rápido, com revisões recebidas em poucos meses e anais publicados rapidamente, as conferências são em geral de grande importância, sendo algumas muito concorridas com taxas de aceitação próximas a 10%, e patrocinadas por

[1] Denominaremos genericamente de conferência qualquer tipo de evento científico, tais como, congresso, simpósio, workshop, etc.

entidades de prestígio como o *Institute of Electrical and Electronics Engineers*[2] (IEEE) e a *Association for Computing Machinery*[3] (ACM). Ademais, as conferências são de grande impacto e visibilidade, com alto padrão de novidade nos trabalhos publicados. Em alguns casos, os melhores artigos apresentados nas conferências são solicitados em versão estendida para publicação em um periódico. Contudo, não existem critérios e métricas consolidados para avaliar automaticamente a qualidade de conferências.

Outra abordagem muito utilizada para medir a qualidade de veículos de publicação, que pode ser aplicada a periódicos e conferências, é consultar e analisar a opinião de especialistas de um determinado campo do conhecimento. Essa abordagem produz uma classificação muito boa, quando o número de especialistas é significativo, já que os especialistas utilizam toda a sua experiência e conhecimento para prover opiniões precisas sobre a qualidade dos veículos. Contudo, o custo para se consultar e coletar a opinião de um grande número de especialistas pode ser muito alto, de modo que em campos do conhecimento muito dinâmicos, nos quais veículos deixam de existir, são criados e mudam de qualidade freqüentemente, esta abordagem torna-se impraticável. Além disso, essa abordagem pode causar algumas distorções entre os sub-campos de um campo do conhecimento muito amplo, em que existem um número desproporcional de pessoas e veículos de publicação em cada sub-campo, dependendo do modo que essa consulta será conduzida, como vamos mostrar em nossos experimentos no Capítulo 5.

Neste trabalho propomos duas abordagens para classificação automática de conferências científicas de Ciência da Computação. Na primeira, estudamos a aplicação das métricas mais tradicionais de periódicos na classificação de conferências. As deficiências encontradas motivaram-nos a propor um conjunto de novas métricas projetadas especificadamente para capturar aspectos intrínsecos e importantes para conferências como, por exemplo, longevidade, popularidade, prestígio, e periodicidade.

Na segunda abordagem, primeiramente caracterizamos alguns dos critérios que um especialista utiliza para dar a sua opinião sobre a qualidade das conferências, dentre os quais podemos citar tradição, visibilidade, taxa de aceitação, reputação dos membros do comitê de programa, entre outros. Então, estudamos o poder discriminativo de diversos desses critérios, assim como a combinação deles, para distinguir conferências de alta, média e baixa qualidade. Através de classificadores baseados em técnicas de aprendizado de máquina (Mitchell, 1997), combinamos esses critérios de forma que pudéssemos encontrar a melhor combinação de evidências para a tarefa de avaliar a qualidade de conferências.

Para conduzir o nosso estudo, coletamos dados sobre 194 Conferências de Ciência da Computação de bibliotecas digitais, sítios de conferências e outras fontes de dados da Web. Em nossos experimentos, para avaliar a efetividade de nossas abordagens, contrastamos os resultados com um gabarito produzido por um grande número de especialistas. Uma descrição aprofundada desse gabarito é apresentada na Seção 3.1.1.1.

[2] www.ieee.org
[3] www.acm.org

1.1 Objetivos e Contribuições

O principal objetivo desta dissertação é apresentar duas abordagens para classificação automática de conferências. A primeira é baseada em citações bibliográficas. Essa é uma abordagem muito prática que pode ser facilmente empregada por bibliotecas digitais e sistemas similares, visto que os dados de citações bibliográficas podem comumente ser encontrados nesses repositórios. A segunda abordagem é inspirada no método executado implicitamente por especialistas ao julgar manualmente uma conferência. Mensuramos quais são as características mais importantes para o processo de classificação e estudamos como combinar essas características para melhor avaliar a qualidade das conferências.

As principais contribuições desta dissertação são: (*i*) estudar a eficácia, para avaliação de conferências, de métricas baseadas em citações bibliográficas projetadas para periódicos; (*ii*) apresentar um conjunto de novas métricas baseadas em citações bibliográficas projetadas especificamente para avaliação de conferências e que capturam aspectos importantes que não são considerados pelas métricas existentes (para periódicos); (*iii*) apresentar e detalhar um conjunto de características que podem ser utilizadas como indicadores de qualidade para conferências científicas; (*iv*) estudar como essas características podem ser combinadas através de técnicas de aprendizado de máquina para automática e efetivamente classificar conferências de acordo com sua a qualidade; e (*v*) apresentar uma análise detalhada das dificuldades inerentes ao problema de classificação de conferências de acordo com a sua qualidade.

As três primeiras contribuições foram apresentadas em (Martins et al., 2008) e estão detalhadas em (Martins et al., 2009b). As duas contribuições finais são apresentadas em (Martins et al., 2009a).

1.2 Trabalhos Relacionados

Como mencionado anteriormente, a maioria dos trabalhos existentes para avaliar a qualidade de veículos de publicação são focados em periódicos. Isso ocorre porque esse tipo de veículo é o mais importante na maioria dos campos do conhecimento. Entretanto, em campos que necessitam de uma disseminação rápida do conhecimento, como a Ciência da Computação e a Engenharia Elétrica, conferências assumem um importante papel, visto que o processo de publicação nesse tipo de veículo normalmente é mais ágil do que em periódicos. Contundo, não existem métricas ou critérios consolidados para avaliar a qualidade de conferências.

A forma mais comum de se avaliar a reputação e a qualidade de um veículo de publicação é através da análise de citações. Na análise de citações, a qualidade do veículo está diretamente relacionada com as citações recebidas pelos artigos publicados nesse veículo. Existem diversas métricas baseadas em citações propostas na literatura, por exemplo, o Fator de Impacto da Thompson (Amin e Mabe, 2000), o *Y-Factor* (Bollen et al., 2006) e o *H-index* (Hirsch, 2005). Entre essas métricas, o Fator de Impacto é a mais largamente utilizada e tem sido adotada por diversas bibliotecas digitais. Entretanto, desde a sua proposta, o Fator de Impacto é largamente criticado devido à sua dependência exclusiva da contagem de citações (Saha et al.,

2003). Em (Seglen, 1997) são discutidos algumas limitações da validade e aplicabilidade do Fator de Impacto. Para lidar com as limitações do Fator de Impacto, diversas outras métricas foram propostas. Discutiremos algumas dessas ao longo deste e do Capítulo 4.

Uma análise profunda das citações das duas principais conferências de Bancos de Dados (SIGMOD e VLDB) e de três dos seus principais periódicos (TODS, VLDB Journal e SIGMOD Record) pode ser encontrada em (Rahm e Thor, 2005). Nesse trabalho é mostrado que essas conferências possuem um número total de citações maior do que os periódicos, um resultado que reforça a importância das conferências. Um trabalho baseado na análise de citações é apresentado em (Larsen e Ingwersen, 2006), no qual é proposto uma abordagem que utiliza as citações bibliográficas para construção de *rankings* em bibliotecas digitais.

Alguns trabalhos apresentam alternativas à análise de citações para classificação de veículos de publicação. Uma abordagem para determinar o impacto de periódicos baseada em métricas de centralidade de redes sociais é apresentada em (Bollen et al., 2005). A partir do *log* de *downloads* de artigos (*usage data*) foi construída uma rede de relacionamento entre os periódicos. Nessa rede, um relacionamento entre dois periódicos (X e Y) é estabelecido quando um usuário realiza o *download* de um artigo pertencente ao periódico X e também realiza o *download* de um artigo pertencente ao periódico Y em um intervalo de tempo pré-definido. Os resultados mostram que essa rede possui características ligeiramente diferentes da rede construída a partir das citações entre os artigos, o que fornece um ângulo diferente para classificação de veículos de publicação, sendo, portanto, uma alternativa às métricas tradicionais baseadas em citações.

Em (Bollen et al., 2008) e (Bollen e de Sompel, 2008), os autores apresentam um estudo comparativo das métricas baseadas em citações e em *usage data*. Uma grande quantidade de eventos de *usage data* foram extraídos do projeto MESUR[4] para validar a utilização de métricas baseadas em *usage data* na classificação de periódicos. Um resultado interessante é que as métricas baseadas em *usage data* estão correlacionadas com diversas métricas baseadas em citações. Entretanto, o grande potencial dessas métricas está na riqueza da estrutura, na densidade das conexões e na capacidade de modelar diferentes áreas do conhecimento através da rede contruída com *usage data*.

Em (Yan e Lee, 2007) um método para avaliar a qualidade de conferências é proposto, no qual os autores consideram que em cada campo do conhecimento existe um conjunto reconhecível de artigos de boa qualidade ("semente"). Assim, para determinar a qualidade de uma conferência basta verificar a quantidade de artigos de qualidade similar às sementes. Para isso, são propostas três métricas que basicamente consideram o número de autores que um artigo tem em comum com as sementes. Em (Souto et al., 2007) os autores desenvolveram um modelo de classificação para auxiliar na avaliação (semi-)automática de conferências de Ciência da Computação com base em ontologias e regras de inferência.

Em (Zhuang et al., 2007) é apresentado um trabalho que utiliza classificadores baseados em técnicas de aprendizado de máquina para medir a qualidade de conferências de Ciência da Computação. As técnicas de aprendizado de máquina são capazes de aprender através de

[4] http://www.mesur.org/MESUR.html

experiências e observações analíticas, sendo amplamente empregadas em problemas de classificação na área de recuperação da informação (Harrison e Worden, 2007; Joachims, 1998; Smith e Jiang, 2007). Esse trabalho assume a hipótese de que a qualidade de uma conferência está diretamente relacionada à qualidade do comitê de programa. Um conjunto de características foi minerado a partir dos dados do comitê de programa e combinado através das técnicas de aprendizado de máquina. Essa abordagem é capaz de distinguir conferências de alta qualidade de "outras". Essa é outra abordagem alternativa à análise de citações. Contudo, esse trabalho limita-se a avaliar conferências de Bancos de Dados e estuda apenas a tarefa de separar conferências de alta qualidade das demais.

Nosso trabalho incluí duas abordagens distintas para o problema de avaliar a qualidade de conferências automaticamente. Na primeira foram desenvolvidas métricas, baseadas em citações, específicas para avaliação de conferências. Abordagens que exploram a análise de citações são muito práticas, principalmente em comparação com outras abordagens que utilizam dados difíceis de serem obtidos como, por exemplo, taxa de aceitação, comitê de programa, patrocinadores. Nosso estudo também é pioneiro na aplicação de métricas de periódicos na classificação de conferências e na criação de métricas baseadas em citações projetadas especificadamente para avaliação de conferências, visto que não fomos capazes de encontrar trabalhos na literatura com esse foco.

A segunda abordagem foi inspirada em (Zhuang et al., 2007). Porém, abordamos um problema mais difícil, já que não ficamos restritos a diferenciar apenas conferências de alta qualidade das demais, mas focamos em distinguir conferências de alta, média e baixa qualidade. Além disso, nossa abordagem não é limitada apenas aos critérios baseados no comitê de programa, mas explora um grande número de características como critérios para avaliar a qualidade de conferências, as quais são combinadas através de técnicas de aprendizado de máquina para automática e eficientemente classificá-las. Também realizamos uma análise detalhada das dificuldades inerentes ao problema com base nos resultados de diferentes experimentos. Nessa abordagem assumimos que o problema de se obter diferentes conjuntos de dados sobre as conferências está resolvido. Assim, exploramos um grande leque de características, além daquelas baseadas no comitê de programa ou nas citações, e incluímos conferências de diversos grupos da Ciência da Computação.

1.3 Organização da Dissertação

O restante desta dissertação está organizado da seguinte forma. O Capítulo 2 apresenta alguns conceitos básicos importantes. O Capítulo 3 descreve os conjuntos de dados utilizados em nossos experimentos e apresenta um estudo sobre as características que podem ser utilizadas para avaliar a qualidade de conferências. O Capítulo 4 apresenta nossa abordagem baseada em citações bibliográficas, para a qual foi proposto um conjunto de novas métricas projetadas especificamente para avaliação da qualidade de conferências. O Capítulo 5 apresenta nossa abordagem para avaliação de conferências baseada nas técnicas de aprendizado de máquina e os resultados experimentais desta abordagem. Por fim, o Capítulo 6 apresenta as conclusões

desta dissertação e sugere temas para trabalhos futuros.

Capítulo 2

Conceitos Básicos

Este capítulo introduz alguns conceitos básicos usados nos capítulos subseqüentes. A Seção 2.1 apresenta a rede de citações entre conferências modelada como um grafo. Na Seção 2.2 descrevemos sucintamente alguns classificadores baseados em técnicas de aprendizado de máquina. Por fim, a Seção 2.3 apresenta algumas das principais métricas de avaliação utilizadas em nossos experimentos.

2.1 Redes de Citações entre Conferências

Coleções de artigos científicos em que os artigos são citados diretamente podem ser modeladas como um grafo direcionado $G_A = (V_A, A_A)$, onde V_A é o conjunto de vértices e representa os artigos, e A_A é o conjunto de arestas direcionadas que representa as citações entre esses artigos. Essas arestas não possuem pesos, visto que um artigo pode citar outro no máximo uma vez. Na rede de citações entre conferências, os artigos são agrupados de acordo com a conferência em que foram publicados. Podemos representar essa rede como um grafo direcionado $G_C = (V_C, A_C)$, porém esse é um grafo ponderado, onde o conjunto de vértices V_C representa o conjunto de conferências e o conjunto de arestas A_C representa as citações entre as conferências. Existe uma aresta entre as conferências X e Y, se algum artigo da conferência X cita algum artigo da conferência Y. Note que um mesmo artigo da conferência X pode citar vários artigos da conferência Y e outros artigos da conferência X podem, também, citar artigos da conferência Y. O número de citações dos artigos da conferência X para os artigos da conferência Y determina o peso dessa aresta.

A Figura 2.1 mostra uma rede citações entre artigos e exemplifica o processo de transformação dessa rede em uma rede de citações entre conferências. As métricas existentes tentam capturar diferentes características da rede de citações entre conferências. Nessa figura, as conferências B, C e D, por exemplo, possuem dois, seis e três artigos publicados, respectivamente, e exatamente três citações cada. Quais dessas conferências são as de maior qualidade, as conferências que possuem uma alta relação de citações por artigo ou as conferências que possuem mais artigos publicados e citações? Para estudar essas e outras questões relativas às rede de citações entre conferências empregamos diversas métricas existentes para redes citações de pe-

riódicos (Seção 4.1) e propomos um novo conjunto de métricas desenhadas especificadamente para redes de citações entre conferências (Seção 4.2).

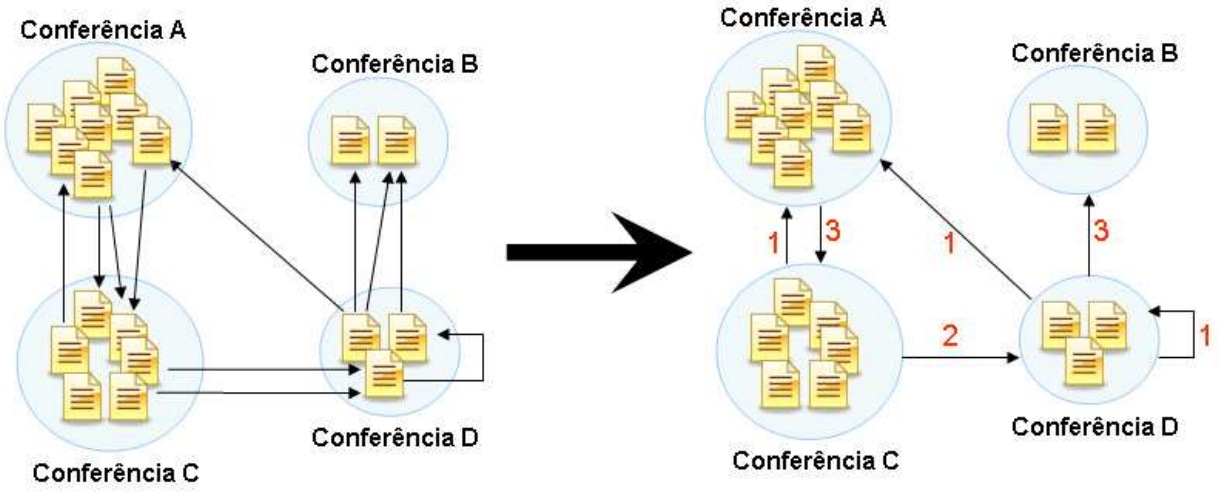

Figura 2.1: Exemplo de uma rede de citações entre conferências.

2.2 Classificadores Baseados em Técnicas de Aprendizado de Máquina

Aprendizado de Máquina é um sub-campo da Inteligência Artificial que estuda como desenvolver algoritmos e técnicas que melhoram o próprio desempenho em alguma tarefa através de experiências e observações analíticas (Mitchell, 1997). O objetivo principal das técnicas de aprendizado de máquina é descobrir regras gerais em grandes conjuntos de dados, que permitam extrair informação desses dados automaticamente.

De modo geral, as técnicas de aprendizado de máquina podem ser divididas em dois tipos: aprendizado supervisionado e aprendizado não supervisionado. No aprendizado supervisionado um conjunto de instâncias rotulados com suas respectivas classes é fornecido para o algoritmo de aprendizado (ou indutor). Um classificador é construído com base nos atributos das instâncias pertencentes a cada classe, permitindo julgar a classe de outras instâncias, que não pertencem ao conjunto fornecido inicialmente, a partir dos seus atributos. No aprendizado não supervisionado, a classe de cada instância não é conhecida ou não está disponível. O algoritmo de aprendizado analisa os atributos do conjunto de instâncias fornecidos procurando por padrões e tendências que permitam gerar agrupamentos ou *clusters*.

Uma de nossas abordagens utiliza técnicas de aprendizado supervisionado para classificação de conferências de acordo com a sua qualidade. Os indutores recebem como entrada um conjunto de conferências rotuladas em classes (denominaremos de categorias) de acordo com a sua qualidade, sendo que cada conferência possui um conjunto de atributos (características).

Técnicas de aprendizado de máquina vêm sendo empregadas com sucesso em alguns problemas de recuperação de informação, entre eles, o de classificação (Harrison e Worden, 2007;

Smith e Jiang, 2007; Joachims, 1998).

Nesta dissertação, exploraremos diferentes técnicas de aprendizado de máquina, visto que, dependendo das características específicas do problema, cada uma delas alcança melhores resultados. Consideraremos os seguintes métodos de aprendizado de máquina: árvores de decisão, métodos Bayesianos, regras de associação e *Support Vector Machines* (Schölkopf et al., 1999). Além disso, utilizamos uma técnica conhecida como *bagging* (*bootstrap aggregating*) (Quinlan, 1996) para tentar melhorar os resultados dos classificadores. *Bagging* é um meta-algoritmo que busca melhorar a acurácia na classificação produzida por algoritmos de aprendizado de máquina, ajudando, também, a diminuir a variância e a possibilidade de *overfitting*. *Overfitting* é quando um algoritmo de aprendizado se especializa demais no conjunto de treinamento e perde a capacidade de generalização, ou seja, os padrões aprendidos não representam o comportamento de outros conjuntos de dados. A seguir, descrevemos brevemente esses métodos.

Árvores de decisão: constituem um dos métodos mais utilizados e práticos para realizar inferência indutiva. O modelo de predição é representado utilizando uma estrutura de dados denominada árvore, na qual os nós correspondem aos atributos, uma aresta para os filhos representa um possível valor desse atributo e uma folha indica a classe de uma instância, assim como o caminho (combinação dos valores dos atributos) até a raiz. Essa árvore é formada através da divisão do conjunto de entrada em subconjuntos disjuntos com base em um ou mais valores dos atributos. Esse processo é repetido em cada subconjunto derivado de maneira recursiva. A recursão termina quando todos (ou quase todos) os elementos de um subconjunto são da mesma classe. A escolha dos elementos utilizados nas divisões é conduzidas através de medidas heurísticas, por exemplo, ganho de informação e razão do ganho de informação (Quinlan, 1993). Existem diversos algoritmos baseados em árvores de decisão, entre eles podemos citar: ID3, C4.5 (Mitchell, 1997) e Floresta Aleátoria (*Random Forest*) (Breiman, 2001).

Métodos Bayesianos: são métodos baseados em abordagens probabilistícas para inferência e que assumem que as quantidades de interesse são regidas pela distribuição de probabilidade e que decisões ótimas podem ser feitas pelo racicíonio sobre estas probabilidades. Os métodos Bayesianos são importantes porque provêem uma abordagem quantitativa para ponderar a probabilidade de um evento ocorrer, fornecendo um arcabouço para auxiliar as operações de outros algoritmos que não manipulam explicitamente probabilidades.

Regras de Associação: o principal objetivo das regras de associação é detectar relacionamentos ou associações entre valores específicos de atributos de uma classe em grandes conjuntos de dados. Essa técnica permite descobrir padrões escondidos entre conjuntos de dados muito grandes, por exemplo, "o consumidor que compra o produto A também compra os produtos B e C". Geralmente, um grande conjunto de regras são criadas e, então, apenas um subconjunto contendo as regras de maior qualidade são escolhidas para serem utilizadas no processo de predição. A qualidade de uma regra pode ser determinada, por exemplo, por sua freqüência nesse conjunto de dados. Classificadores baseados em regras de associação são muito utilizados em aplicações reais devido à fácil interpretação das regras.

Support Vector Machines: constituem um método que visualiza os dados de entrada como dois conjuntos de vetores em um espaço N-dimensional (N é o número de atributos). *Support Vector Machines* constroem um hiperplano de separação nesse espaço que maximiza a margem entre as duas classes. Para classificação, basta determinar a posição relativa da instância em relação a esse hiperplano.

2.3 Métricas de Avaliação

Nesta seção descrevemos algumas das métricas de avaliação utilizadas em nossos experimentos para medir a efetividade de nossas abordagens, sendo que essas métricas são comumente empregadas em problemas de Aprendizado de Máquina e Recuperação da Informação (Ribeiro-Neto e Baeza-Yates, 1999). São elas: acurácia, precisão, revocação, e *F-measure* (F1). Acurácia mede a soma total de acertos em todas as categorias. Precisão é a razão dos elementos classificados corretamente na categoria dividido pelo número de elementos classificados nessa categoria, enquanto a revocação é o número de conferências corretamente classificadas na categoria dividido pelo número de elementos da categoria. F1 sumariza ambas as métricas precisão e revocação. Em nossos experimentos utilizamos os valores médios da precisão, revocação e F1 para todas as categorias (Macro-Precisão, Macro-Revocação e Macro-F1).

A Figura 2.2 apresenta uma matriz de confusão (Kohavi e Provost, 1998) usada para definir essas medidas utilizando apenas duas categorias (A e B) para simplificar. Nessa matriz a é o número de conferências classificadas corretamente na categoria A, b é o número de conferências classificadas erroneamente na categoria B, c é o número de conferências classificadas erroneamente na categoria A e d é o número de conferências classificadas corretamente na categoria B. No caso de F1, X representa uma das categorias (A ou B).

		Predição	
		A	B
Rótulo	A	a	b
Verdadeiro	B	c	d

$$\text{Acurácia} = \frac{a+d}{a+b+c+d}$$

$$\text{Revocação}_A = \frac{a}{a+b} \quad — \quad \text{Revocação}_B = \frac{d}{c+d}$$

$$\text{Precisão}_A = \frac{a}{a+c} \quad — \quad \text{Precisão}_B = \frac{d}{b+d}$$

$$F1_X = \frac{2 \times (\text{Precisão}_X \times \text{Revocação}_X)}{\text{Precisão}_X + \text{Revocação}_X}$$

Figura 2.2: Métricas de Avaliação.

Capítulo 3

Conjuntos de Dados e Características das Conferências

Na Seção 3.1 descrevemos os conjuntos de dados utilizados para a realização de nossos experimentos apresentados nos Capítulos 4 e 5. Na Seção 3.2 descrevemos as características que foram utilizadas como indicadores de qualidade das conferências. Essas características foram definidas com base na análise dos dados das fontes descritas nas Seções 3.1.2 e 3.1.3.

3.1 Conjunto de dados

Nesta seção são descritos os três conjuntos de dados utilizados neste trabalho. O primeira é um *ranking*[1] de conferências em Ciência da Computação produzido a partir de uma enquete eletrônica sobre a qualidade de uma lista preestabelecida de conferências, da qual participaram diversos especialistas em Ciência da Computação. Esse *ranking* é parte de um projeto desenvolvido na Universidade Federal de Minas Gerais em conjunto com outras universidades brasileiras denominado Perfil-CC[2], e que tinha como objetivo levantar o perfil da produção científica dos melhores programas de pós-graduação em Ciência da Computação do país (Laender et al., 2008). Esse *ranking* foi utilizado em nossos experimentos como um gabarito para medir a efetividade de nossas duas abordagens. O segundo conjunto de dados é uma coleção de metadados de citações bibliográficas coletada automaticamente da Libra Academic Search[3] (Nie et al., 2007). Finalmente, o terceiro conjunto de dados descreve aspectos específicos (por exemplo, número de edições, lista dos membros do comitê de programa, taxa de aceitação, etc.) referentes a um subconjunto da lista de conferências do Perfil-CC e coletados manualmente da Web.

[1] Optamos por utilizar a palavra *ranking* ao invés de alguma possível tradução para o português por sua utilização ser largamente difundida e expressar melhor o significado em diversas ocasiões ao longo do texto.

[2] http://www.latin.dcc.ufmg.br/perfilccranking

[3] http://libra.msra.cn

3.1.1 Perfil-CC Ranking

Aqui explicaremos a metodologia utilizada para construir o *ranking* de conferências produzido pelo projeto Perfil-CC, denominado *Perfil-CC Ranking*. Inicialmente, foram obtidas e combinadas listas de conferências de diversas fontes. Depois, foi executado um processo de limpeza nessa lista para remover conferências duplicadas e conferências com as seguintes características: (*i*) conferências ainda não consolidadas (menos de quatro edições), (*ii*) conferências com submissão por meio de resumo, ou seja, o artigo não é avaliado na integra, e (*iii*) conferências regionais ou restritas especificamente a um país, principalmente conferências que não são substancialmente importantes para a comunidade brasileira por serem realizadas em países distantes, por exemplo, as pacífico-asiáticas. Para facilitar o processo de votação, as conferências que permaneceram após o processo de limpeza foram divididas em 27 grupos, que representam uma possível divisão do campo de Ciência da Computação (Laender et al., 2008). Os grupos que compõem essa divisão estão listados no Apêndice A.

Para validar e aumentar a qualidade e cobertura da lista inicial de conferências, foi convidado um grande número de especialistas, incluindo todos os pesquisadores em Ciência da Computação bolsistas em produtividade do CNPq (Conselho Nacional de Desenvolvimento Científico e Tecnológico)[4] e todos os docentes dos programas de pós-graduação em Ciência da Computação do país. Todos os participantes puderam sugerir adições e remoções de conferências da lista, assim como mudanças de grupos. No final desse processo chegou-se a uma lista com pouco mais de mil conferências.

Na próxima etapa, foi conduzida uma enquete eletrônica sobre a qualidade de cada conferência na lista. Nessa enquete, cada especialista poderia votar apenas uma vez em cada grupo, porém era permitido votar em todos os grupos. Cada conferência poderia ser classificada em três categorias (A, B ou C) de acordo com a sua qualidade, sendo a categoria A a de maior qualidade. Para evitar excessos na avaliação das conferências, como, por exemplo, todas as conferências de um determinado grupo serem classificadas como A, foi estabelecido que, em cada grupo, não mais de 40% das conferências poderiam ser classificadas como A e a soma de conferências A e B não poderia exceder 80%. Além disso, existiam duas outras categorias: NA ("não avaliada"), quando um pesquisador escolhesse não votar em uma determinada conferência, por não se considerar apto a julgar a qualidade dessa conferência (por falta de conhecimento sobre a conferência, por exemplo), e NC ("não considerar"), para desqualificar uma conferência cuja qualidade fosse considerada muito baixa para estar listada. Essa etapa envolveu 312 participantes, dos quais 147 eram pesquisadores bolsistas de produtividade do CNPq, totalizando 875 votos em todos os grupos, o que mostra que em média cada pesquisador votou em aproximadamente três grupos.

A última etapa foi tornar esses votos em um número que traduzisse a qualidade da conferência. Diversos cenários, considerando diferentes critérios para contar e ponderar os votos, foram considerados. As diferenças entre esses cenários não eram significantes e aquele que foi considerado o melhor por alguns especialistas consultados foi o escolhido. Deste modo, os

[4]Esses pesquisadores recebem essa bolsa baseado na qualidade de sua produção científica e são considerados como os mais importantes pesquisadores ou líderes em seus respectivos campos de pesquisa.

pesos de 3, 2 e 1 foram atribuídos para os votos A, B e C, respectivamente, enquanto que os votos em NA e NC receberam peso zero. Para normalizar a pontuação obtida por uma conferência em um número entre 0 e 1, a pontuação foi divida por três vezes o número de votos válidos recebidos pela conferência. Foram considerados votos válidos, todos os votos diferentes de NA e NC. Assim, os votos NA e NC não foram contados e não influenciaram o resultado final.

Conferências que tiveram um número de votos válidos menor que 40% foram removidas do resultado final. Essa porcentagem foi escolhida experimentalmente para evitar casos de conferências bem classificadas, mas com poucos votos válidos. Depois dessa etapa, a lista final de conferências avaliadas chegou ao número exato de mil conferências, que foram ordenadas de acordo com a pontuação obtida, assim originando o *Perfil-CC Ranking*. O projeto Perfil-CC também prevê uma divisão das conferências em três níveis de qualidade, A (Alta Qualidade), B (Média Qualidade) e C (Baixa Qualidade). Assim, as 40% primeiras conferências do *Perfil-CC Ranking* foram rotuladas como A, as próximas 40% foram rotuladas como B e as conferências restantes foram rotuladas como C. O *ranking* final gerado, assim como a divisão das conferências nas três categorias, pode ser acessado em `http://www.latin.dcc.ufmg.br/perfilccranking`.

O grande número de conferências encontradas por esse projeto reforça a importância da existência de um processo automático para avaliação da qualidade de conferências, visto que a execução de um processo como esse é muito custoso. Todo o processo, incluindo a coleta de dados, o processamento (padrozinação e limpeza), a enquete e a apuração dos resultados, durou cerca de seis meses, contando com a ajuda de vários voluntários. Outro ponto interessante foi o grande número de conferências criadas recentemente encontradas na listagem inicial. Como essas conferências não fazem parte da lista final, futuramente diversas delas tornar-se-ão conferências consolidadas e, portanto, a lista produzida precisará ser atualizada, sendo esse um fator que dificulta muito a utilização dessa abordagem para classificação de conferências.

3.1.1.1 Perfil-CC Ranking versus CORE Ranking

Para validar o Perfil-CC *ranking* como nosso gabarito, o comparamos com o *CORE Ranking of ICT Conferences*[5]. Esse *ranking* foi gerado como resultado de um processo similar ao usado no projeto Perfil-CC, i.e., a CORE, *The COmputing REsearch Association of Australasia*, conduziu um processo para construir um *ranking* através de uma série de encontros entre pesquisadores australianos, envolvendo aproximadamente 1500 conferências de Ciência da Computação. Existe uma grande interseção entre as listas de conferências dos dois projetos (CORE e Perfil-CC), as principais exceções sendo as conferências regionais e locais (e.g., conferências pacífico-asiáticas) que estão presentes na lista do CORE. Uma importante diferença entre os dois projetos é o número de categorias e a distribuição das conferências em cada categoria. O CORE divide as conferências em cinco categorias: A+ ou altíssima qualidade,

[5] http://www.core.edu.au

A ou alta qualidade, B ou média qualidade, C ou baixa qualidade e L ou conferências regionais de baixa qualidade. A distribuição das conferências na lista do CORE é apresentada na Tabela 3.1.

Category:	A+	A	B	C	L
Size(%):	6	27	31	29	6

Tabela 3.1: Porcentagem de conferências em cada categoria do CORE *ranking*.

Para comparar os *rankings* gerados pelos projetos Perfil-CC e CORE, consideramos uma amostra de 194 conferências extraídas da lista do Perfil-CC, que é utilizado em nossos experimentos (uma explicação sobre a escolha dessas conferências pode ser encontrada na Seção 3.1.2). Das 194 conferências, 133 são comuns a ambas listas. Mapeamos as conferências das categorias A+ e A para uma única categoria rotulada A, contendo apenas conferências de alta qualidade, e desconsideramos a categoria L, já que nenhum conferência regional está presente na lista do Perfil-CC. Como resultado, ambas as listas têm três categorias depois dessas modificações. A Figura 3.1 mostra a matriz de confusão contrastando as categorias do Perfil-CC contra as categorias do CORE. Como podemos ver, 79 (59,4%) das 133 conferências são classificadas na mesma categoria nos dois *rankings*. Esse valor pode ser considerado alto, se considerarmos as diferentes distribuições em cada categoria e os fatores culturais das comunidades consultadas que pode influenciar a opinião sobre algumas conferências. Note, por exemplo, que o número de conferências classificadas erroneamente nos extremos, i.e., A como C e C como A, é muito pequeno. Além disso, acreditamos que os *rankings* produzidos são consistentes e refletem uma classificação precisa da qualidade de um número considerável de conferências.

		CORE		
		A	B	C
Perfil-CC	A	55	16	4
	B	19	15	7
	C	3	5	9

Figura 3.1: Matriz de confusão contrastando a categoria das conferências do Perfil-CC contra a categoria da CORE.

3.1.2 Metadados da Libra

A *Libra Academic Search* ou simplesmente Libra (Nie et al., 2007) é uma biblioteca digital focada no campo da Ciência da Computação. A Libra permite busca de dados bibliográficos gratuitamente, ajudando usuários a encontrar artigos científicos de interesse. Atualmente possui mais de três milhões de documentos organizados de acordo com o veículo de publicação. Os veículos são divididos em 23 grupos, similares aos grupos existentes no Perfil-CC.

A escolha da Libra como fonte dos metadados de citações ocorreu devido a duas razões principais. A primeira razão é o volume de informação existente na Libra. O número de

conferências é considerável e os artigos dessas conferências possuem uma grande quantidade de metadados, incluindo citações bibliográficas que são uma informação muito importante para o nosso trabalho. A DBLP[6] (Ley, 2002), uma das maiores bibliotecas digitais de Ciência da Computação, poderia ser uma boa alternativa, visto que possui uma grande cobertura de conferências e a qualidade dos seus metadados é muito alta; entretanto poucas entradas da DBLP possuem os dados de citações. A segunda razão para escolha da Libra se deve à sua cobertura em relação à lista de conferências do Perfil-CC, que para alguns grupos é maior que 80%, como pode ser visto na Tabela 3.2. CiteSeer[7] (Giles et al., 1998), uma terceira opção, foi a primeira biblioteca digital no campo de Ciência da Computação a indexar e conectar citações automaticamente. Contudo, o CiteSeer não provê uma organização dos artigos por veículo de publicação, o que torna difícil um processo de coleta automática focada nas conferências de interesse, assim como determinar a cobertura de conferências em relação à lista de conferências do Perfil-CC. Além disso, o número de documentos existentes no CiteSeer, pouco mais de 760 mil, é muito menor do que na Libra.

Grupo	**Cobertura da Libra**(%)
Aprendizado de Máquina	84,2
Bancos de Dados, Recuperação de Informação, Bibliotecas Digitais, Mineração de Dados	88,9
Biologia Computacional	83,3
Interação Humano Computador, Sistemas Colaborativos	88,9
Redes de Computadores, Sistemas Distribuídos, Sistemas P2P	68,3
Web, Sistemas Multimídia e Hipermídia	79,1

Tabela 3.2: Cobertura da Libra em relação às conferências de cada grupo do Perfil-CC.

Para executar nossos experimentos, coletamos da Libra a informação sobre artigos e citações das conferências de seis grupos da lista do Perfil-CC, denominados: Aprendizado de Máquina (AM); Bancos de Dados, Recuperação de Informação, Bibliotecas Digitais, Mineração de Dados (BD); Biologia Computacional (BIO); Interação Humano Computador, Sistemas Colaborativos (IHC); Redes de Computadores, Sistemas Distribuídos, Sistemas P2P (REDES); Web, Sistemas Multimídia e Hipermídia (WEB). Esses grupos foram escolhidos priorizando aqueles que possuíam maior cobertura em relação à lista de conferências do Perfil-CC. A Tabela 3.2 mostra a cobertura para cada um desses grupos.

Os resultados dessa coleta está sumarizado na Tabela 3.3. A interseção entre as listas de conferências do Perfil-CC e da Libra para esses seis grupos totalizam 194 conferências. A lista das 194 conferências coletadas pode ser encontrada no Apêndice B. Uma análise baseada em cada grupo também foi executada. A Figura 3.2 mostra a porcentagem de artigos e citações para cada grupo. O grupo BD possui o maior número de artigos e citações. A diferença para outros grupo é significativa e pode favorecer conferências desse grupo em alguns dos experimentos realizados.

[6] http://dblp.uni-trier.de
[7] http://citeseer.ist.psu.edu

Conferências Coletadas	194
Artigos Coletados	109.969
Total de citações (apenas internas)	145.282
Artigos que possuem os dados de referências (apenas internas)	27.759
Artigos que recebem pelo menos uma citação	43.749

Tabela 3.3: Descrição dos dados coletados da Libra. Citações para artigos que não estão no nosso conjunto de dados foram removidas.

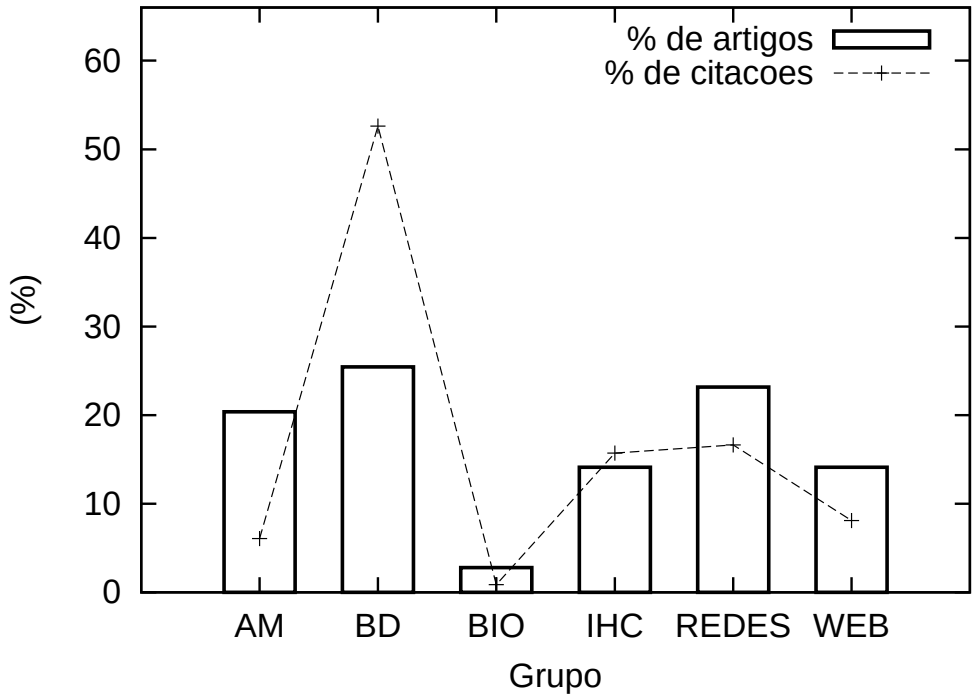

Figura 3.2: Porcentagem de artigos e citações para cada grupo.

3.1.3 Dados Adicionais

Além das citações e dados bibliográficos coletados da Libra, coletamos manualmente de bibliotecas digitais, sítios de conferências e outros repositórios encontrados na Web, um grande volume de dados sobre as conferências pertencentes aos seis grupos do Perfil-CC listados na Tabela 3.2. Esses dados compreendem, para cada conferência, a periodicidade (anual, bianual, trianual), sua designação (workshop, simpósio, conferência, etc.) e o número de edições realizadas até 2007, bem como taxas de submissão e aceitação e a lista de membros do comitê de programa das edições realizadas no triênio 2005-2007. Essa coleta forneceu uma grande quantidade de dados que, juntamente com as citações bibliográficas, possibilitou-nos explorar diversos grupos de características e analisar seu potencial para classificar conferências de acordo com as três categorias do Perfil-CC. Essas características são apresentadas detalhadamente na próxima seção.

3.2 Estudo das Características das Conferências

A partir dos conjuntos de dados descritos anteriormente, exploramos um total de 21 características relativas às conferências. Essas características podem ser utilizadas como indicadores da qualidade das conferências e foram divididas em quatro grupos, de acordo com o aspecto a que elas estão relacionadas: contagem de citações, taxas de submissão e aceitação, tradição e características do comitê de programa. Outras características mais gerais, como a periodicidade, foram indiretamente utilizadas. Esta seção analisa e discute esses diferentes grupos de características e seu potencial para classificar conferências em três categorias, rotuladas A, B e C (do maior nível de qualidade para o menor). A distribuição das 194 conferências de nossa amostra nas categorias A, B e C de acordo com a divisão provida pelo projeto Perfil-CC é, respectivamente, de: 88 (45,36%), 71 (36,60%) e 35 (18,04%). Também apresentamos os gráficos de distribuição acumulada que ilustram o poder de discriminação de algumas dessas características em termos dos dados coletados das 194 conferências consideradas em nosso estudo.

3.2.1 Características baseadas em Citações

Citações bibliográficas constituem a característica bibliométrica mais utilizada para medir a qualidade de veículos de publicação, sendo que existem diversas métricas baseadas em citações propostas na literatura. A partir da rede de citações entre conferências, exploramos um grande subconjunto dessas métricas, incluindo as métricas mais populares para classificação de periódicos, *H-index* (Braun et al., 2006), Contagem de Citações (CC), Número Médio de Citações (AVGCC), *Weighted PageRank* (WPR) (Bollen et al., 2006), *Y-Factor* (Bollen et al., 2006), e Fator de Impacto (FI) (Amin e Mabe, 2000) (discutiremos algumas dessas métricas mais profundamente na Seção 4.1). Além disso, utilizamos algumas métricas projetadas especificamente para conferências, *Conference Impact Factor* (CIF), *Conference Citation Impact* (CCI), *Conference Combined Factor* (CCF) e *Conference Factor* (C-Factor), sendo que essas métricas são uma contribuição desta dissertação e serão apresentadas detalhadamente na Seção 4.2.

Cada uma dessas métricas (para periódicos e conferências) busca promover diferentes características desejáveis de veículos de alta qualidade, por exemplo, CC mede a popularidade do veículo, WPR promove veículos que recebem citações de veículos de alta qualidade valorizando o prestígio do veículo e FI mede a popularidade atual do veículo.

Para demonstrar como as métricas baseadas em citações se comportam, a Figura 3.3 apresenta a distribuição acumulada do número total de citações (Contagem de Citações - CC) por categoria. Nas Figuras 3.3 a 3.9, dado um valor a no eixo x, as curvas mostram a fração de conferências que têm algum valor menor ou igual a a. As exceções são as Figuras 3.3 e 3.5; a primeira está em escala logarítmica (log_{10}) e a última possui uma interpretação oposta: $P(X > x)$. Como podemos ver, as conferências da categoria A tendem a receber mais citações que as conferências da categoria B, que por sua vez são mais citadas que as conferências da categoria C. Além disso, as curvas são bastante distintas, sem interseções. Assim, podemos

concluir que CC é um bom discriminador para as três categorias de conferências.

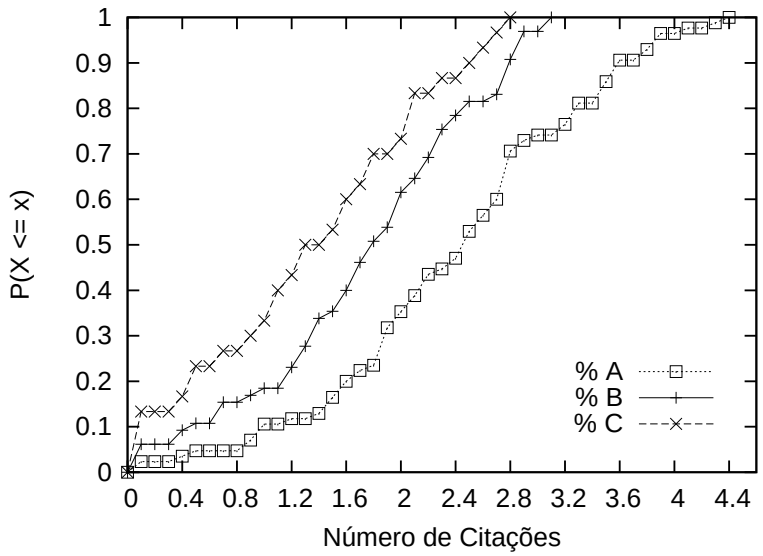

Figura 3.3: Distribuição acumulada do número de citações. O número de citações está em escala logarítmica (log_{10}).

Apesar de estudarmos diversas métricas baseadas em citações, optamos por exibir apenas a distribuição acumulada de CC pela simplicidade e fácil entendimento, além de ser direta e indiretamente utilizada para definir as outras métricas baseadas em citações.

Embora as métricas baseadas em citações nos permitam medir diferentes e importantes aspectos sobre a qualidade das conferências, essas métricas possuem algumas limitações: problemas como citações induzidas e auto-citações, falta de citações, entre outros (MacRoberts e MacRoberts, 1989) ainda não estão resolvidos e podem ter grande impacto nessas métricas.

3.2.2 Características baseadas nas Taxas de Submissão e Aceitação

O número de trabalhos submetidos para uma conferência é um indicativo de sua popularidade e do seu porte. Conferências de alta qualidade normalmente são de grande porte e atraem a atenção de um grande número de pesquisadores que se interessam em publicar ou apenas participar dessas conferências. Contudo, essas conferências comumente possuem taxas de aceitação muito baixas. A taxa de aceitação reflete a dificuldade de se ter um artigo aceito para apresentação na conferência. Em nosso estudo, utilizamos as taxas de submissão e aceitação média no triênio 2005-2007, como características para avaliar a qualidade da conferência.

A Figura 3.4 mostra a distribuição acumulada da taxa de submissão. Como pode ser observado, conferências das categorias B ou C possuem valores mais baixos para a taxa de submissão, enquanto as conferências da categoria A possuem valores mais altos, como esperado. Entretanto, as curvas de distribuição acumulada das conferências da categoria B e C são muito similares e possuem interseções em alguns pontos, impedindo uma distinção clara entre as conferências das duas categorias. Além disso, a distinção entre essas duas categorias

pode ser afetada pela quantidade de dados disponíveis. Em nosso conjunto de dados, para 11,3% das conferências classificadas como B e 40,0% das conferências classificadas como C não foi possível encontrar a taxa de submissão, contra apenas 4,5% das conferências classificadas como A. Contudo, note que a curva da categoria B mostra que existe uma pequena quantidade de conferências com taxas de submissão muito alta, inclusive maiores que as taxas da categoria A. Entretanto, como essa porcentagem é muito pequena não influenciam nossa análise. Deste modo, a taxa de submissão é uma boa caraterística para distinguir conferências da categoria A, porém não é uma boa caraterística para distinguir entre conferências B e C.

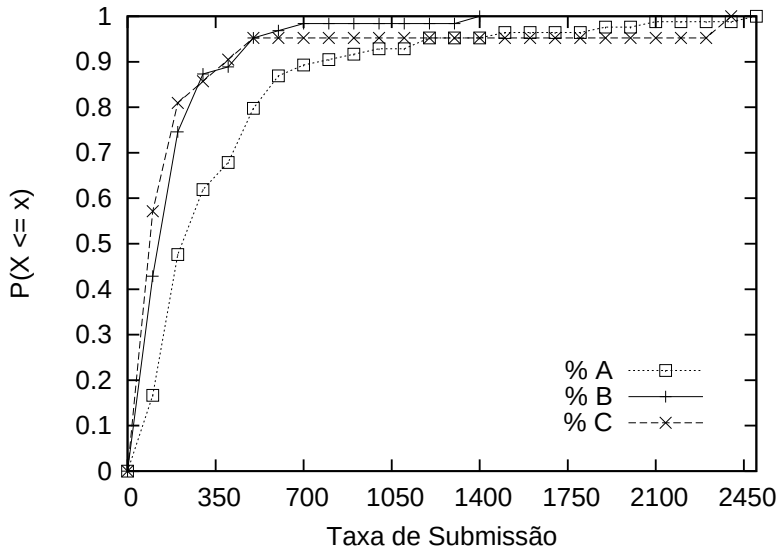

Figura 3.4: Distribuição acumulada da taxa de submissão.

A Figura 3.5 mostra a distribuição acumulada para a taxa de aceitação. Note que a taxa de aceitação das conferências da categoria A normalmente são menores que a taxa de aceitação das conferências da categoria B, que por sua vez possuem valores menores para a taxa de aceitação que as conferências da categoria C, sendo que apenas 16,6% das conferências da categoria A possuem taxa de aceitação maior que 40%, enquanto que nas categorias B e C são aproximadamente 35% e 48%, respectivamente. Ademais, praticamente todas as conferências da categoria C (92%) possuem taxa de aceitação maior que 30%. Essa figura mostra que a taxa de aceitação pode confundir os classificadores em algumas situações, como as conferências das categorias B e C com baixa taxa de aceitação. Todavia, a taxa de aceitação pode ser considerada uma boa característica para avaliação da qualidade de conferências.

3.2.3 Características baseadas na Tradição

O número de edições tem uma relação direta com a tradição da conferência. Conferências mais antigas e tradicionais são, normalmente, conferências de maior qualidade. Entretanto, existem conferências novas de alta de qualidade. Assim, embora o número de edições claramente reflita a tradição de uma conferência, isoladamente pode não ser uma boa característica para avaliar a sua qualidade.

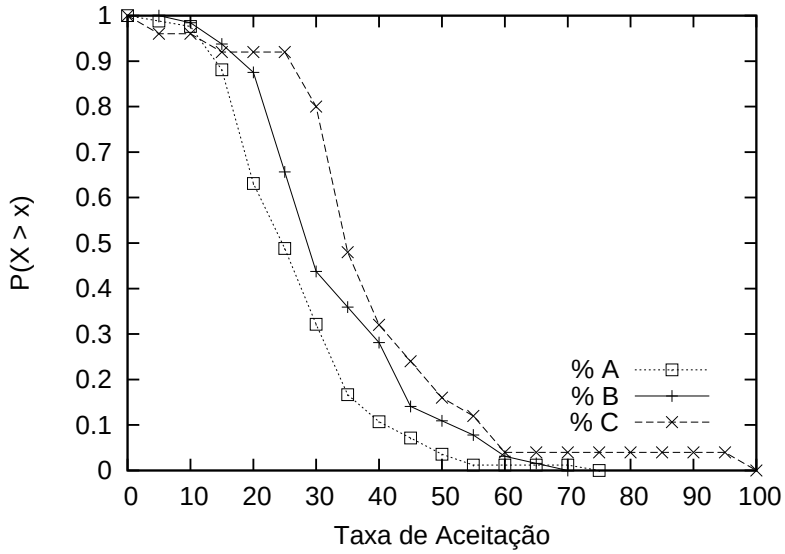

Figura 3.5: Distribuição acumulada da taxa de aceitação.

A Figura 3.6 mostra a distribuição acumulada do número de edições por categoria. Nesse caso, conferências da categoria A tendem a possuir um grande número de edições, enquanto conferências B e C tendem a possuir um número menor de edições. Cerca de 40% das conferências da categoria A possuem mais de 16 edições. Para as categorias B e C, apenas 10,00% e 16,13% das conferências alcançam esse número, respectivamente. Também podemos notar que não existe uma grande distância entre as curvas B e C, indicando que o número de edições para as conferências dessas categorias é bem próximo.

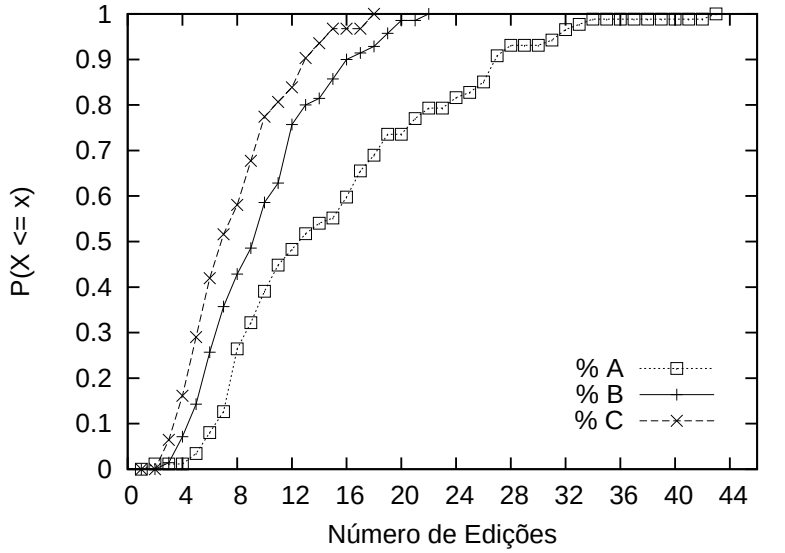

Figura 3.6: Distribuição acumulada do número de edições.

3.2.4 Características baseadas no Comitê de Programa

Diversas características baseadas no comitê de programa podem ser utilizadas para medir a qualidade de conferências, assumindo que a qualidade de uma conferência está diretamente relacionada com a reputação do comitê de programa. Em (Zhuang et al., 2007), por exemplo, os autores avaliam a qualidade de conferências considerando o tamanho do comitê de programa, o número médio de artigos publicados pelos membros do comitê de programa e o número médio de coautores desses membros. Também são utilizadas duas medidas de centralidade de redes, *closeness* e *betweenness centrality* (Wasserman e Faust, 1994). Essas duas últimas medidas são calculadas utilizando a rede de coautoria. Note que nenhuma das características citadas acima é baseada nas citações recebidas pelos membros do comitê de programa, sendo essa uma fonte de informação comumente usada quando pesquisadores são avaliados.

Nosso trabalho estende o conjunto de características proposto em (Zhuang et al., 2007) e ainda considera o número de citações, o número médio de citações por artigo e o H-index (Hirsch, 2005) de cada membro do comitê de programa. Note que todas as características baseadas no comitê de programa, exceto o tamanho do comitê, estão diretamente relacionadas com a produção científica de seus membros.

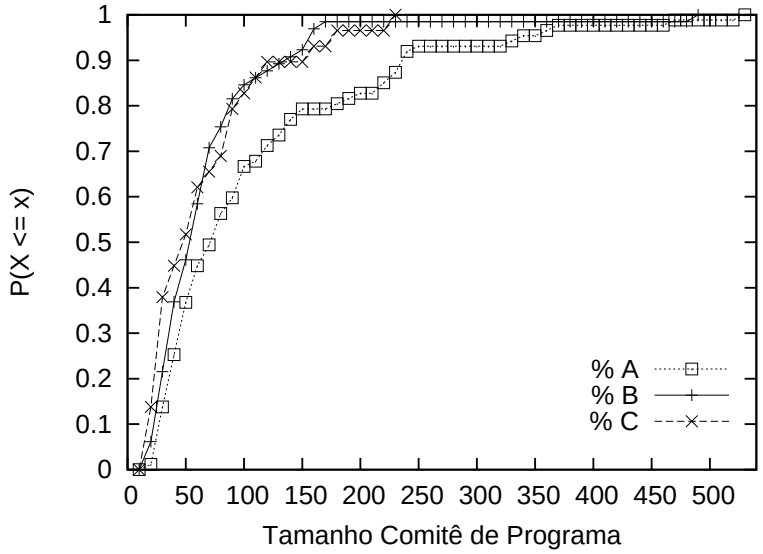

Figura 3.7: Distribuição acumulada do tamanho do comitê de programa.

A Figura 3.7 mostra a distribuição acumulada do tamanho do comitê de programa. A curva que representa a categoria A indica que essas conferências possuem comitês de programas maiores que as conferências das categorias B e C. Novamente, as curvas das categorias B e C são muito similares, mostrando que essa característica pode não ser efetiva para distinguir conferências dessas duas categorias. O número médio de artigos publicados pelos membros do comitê de programa também é uma característica fraca para classificação de conferências, como mostrado na Figura 3.8. Note a interseção entre as três curvas.

Existe uma grande similaridade entre o número médio de citações recebidas pelos membros do comitê de programa das conferências classificadas como A e B, como pode ser observado

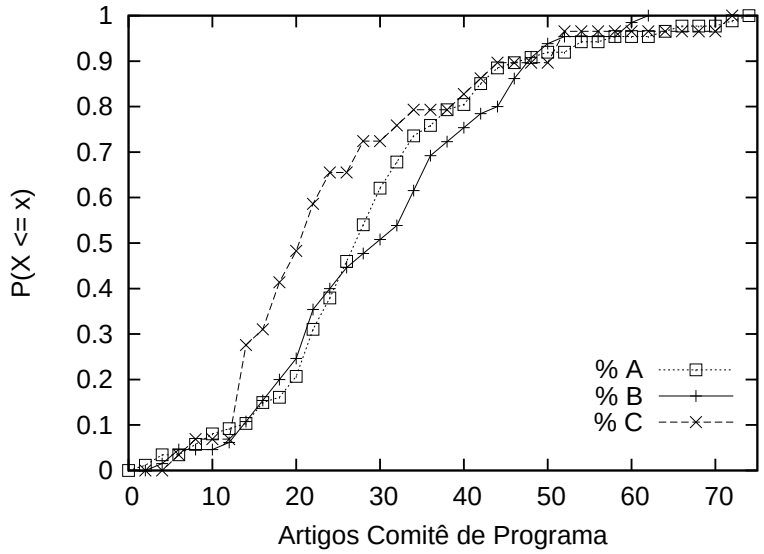

Figura 3.8: Distribuição acumulada do número médio de artigos publicados pelos membros do comitê de programa.

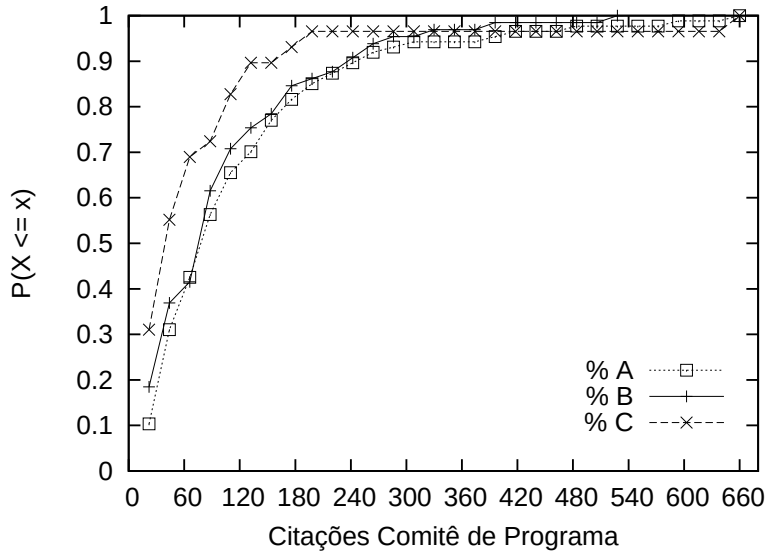

Figura 3.9: Distribuição acumulada do número de citações recebidas pelos membros do comitê de programa

na Figura 3.9. Por outro lado, a curva das conferências da categoria C indica que os membros do comitê de programa dessas conferências são menos citados.

3.2.5 Outras Características

Além das características apresentadas nas seções anteriores, usamos em nosso estudo três outras características que não podem ser consideradas diretamente como indicadores de qualidade, mas acreditamos que podem auxiliar no processo de classificação. A primeira característica é a forma como o evento é designado. Utilizamos essa informação porque, normalmente,

ela indica o porte do evento. De modo geral, *workshops* são de menor porte que simpósios e simpósios são de menor porte que conferências. Entretanto, podemos facilmente encontrar exceções a essa regra. A segunda característica é a periodicidade da conferência. Pretendemos analisar se conferências anuais, bianuais e trianuais apresentam diferentes padrões para as características estudadas. Por último, utilizamos a informação sobre o grupo temático a que a conferência pertence, i.e., AM, BD, BIO, IHC, REDES, e WEB, conforme considerado pelo projeto Perfil-CC (veja a Tabela 3.2). Essa informação é importante porque diferentes grupos podem ter diferentes padrões de citação e critérios de autoria, assim como possuir um número diferente de conferências e pesquisadores.

Capítulo 4

Abordagem Baseada em Citações Bibliográficas

Neste capítulo propomos uma abordagem baseada em citações bibliográficas para classificar automaticamente conferências de acordo com a sua qualidade. Apresentamos um conjunto de métricas para avaliação de periódicos e analisamos a sua aplicabilidade para classificação de conferências. As deficiências e os problemas encontrados nessas métricas motivaram-nos a propor um conjunto de novas métricas especialmente projetadas para capturar aspectos intrínsecos e importantes relacionados às conferências, como longevidade, popularidade, prestígio e periodicidade. Realizamos uma série de experimentos para medir a eficácia dessas métricas, obtendo ganhos de até 8,4% sobre os melhores resultados das métricas para periódicos e de até 58% sobre o Fator de Impacto, métrica mais tradicional para periódicos.

4.1 Métricas Baseadas em Citações para Avaliação de Periódicos

Nesta seção, apresentaremos algumas das métricas mais populares, baseadas em citações, usadas para estimar a qualidade de periódicos. Na discussão seguinte, analisamos as características de cada uma dessas métricas e mostramos que nenhuma delas captura todos os aspectos necessários para uma boa métrica para avaliação de conferências.

4.1.1 Contagem de Citações

A Contagem de Citações (CC) é a métrica mais simples utilizada para medir a qualidade de periódicos, sendo também utilizada para estimar o impacto de artigos, assim como de pesquisadores ou de grupo de pesquisadores. A CC de um periódico X (CC_X) é calculada contando o número de citações recebidas por todos os artigos publicados em X, isto é,

$$CC_X = |I_X|,$$

onde I_X é o conjunto de artigos que citam artigos publicados em X. Apesar de ser largamente utilizada, podemos mencionar dois pontos negativos dessa métrica. O primeiro é que as citações vindas de um periódico de baixa qualidade possuem o mesmo valor das citações vindas de periódicos de alta qualidade, ou seja, a CC não pondera as citações de acordo com a qualidade do periódico do qual vem a citação. O segundo ponto negativo é que periódicos mais antigos podem ser favorecidos quando comparados com periódicos novos, simplesmente porque possuem artigos publicados há mais tempo e que, portanto, podem ter recebido mais citações ao longo do tempo do que os artigos publicados em periódicos mais novos.

4.1.2 Fator de Impacto

O Fator de Impacto (FI) é a métrica proposta pelo *Institute for Scientific Information*[1]. FI mede a qualidade de um periódico com base no número de citações recebidas em um período de um ano pelos artigos publicados nos dois anos anteriores (a janela temporal final é de três anos, dois anos de artigos publicados mais um ano contabilizando as citações para esses artigos). FI é a métrica mais popular para medir a qualidade de periódicos, sendo largamente aplicada. Um estudo profundo do FI pode ser encontrado em (Amin e Mabe, 2000). Entretanto, não conseguimos encontrar qualquer estudo sobre a aplicação de FI para avaliação de conferências.

Mais especificamente, o FI de um periódico X em um dado ano Y é calculado contando o número de citações recebidas dos artigos publicados no ano Y pelos artigos publicados em X nos dois anos anteriores a Y. Esse número é dividido pelo número de artigos publicados em X nos dois anos anteriores a Y.

Como CC, FI não considera o prestígio ou a qualidade de um periódico do qual se origina a citação. A principal diferença da métrica FI quando comparada com CC, é que CC promove periódicos mais antigos que receberam um grande número de citações ao longo do tempo e FI mede a popularidade atual dos artigos publicados em um periódico, de modo que periódicos mais novo não são penalizados.

4.1.3 Weighted PageRank

Uma variação do algoritmo PageRank (Brin e Page, 1998) adaptada para análise de citações é proposta em (Bollen et al., 2006). O algoritmo proposto, denominado de Weighted PageRank (WPR), expressa o prestígio de um determinado periódico considerando o prestígio dos artigos que citam esse periódico. Arestas são formadas quando algum artigo de um periódico cita outro e são atribuídos pesos para essas arestas que determinam a quantidade de prestígio transferida. Esse peso é calculado de acordo com o número de vezes que um periódico cita outro. A definição formal é dada pelas fórmulas:

$$WPR_X = (1 - d) + d \sum_{\forall Y \in J_X} (WPR_Y \times w_{Y,X})$$

e

[1] Agora denominado *Thomson Scientific*.

$$w_{Y,X} = \frac{W(Y,X)}{\sum_{\forall Z} W(Y,Z)},$$

onde $W(Y,X)$ e $W(Y,Z)$ representam o número de citações do periódico Y para o periódico X e o número de citações do periódico Y para o periódico Z, respectivamente, J_X é o conjunto de periódicos que citam X e d é uma constante que pode variar entre 0 e 1, sendo responsável por atenuar a quantidade de prestígio transferido de um periódico para outro. Em nossos experimentos exibidos na Seção 4.3, usamos o valor de 0,85, que é o mesmo empregado no trabalho original (Bollen et al., 2006).

4.1.4 Y-Factor

A métrica Y-Factor, também proposta em (Bollen et al., 2006), combina as métricas FI e WPR. O Y-Factor de um periódico X é dado por:

$$Y\text{-}Factor_X = WPR_X \times FI_X.$$

Como FI envolve a contagem de citações que um periódico recebeu nos últimos dois anos, essa é uma métrica que expressa a popularidade de um periódico. WPR, por outro lado, é uma métrica que expressa o prestígio de um periódico. Assim, Y-Factor é uma métrica que combina o prestígio e a popularidade de um periódico, ou seja, um periódico com alto valor de Y-Factor deve ser popular e ter um alto prestígio.

4.1.5 H-index

Uma métrica que captura a qualidade da produção científica de um indivíduo, conhecida como H-index, é proposta em (Hirsch, 2005). Com o sucesso do H-index, extensões para avaliação de periódicos foram propostas (Braun et al., 2006). De maneira análoga à definição original, um periódico tem H-index h se existem pelo menos h artigos publicados nesse periódico que tenham recebido pelo menos h citações. O H-index captura o número de artigos publicados em um periódico com mais impacto ao longo do tempo, deste modo periódicos novos são prejudicados por essa métrica. Alguns trabalhos (por exemplo, (Sidiropoulos et al., 2007), (Katsaros et al., 2007)) propõem variações da definição acima para incluir aspectos temporais.

4.1.6 Análise das Métricas para Avaliação de Periódicos

Para analisar as maiores deficiências das métricas de periódicos para avaliação da qualidade de conferências, primeiro consideramos os aspectos intrínsecos das conferências e discutimos como algumas dessas características podem ser consideradas por métricas específicas para avaliação de conferências:

- *Longevidade versus Popularidade Atual.* A longevidade de uma conferência traz importantes informações sobre a qualidade da conferência. Uma conferência mais antiga e tradicional normalmente possuirá um número maior de citações absolutas, tendo, por

exemplo, alto valor de CC. Entretanto, uma conferência pode perder a importância com o tempo ou mesmo deixar de existir. Outro ponto negativo de se considerar apenas a longevidade da conferência é a existência de conferências novas de alta qualidade. Assim, uma boa métrica também deve considerar a importância e popularidade atual de uma conferência.

- *Prestígio versus Popularidade.* Algumas métricas para periódicos capturam apenas a popularidade, valorizando assim periódicos que possuem um grande número de citações. Embora a popularidade seja uma informação valiosa, métricas que consideram o prestígio, ou seja, valorizam citações vindas de veículos de mais alta qualidade, podem também capturar informações importantes sobre a qualidade dos veículos avaliados. Assim, uma métrica que combine essas duas características provavelmente irá produzir melhores resultados.

- *Tamanho da Conferência.* O tamanho de uma conferência pode ser considerado um possível indicador de qualidade, porque conferências de alta qualidade usualmente atraem um grande número de submissões e participantes. Deste modo, essa informação deve ser capturada por uma boa métrica para avaliação de conferências. Certamente, existem conferências de grande porte de baixa qualidade e conferências de pequeno porte de alta qualidade, como SIGCOMM e SIGMETRICS[2]. Assim o tamanho de uma conferência não deve ser usado como único critério de avaliação.

- *Periodicidade.* Conferências podem ser anuais, bianuais ou trianuais. Uma métrica que não considera uma janela temporal compatível com essa periodicidade pode prejudicar uma conferência ou um grupo particular de conferências.

- *Cobertura e Esparsidade dos Dados.* Ao contrário do que ocorre com os periódicos, para os quais existem coleções de metadados muito organizadas, a quantidade de conferências e as edições de uma conferência cobertas pelas bibliotecas digitais são limitadas. Isso significa que os dados de uma determinada conferência podem ter sido originados de diversas fontes. Assim, muito freqüentemente, uma conferência pode estar indexada em uma biblioteca digital, mas sem que todas as suas edições ou ocorrências estejam cobertas. Isso leva a situações nas quais os dados de citações disponíveis para uma conferência não cobrem uma janela temporal contínua. Similar ao que ocorre com a periodicidade, uma métrica que não considera uma janela temporal que seja suficientemente flexível para amenizar os problemas de cobertura e esparsidade dos dados, devido à ausência de algumas edições, pode prejudicar determinadas conferências.

A Tabela 4.1 mostra quais características são consideradas por cada uma das métricas de periódicos descritas anteriormente. Como podemos ver nenhuma das métricas considera todas as características. Embora alguém possa argumentar que essas características podem

[2]SIGCOMM - ACM SIGCOMM Conference (http://www.sigcomm.org) e SIGMETRICS - ACM - SIGMETRICS Conference on Measurement and Modelling of Computing Systems (http://www.sigmetrics.org) são duas das conferências mais importantes em seus respectivos grupos.

ser consideradas para classificação de periódicos, é evidente que essas características são fundamentais para prover uma classificação mais precisa sobre a qualidade de conferências. Por exemplo, uma métrica como FI, que considera uma janela temporal curta e fixa, pode prejudicar conferências cujos dados de citações não estejam disponíveis nesse intervalo de tempo específico. E também, nenhuma das métricas de periódicos considera diretamente o tamanho da conferência como uma característica importante para avaliação. Todas essas observações indicam que novas métricas específicas para avaliação da qualidade de conferências são necessárias.

	CC	FI	WPR	Y-Factor	H-index
Longevidade	Sim	Não	Sim	Sim	Sim
Popularidade Atual	Não	Sim	Não	Sim	Não
Prestígio	Não	Não	Sim	Sim	Não
Tamanho da Conferência	Não	Não	Não	Não	Não
Periodicidade	Sim	Não	Sim	Sim	Sim
Cobertura e Esparsidade dos Dados	Sim	Não	Sim	Sim	Sim

Tabela 4.1: Característica coberta por cada métrica (Sim: significa que a métrica cobre a característica; Não: caso contrário).

4.2 Novas Métricas Propostas

Nesta seção propomos um conjunto de novas métricas baseadas em citações bibliográficas específicas para medir a qualidade de conferências e que foram desenvolvidas para suprir as deficiências das métricas para periódicos encontradas na análise da seção anterior (Martins et al., 2008, 2009b).

4.2.1 Conference Impact Factor

O FI é a métrica mais popular para medir a qualidade de periódicos, mas sua utilização para classificação de conferências deve ser investigada mais profundamente devido a dois importantes aspectos que não são considerados por essa métrica: a periodicidade da conferência e a cobertura e esparsidade dos dados. Utilizando a métrica FI como foi proposta originalmente podemos prejudicar conferências que não são cobertas pela biblioteca digital na janela temporal específica usada por FI, que considera as citações dos artigos publicados em um determinado ano para os artigos publicados nos dois anos anteriores. Por exemplo, para uma conferência bianual ou trianual, apenas uma edição poderia influenciar no cálculo de FI, em termos de citações recebidas e realizadas. Como discutimos anteriormente, os dados de algumas edições de uma conferência podem não estar cobertos pela biblioteca digital. Assim, mesmo conferências anuais podem estar sujeitas aos mesmos problemas das conferências bianuais ou trianuais. Para abordar essas questões, propomos uma redefinição de FI, denominada *Conference Impact Factor* (CIF), que emprega uma janela temporal maior aumentando a probabilidade de se obter dados para as conferências inicialmente classificadas.

O CIF de uma conferência X em um dado ano Y é calculado contando o número de citações feitas pelos artigos das conferências publicados entre os anos Y e Y-3 para os artigos publicados em X nos três anos anteriores a Y-3, sendo esse valor dividido pelo número total de artigos publicados em X nos três anos anteriores a Y-3. Assim, aumentamos a janela temporal para um período de seis anos. Deste modo, mesmo conferências trianuais terão duas edições nesse período. Com uma janela temporal maior, a probabilidade de alguma conferência ser prejudicada é menor, mesmo existindo problemas de cobertura. Por outro lado, considerando uma janela temporal maior, CIF perde um pouco da "recenticidade".

Deve ser salientado que CIF é apenas uma das novas métricas propostas e que apesar de ter por base algumas das idéias das métricas de periódicos apresentadas anteriormente, ele é uma métrica mais bem adaptada para o contexto das conferências. Como discutido na Seção 4.1.6, existem vários outros aspectos intrínsecos das conferências que devem ser capturados. Isso é feito em cada uma das novas métricas propostas, discutidas a seguir.

4.2.2 Conference Citation Impact

CIF é a razão entre o número total de citações e o número total de artigos publicados em uma conferência em um certo período de tempo. Utilizando uma idéia similar à aplicada a CIF, exploramos uma janela temporal de seis anos e propomos o *Conference Citation Impact* (CCI). CCI mede a razão entre o número total de citações recebidas por uma conferência em um período de tempo e o total de citações recebidas por todas as conferências nesse mesmo período de tempo. CCI tem as mesmas vantagens e desvantagens da métrica CIF, mas tende a promover conferências que possuem um número maior de artigos. Apesar de utilizar a mesma idéia da métrica CIF, isto é, uma janela temporal maior, as modificações propostas para CCI, que considera a proporção entre o número de citações recebidas pela conferência e a quantidade total de todas as citações, produz uma métrica mais adequada para conferências e melhora significativamente os resultados quando comparados com a métrica CIF, como será mostrado na Seção 4.3.

Mais especificadamente, o CCI de uma conferência X em um dado ano Y é calculado contando o número de citações feitas pelos artigos das conferências publicados entre os anos Y e Y-3 para os artigos publicados em X nos três anos anteriores a Y-3, sendo esse valor dividido pelo número total de citações feitas por todos os artigos publicados entre os anos Y e Y-3 para todos os artigos, de todas as conferências, publicados nos três anos anteriores a Y-3. O valor obtido é multiplicado pelo número de conferências (NC) para obter um valor com a mesma magnitude da métrica CIF. Essa multiplicação não afeta a avaliação gerada pela métrica CCI, entretanto permite combinar essa com outras métricas com a mesma importância ou peso para o resultado final.

4.2.3 Combined Conference Factor

Duas questões importantes que não são explicitamente tratadas pelas métricas CIF ou CCI são a longevidade e o tamanho das conferências. Para tentar capturar explicitamente o ta-

manho de uma conferência, propomos uma métrica chamada de *Conference Size* (CS), que estima o tamanho de uma conferência X analisando o número de artigos publicados em X, visto que normalmente apenas as conferências mais importantes e de grande porte possuem infra-estrutura para receber e publicar uma grande quantidade de artigos. Novamente, é importante salientar que existem vários exemplos de conferências de grande porte de baixa qualidade e essa métrica não deve ser utilizada como único critério para avaliação de conferências. Entretanto, na prática, conferências de baixa qualidade tendem a ser de menor porte, enquanto conferências de alta qualidade tendem a ser de maior porte. Essa tendência é confirmada em nossos experimentos apresentados na Seção 4.3. Como CS utiliza o número de artigos para estimar o tamanho da conferência e CCI tende a favorecer conferências com um número maior de artigos, qual a necessidade da métrica CS? Como já dissemos, CCI não captura explicitamente o tamanho da conferência, já que podem existir conferências com poucos artigos, mas com muitas citações e assim possuir um alto valor dessa métrica. Portanto, conferências de menor porte (com poucos artigos) podem ser bem avaliadas pela métrica CCI.

Definimos o CS de uma conferência X como a quantidade de artigos publicados em X em todos os anos (NA_X) dividido pelo número total de artigos publicados em todas as conferências em todos os anos (TNP). O resultado é multiplicado por NC, para que esse valor tenha a mesma magnitude do CIF. Formalmente, o CS de uma conferência X é definido como:

$$CS_X = \frac{NC \times NA_X}{TNP}.$$

Para medir a importância de uma conferência ao longo do tempo, ou seja, a longevidade da conferência, propomos outra métrica simples denominada *Conference Longevity* (CL). O CL de uma conferência X consiste da CC de X dividido pelo número total de citações recebidas por todas as conferências em todos os anos (TNC). Esse resultado é multiplicado por NC, como é feito para CS, também para que o valor obtido tenha a mesma ordem de magnitude da métrica CIF. Formalmente, o CL de uma conferência X é definido como:

$$CL_X = \frac{NC \times CC_X}{TNC}.$$

Combinando CIF e CCI, e as duas métricas descritas acima, propomos uma nova métrica, denominada *Combined Conference Factor* (CCF), definida como:

$$CCF_X = CIF_X + CCI_X + CS_X + CL_X.$$

CCF é uma métrica nova e original que cobre a maior parte das características intrínsecas das conferências, as quais não são explicitamente cobertas por nenhuma das métricas tradicionais para periódicos. Essa métrica promove conferências tradicionais, de grande porte e populares (longevidade e popularidade atual), ou seja, conferências que possuem valores altos para os indicadores de qualidade. É importante ressaltar que uma conferência não precisa cobrir necessariamente todos os critérios capturados pela métrica CCF para ser classificada como uma conferência de alta qualidade. É fácil notar que, por exemplo, uma conferência de alta qualidade e que seja de pequeno porte, tradicional e popular (longevidade e popularidade

atual) muito possivelmente será classificada como de alta qualidade pela métrica CCF. CCF cobre a maioria das questões discutidas na Seção 4.1.6, sendo o prestígio da conferência o único aspecto que não é explicitamente coberto por essa métrica. Isso motivou-nos a propor a próxima métrica, o *Conference Factor*.

4.2.4 Conference Factor

Similar ao que foi feito para o Y-Factor, o *Conference Factor* (C-Factor) é uma métrica que combina prestígio e popularidade. Entretanto, enquanto o Y-Factor é calculado multiplicando WPR por FI, o C-Factor substitui FI por CCF, pois acreditamos que CCF é uma métrica mais adequada para avaliação de conferências. O C-Factor de uma conferência X é definido como:

$$C\text{-}Factor_X = WPR_X \times CCF_X.$$

C-Factor é a única métrica que atende todas características desejáveis para uma boa métrica para avaliação de conferências. C-Factor e CCF, que cobrem praticamente todas as características, são as duas métricas mais completas para avaliação de conferências, e como será apresentado na próxima seção, são as métricas mais adequadas para avaliação de conferências entre todas as métricas consideradas.

4.2.5 Análise das Novas Métricas

De forma a resumir a discussão sobre as novas métricas a Tabela 4.2 mostra quais características discutidas anteriormente são consideradas por cada uma das novas métricas. Como já dissemos, o CCF considera praticamente todas as características fundamentais para avaliar a qualidade de conferências, apenas o prestígio não é considerado por essa métrica. O C-Factor é a única métrica que considera todas as características. Note que todas as novas métricas consideram o problema da periodicidade e da cobertura e esparsidade dos dados.

	CCF	CCI	CCF	C-Factor
Longevidade	Não	Não	Sim	Sim
Popularidade Atual	Sim	Sim	Sim	Sim
Prestígio	Não	Não	Não	Sim
Tamanho da Conferência	Não	Não	Sim	Sim
Periodicidade	Sim	Sim	Sim	Sim
Cobertura e Esparsidade dos Dados	Sim	Sim	Sim	Sim

Tabela 4.2: Característica coberta por cada métrica (Sim: significa que a métrica cobre a característica; Não: caso contrário).

4.3 Experimentos

Nesta seção apresentaremos dois conjuntos de experimentos realizados para avaliar a efetividade das novas métricas propostas para classificação de conferências. No primeiro ex-

perimento, todas as métricas apresentadas foram utilizadas para criar um *ranking* das 194 conferências que foram coletadas da Libra e os resultados comparados com o gabarito provido pelo *Perfil-CC Ranking*. No segundo experimento, utilizamos novamente todas as métricas apresentadas para classificar essas conferências em categorias de acordo com a sua qualidade, também conforme o *Perfil-CC Ranking*. Nos dois conjuntos de experimentos, nossas métricas obtiveram, em geral, resultados melhores do que as métricas existentes, obtendo ganhos de até 8,4% na comparação de similaridade dos *rankings* e 7,8% na acurácia para a classificação em categorias, em relação as métricas existentes.

4.3.1 Comparação dos *Rankings*

Para cada uma das métricas apresentadas foi gerado um *ranking*, o qual foi comparado com o *Perfil-CC Ranking* utilizando algumas métricas de avaliação. A primeira dessas métricas, denominada Top 30, é baseada na comparação dos 30 elementos do topo de cada *ranking* gerado, com os 30 elementos do topo do *Perfil-CC Ranking*, isto é, a quantidade de conferências em comum. Contudo, comparar apenas os elementos das 30 primeiras posições mostra apenas a eficácia de uma métrica para as conferências de maior qualidade. Outro modo de comparar a similaridade entre dois *rankings* é calcular a distância entre a posição de cada elemento correspondente nos *rankings*. A Distância Simples, segunda métrica de avaliação utilizada, é definida como:

$$D(m_1, m_2) = \frac{1}{N} \sum_{\forall i} \frac{|P_{m1}(i) - P_{m2}(i)|}{N},$$

onde N é o número de elementos no *ranking*, $P_{m1}(i)$ e $P_{m2}(i)$ são os elementos na $i^{\underline{a}}$ posição no *ranking* construído com as métricas m_1 e m_2, respectivamente. A função $D(m_1, m_2)$ produz como saída um número entre 0 (quando os *rankings* são iguais) e 0,5 (quando os *rankings* estão em ordem inversa).

Quando dois *rankings* são comparados, diferenças nas posições mais altas normalmente são mais importantes que nas posições mais baixas, ou seja, é mais importante melhorar o *ranking* gerado por uma métrica de forma que as melhores conferências sejam colocadas nas posições de topo. A Distância Simples é uma métrica que não captura esse fator, visto que as distâncias entre todos os elementos possuem o mesmo peso. Por essa razão, utilizamos outra métrica de avaliação, a Distância Balanceada (*Weighted Distance*), definida em (Sidiropoulos e Manolopoulos, 2006) como:

$$D_w(m_1, m_2) = \frac{1}{N \sum_{\forall i \in V} w(m_1, m_2, i)} \sum_{\forall i \in V} d_w(m_1, m_2, i),$$

$$d_w(m_1, m_2, i) = |P_{m1}(i) - P_{m2}(i)| \times w(m_1, m_2, i),$$

$$w(m_1, m_2, i) = \frac{1}{min(P_{m1}(i), P_{m2}(i))}.$$

Finalmente, utilizamos uma métrica proposta pela comunidade de recuperação da informação para computar a similaridade entre dois *rankings*, o coeficiente Tau de Kendall (τ) (Kendall, 1938), definido como:

$$\tau = \frac{2P}{\frac{1}{2}N(N-1)},$$

onde N é o número de elementos nos *rankings* e P é o número de pares que estão na mesma ordem nos dois *rankings* subtraído do número de elementos que estão fora de ordem, ou seja, para cada par de elementos analisado, se o par aparece no segundo *ranking* na mesma ordem do primeiro, adicionamos 1 a P, e se o par não aparece na mesma ordem, subtraímos 1. O valor de τ é +1 quando os *rankings* são iguais e -1 quando estão exatamente na ordem inversa.

Os resultados da comparação dos *rankings* gerados por todas as métricas e usando todas as métricas de avaliação consideradas podem ser vistos na Tabela 4.3. Para a Distância Simples e a Distância Balanceada, menores valores indicam maior similaridade e, portanto, melhores resultados. Como pode ser visto, CC obteve o melhor resultado entre as métricas existentes. C-Factor obteve o *ranking* mais similar ao *Perfil-CC Ranking* entre todas as métricas avaliadas, exceto para Top 30, com um ganho, considerando o melhor resultado das métricas existentes, de 7,5% na Distância Simples, 5,0% na Distância Balanceada e 8,4% no coeficiente τ. CCF foi a métrica que obteve o melhor resultado para Top 30. Contra a segunda melhor métrica utilizada, WPR, nossos ganhos chegam a cerca de 14% para o coeficiente τ, por exemplo. Se considerarmos a métrica mais tradicional e largamente utilizada, FI, nossos ganhos chegam a mais de 58% (para o coeficiente τ). Esses resultados, quando comparados com as métricas existentes, mostram a importância de se considerar os aspectos intrínsecos das conferências na proposta de C-Factor e CCF.

	Top 30	D	D_w	τ
CC	19	0,2109	0,1191	0,3840
FI	13	0,2490	0,3233	0,2626
WPR	18	0,2159	0,1224	0,3652
Y-Factor	16	0,2359	0,1298	0,3053
H-index	18	0,2167	0,1257	0,3640
CIF	12	0,2352	0,2586	0,2973
CCI	19	0,2077	0,1210	0,3718
CCF	**21**	0,2031	0,1166	0,3915
C-Factor	19	**0,1951**	**0,1132**	**0,4161**

Tabela 4.3: Resultados da comparação dos *rankings* usando Top 30, Distância Simples, Weighted Distance e o coeficiente Tau de Kendall. Os melhores resultados estão em negrito.

Analisando a Tabela 4.3 mais detalhadamente, podemos ver que CIF produz um *ranking* mais similar ao *Perfil-CC Ranking* do que FI. Esse era um resultado esperado, já que ao aumentarmos a janela temporal, diminuímos a probabilidade de que algumas conferências sejam prejudicadas por falta de dados. CC e H-index, que exploram a longevidade da conferência, sofrem menos com os problemas de cobertura e periodicidade e obtiveram bons resultados.

Uma das razões que justificam os bons resultados obtidos pelas métricas que exploram a longevidade é que são raros os casos que uma conferência perde importância ao longo dos anos. A tendência é que as conferências se consolidem com o tempo e, conseqüentemente, sua importância aumente. Os bons resultados de WPR mostram a importância do prestígio das conferências. Esse fato também pode ser observado na pequena melhora obtida pela métrica C-Factor em relação a CCF e pela superioridade dos resultados do Y-Factor quando comparados com FI.

Olhando mais profundamente cada *ranking* gerado, encontramos problemas de cobertura que podem explicar alguns dos resultados. Das 194 conferências consideradas, 14 (7,22%) não possuem nenhuma citação. Para FI, que considera uma janela temporal de apenas três anos, o número de conferências que não possuem citações é de 74 (38,14%); utilizando uma janela temporal de seis anos, como proposta para CIF, esse número diminui para 47 (24,23%). Conseqüentemente, essas conferências, sem nenhuma citação, serão classificadas nas últimas posições por praticamente todas as métricas. Essa análise mostra a importância de se considerar a longevidade da conferência e é uma justificativa para os bons resultados das métricas que consideram esse aspecto. As métricas que incorporam CS e, portanto, consideram o número de artigos publicados pela conferência, podem classificar melhor conferências sem nenhuma citação. A ausência dos dados de citações irá prejudicar a classificação dessas conferências, porém elas não serão obrigatoriamente classificadas nas últimas posições, visto que existe algum tipo de informação para avaliá-las. Essa análise também é válida para conferências que possuem uma cobertura ruim dos dados de citações.

Para se ter uma melhor idéia sobre os *rankings* gerados por nossas métricas, a Tabela 4.4 mostra as conferências nas dez primeiras posições (Top 10) no *Perfil-CC Ranking* e nos *rankings* gerados pelas métricas FI, CCF e C-Factor. Como podemos ver, os *rankings* gerados por CCF e C-Factor são muito similares nas primeiras dez posições, com nove conferências em comum. Apesar de existirem pequenas diferenças nesses *rankings* em relação ao *Perfil-CC Ranking*, todas as conferências que aparecem nas dez primeiras posições são definitivamente conferências de alta qualidade. Essas duas métricas favorecem conferências do grupo BD e incluem nas dez primeiras posições seis e cinco conferências desse grupo, respectivamente. Isso ocorre devido ao grupo BD possuir a melhor cobertura na Libra. No entanto, o conjunto de conferências nas dez primeiras posições do *ranking* gerado por FI é muito diferente daquele do Perfil-CC Ranking, incluindo três conferências classificadas acima da posição 50 nesse *ranking* (EDBT 56ª, NSDI 100ª, e CIDR 183ª). Isso mostra que as novas métricas são mais consistentes e promovem grandes melhorias em relação a FI, métrica mais tradicional para periódicos.

Para demonstrar a importância da métrica CS e, conseqüentemente, do tamanho da conferência como um indicador de qualidade, exibimos na Tabela 4.5 as dez primeiras (Top 10) e as dez últimas (Bottom 10) conferências classificadas por CS. Como podemos ver, o conjunto Top 10 inclui duas conferências classificadas acima da 100ª posição no *Perfil-CC Ranking*. Por outro lado, CHI, INFOCOM, ICDE, SIGMOD e VLDB são excelentes exemplos de conferências de grande porte e de alta qualidade. Além disso, no conjunto Bottom 10 existem

	Perfil-CC Ranking	**FI**		**CCF**		**C-Factor**	
Posição	Conf	Conf	Pos	Conf	Pos	Conf	Pos
1	INFOCOM	CIDR	183	SIGMOD	9	SIGMOD	9
2	SIGCOMM	SIGCOMM	2	VLDB	12	VLDB	12
3	CSCW	SIGMOD	9	CHI	4	CHI	4
4	CHI	VLDB	12	SIGCOMM	2	SIGCOMM	2
5	KDD	PODS	18	INFOCOM	1	INFOCOM	1
6	WWW	NSDI	100	WWW	6	ICDE	13
7	ICML	IMC	31	ICDE	13	SIGIR	32
8	ACM-MM	KDD	5	PODS	18	PODS	18
9	SIGMOD	EDBT	56	SIGIR	32	ICML	7
10	ICC	UIST	17	KDD	5	WWW	6

Tabela 4.4: Conferências nas dez primeiras posições (Top 10) do *Perfil-CC Ranking* e dos *rankings* gerados pelas métricas FI, CCF e C-Factor, e suas respectivas posições no *Perfil-CC Ranking*.

apenas conferências classificadas entre as posições 100 e 193 no *Perfil-CC Ranking*, o que comprova que conferências de menor qualidade tendem a ser de menor porte. Isso indica que o tamanho da conferência é um valioso indicador de qualidade, mas que, entretanto, não pode ser utilizado isoladamente.

Top 10			**Bottom 10**		
Rank	Conf	Pos	Rank	Conf	Pos
1	CHI	4	185	SAPIR	193
2	INFOCOM	1	186	DIWeb	178
3	ICPR	52	187	SWWS	151
4	NIPS	24	188	NSDI	100
5	ICDE	13	189	W4A	174
6	SIGMOD	9	190	EuroITV	142
7	VLDB	12	191	PDC	175
8	WebNet	176	192	WCW	188
9	ICME	23	193	GIR	167
10	KES	110	194	Web3D	158

Tabela 4.5: As dez primeiras (Top 10) e as dez últimas (Bottom 10) conferências no *Perfil-CC Ranking* e suas respectivas posições de acordo com a métrica CS.

Um resultado interessante que exigiu uma análise mais profunda foi o conjunto Top 30. As conferências classificadas nas 30 primeiras posições pelo *Perfil-CC Ranking* são, provavelmente, conferências muito populares e importantes, o que implicaria em uma maior cobertura pela Libra e, conseqüentemente, uma melhor classificação por nossas métricas. Contudo, o melhor resultado no conjunto Top 30 foi obtido pela métrica CCF, com 21 conferências classificadas corretamente nesse conjunto, ou seja, o melhor resultado classificou nove conferências fora do Top 30. Uma análise dessas conferências mostrou que sete conferências do topo não foram classificadas nas 30 primeiras posições por nenhuma das métricas. Investigamos duas possibilidades: (*i*) a Libra não possui cobertura suficiente para essas sete conferências, impe-

dindo a classificação dessas conferências entre as Top 30 e (*ii*) os dados de citações não são suficientes para classificar corretamente essas conferências e mais informação é necessária. Essas sete conferências são: *International Conference on Fuzzy Systems (FUZZ)*, *International Conference on Communications (ICC)*, *International Joint Conference on Neural Networks (IJCNN)*, *European Conference on Machine Learning (ECML)*, *International Conference on Web Services (ICSW)*, *IEEE/WIC/ACM International Conference on Web Intelligence (WI)*, e *ACM/IFIP/USENIX International Middleware Conference (MIDDLEWARE)*.

Das sete conferências que não foram classificadas no conjunto Top 30 por nenhuma das métricas, cinco (FUZZ, ICC, IJCNN, ICSW, WI) têm sérios problemas de cobertura. O problema de cobertura ocorre por duas razões. Primeiro, a Libra não possui uma quantidade de dados suficiente sobre a conferência e, segundo, o subconjunto coletado não é grande e abrangente o suficiente. Por exemplo, conferências relacionadas a Aprendizado de Máquina podem ser afetadas pela ausência das conferências de Inteligência Artificial, visto que existe, possivelmente, uma forte conexão entre esses dois grupos de conferências. As duas conferências restantes, ECML e MIDDLEWARE, possuem uma quantidade de informação considerável na Libra, mas ainda assim não foram incluídas no conjunto Top 30 por nenhuma das métricas. Nesses dois casos, informações adicionais (diferentes das citações) são necessárias, provavelmente, para se obter uma classificação correta.

Para entender como cada métrica se comporta quando classifica erroneamente uma conferência fora do conjunto Top 30, calculamos a Distância Simples entre a posição estabelecida pelas métricas e a posição 30, para determinar quantas posições acima da posição 30 as conferências foram classificadas. Analisando os resultados mostrados na Tabela 4.6, podemos ver que em média Y-Factor e C-Factor foram as métricas que tiveram melhor predição para as conferências classificadas erroneamente, classificando-as aproximadamente 48 posições acima da posição 30, ou seja, na 78ª posição. Desse resultado podemos concluir que essas conferências, que são de alta qualidade, estão sendo classificadas muitas posições acima de sua posição de acordo com o Perfil-CC Ranking. Portanto, isso indica que falta informação para que as métricas classifiquem corretamente as conferências entre as 30 primeiras. Essa falta de informação pode ser devida a problemas de cobertura ou pela dependência exclusiva dos dados de citações em nossas métricas.

	CC	FI	WPR	Y-Factor	H-index	CIF	CCI	CCF	C-Factor
NEr	11	17	12	14	12	18	11	9	11
DMTop30	59,09	68,41	49,33	47,43	65,75	76,61	60,45	52,78	47,45

Tabela 4.6: Resultados da análise do conjunto Top 30. NEr é o número de conferências classificadas erroneamente fora do conjunto Top 30 e DMTop30 é a Distância Simples média da posição 30 para essas conferências.

4.3.2 Classificação das Conferências de Acordo com a sua Qualidade

O projeto Perfil-CC, além de prover um *ranking* para mil conferências, também provê uma classificação em três categorias (A, B e C) de acordo com a qualidade das conferências, como

explicado anteriormente. Comparamos o desempenho das métricas estudadas (métricas existentes e novas) na tarefa de classificar as 194 conferências presentes em nosso conjunto de dados, que forma nossa amostra, nessas três categorias. Para isso, usamos a posição que a conferência foi classificada por cada métrica e a distribuição das conferências em cada categoria. A distribuição das conferências nas categorias A, B e C em nossa amostra é, respectivamente, de: 88 (45,36%), 71 (36,60%) e 35 (18,04%). Assim, conferências classificadas por cada métrica até a 88ª posição foram consideradas da categoria A, conferências entre as posições 89 e 160 foram consideradas da categoria B e conferências acima da 160ª posição foram consideradas da categoria C. Para comparar o desempenho das métricas nessa tarefa, empregamos as medidas acurácia, precisão e revocação, que foram apresentadas na Seção 2.3.

Em nossos experimentos, os valores de revocação não são apresentados porque são iguais aos de precisão para todas as métricas. Isso era esperado por causa da metodologia empregada para transformar a posição da conferência no *ranking* gerado por cada métrica em uma das três categorias (A, B ou C), que segue a distribuição apresentada anteriormente. Para simplificar essa explicação consideraremos apenas duas categorias (A e B). Como o número de conferências classificadas em cada categoria é fixo e uma conferência classificada erroneamente será classificada necessariamente na outra categoria, o número de conferências da categoria A classificadas na categoria B é igual ao número de conferências da categoria B classificadas na categoria A. Essa relação é justamente medida pela precisão e revocação, e como são iguais, a precisão é igual à revocação.

Acurácia, precisão e revocação são métricas de avaliação que medem o desempenho das métricas apenas para as conferências classificadas na categoria correta, não considerando a "gravidade" do erro. Por exemplo, classificar uma conferência A erroneamente na categoria B por uma métrica representa um erro menos grave do que classificar essa mesma conferência na categoria C. Para comparar as métricas e verificar quais delas provêem classificações com um número menor de erros graves, propomos a métrica de avaliação "Escore de Categoria". Três diferentes cenários foram propostos com diferentes pontuações. Como os resultados em todos os três cenários foram muito similares, apresentaremos apenas um deles. O Escore de Categoria é calculado do seguinte modo: para cada conferência classificada na categoria correta soma-se dois pontos, para cada conferência A ou C classificada como B soma-se um ponto, para cada conferência B classificada como A ou C soma-se um ponto e para conferências A classificadas como C ou C classificadas como A não são somados pontos.

A Tabela 4.7 sumariza os resultados para todas as métricas. Como anteriormente, o pior resultado foi obtido pela métrica FI. CCF e C-Factor foram, novamente, as métricas com melhor desempenho. CCF classificou 57,22% das conferências na categoria correta, apresentado um ganho de 7,8% sobre WPR, melhor resultado das métricas existentes. C-Factor, por outro lado, obteve os maiores valores de precisão/revocação e Escore de Categoria, apresentando ganhos de 5,8% na acurácia, 5,9% na precisão e 3,4% para o Escore de Categoria, sobre os melhores resultados das métricas existentes.

Os resultados da métrica CIF são muito superiores aos de FI, portanto FI é uma métrica que não deve ser utilizada para avaliação de conferências da maneira como foi definida original-

	Acurácia(%)	Precisão(%)	Escore de Categoria
CC	51,03	47,77	278
FI	44,33	39,70	258
WPR	53,09	49,65	278
Y-Factor	47,94	42,63	266
H-index	47,94	45,22	272
CIF	51,03	46,23	274
CCI	53,09	47,93	278
CCF	**57,22**	52,47	286
C-Factor	56,19	**52,58**	**288**

Tabela 4.7: Resultados da classificação em categorias. Os melhores resultados estão em negrito.

mente. Os resultados de CCI foram significativamente melhores quando comparados com os de CIF usando a mesma janela temporal, provavelmente devido a essa métrica promover conferências com um grande número de artigos publicados, estimando indiretamente o tamanho da conferência. H-index obteve resultados intermediários e uma análise mais completa sobre essa métrica precisa ser realizada para extrair alguma característica que possa ser incorporada em outras métricas desenvolvidas para conferências. CC, embora sendo uma métrica muito simples, obteve novamente bons resultados, reforçando a importância da longevidade das conferências. Das métricas existentes, WPR, que captura o prestígio da conferência, foi a métrica que obteve os melhores resultados. Y-Factor obteve resultados ruins, muito provavelmente devido à utilização de FI na sua composição. CCF e C-Factor foram as melhores métricas para classificação de conferências, sendo que essas métricas foram projetadas especialmente para cobrir aspectos específicos relacionados à avaliação da qualidade de conferências.

A Figura 4.1 mostra a matriz de confusão gerada por C-Factor, a métrica que obteve o melhor valor para o Escore de Categoria. Como podemos ver, sete conferências da categoria A foram classificadas na categoria C e oito conferências da categoria C foram classificadas na categoria A. O Escore de Categoria é a única métrica de avaliação que penaliza esses casos. Todas as sete conferências da categoria A classificadas como da categoria C possuem menos de dez citações, sendo que três delas são conferências novas e para as quais ainda não existe muita informação. As outras quatro conferências são conferências consolidadas que possivelmente sofrem de problemas de cobertura. Das oito conferências da categoria C classificadas como da categoria A, três delas são conferências do grupo BD e uma delas do grupo IHC, os dois grupos com melhor cobertura em relação ao número de artigos e citações em nossa amostra coletada da Libra, o que pode explicar o fato dessas conferências terem sido bem classificadas. Entre as outras quatro conferências, duas possuem um grande número de artigos e a outra possui um grande número de citações, o que pode, também, explicar a sua classificação. A última conferência da categoria C classificada como da categoria A tem um número médio de citações, mas todas essas são citações muito recentes (alto valor de popularidade atual), o que gera um alto valor para CIF e CCI, duas métricas que influenciam o C-Factor.

A Tabela 4.8 mostra a acurácia das métricas em cada grupo de conferências. Em negrito

		Predição		
		A	B	C
Rótulo	A	**67,0%**	25,0%	8,0%
Verdadeiro	B	29,6%	**50,7%**	19,7%
	C	22,9%	37,1%	**40,0%**

Figura 4.1: Matriz de confusão gerada pela métrica C-Factor.

estão os melhores resultados para cada grupo. Como podemos ver, nenhuma das métricas foi melhor em mais de dois grupos. Uma análise profunda sobre as características de cada grupo é necessária para entender melhor esse fenômeno e para permitir que possamos sugerir qual métrica deve ser usada em cada situação/grupo. Na ausência desse estudo, sugerimos a utilização de CCF e C-Factor por serem as melhores métricas no geral. Contudo, as novas métricas novamente conseguiram os melhores resultados, obtendo o melhor valor de acurácia em todos os grupos, exceto para o grupo BD. Nesse grupo, que possui a maior cobertura de artigos e citações em nossa amostra, WPR obteve os melhores resultados, seguido por C-Factor. Nesse caso, a rede de citações entre as conferências é muito densa e o prestígio tornou-se uma característica muito importante. Para os grupos BIO e WEB, nos quais a maioria das conferências foi criada há poucos anos (aproximadamente sete anos em média), o melhor resultado foi obtido por CCI, métrica que captura a popularidade atual da conferência.

	ML	DB	BIO	IHC	NT	WEB
CC	37,50	46,30	40,00	68,23	54,55	45,45
FI	40,63	40,00	30,00	66,67	38,18	51,52
WPR	40,63	**60,00**	50,00	75,00	54,55	39,39
Y-Factor	43,75	45,00	40,00	70,83	36,51	54,55
H-index	37,50	47,50	40,00	62,50	49,09	48,48
CIF	46,87	45,00	50,00	62,50	50,90	54,55
CCI	50,00	47,50	**60,00**	66,67	47,27	**60,61**
CCF	**56,25**	52,50	40,00	75,00	56,36	57,58
C-Factor	40,63	55,00	40,00	**83,33**	**58,18**	54,55

Tabela 4.8: Acurácia por grupo (melhores resultados em negrito).

Capítulo 5

Abordagem Baseada em Técnicas de Aprendizado de Máquina

Na Seção 3.2 apresentamos um conjunto de características que podem ser utilizadas como indicadores da qualidade de conferências. Neste capítulo vamos utilizar e combinar essas características para classificar as conferências em níveis de qualidade preestabelecidos pelo *Perfil-CC Ranking* através de técnicas de aprendizado de máquina. Com base nos nossos resultados experimentais (Martins et al., 2009a), (1) avaliamos o potencial dos grupos de características em discriminar conferências em categorias, (2) identificamos os principais desafios que tornam o problema de classificação automática de conferências difícil e (3) estudamos alternativas para amenizar o efeito desses problemas.

Nossa abordagem para avaliar automaticamente a qualidade de conferências combina diversos tipos de característica buscando obter uma classificação mais precisa. Contudo, existe um número muito grande de possíveis combinações, o que torna a busca por uma solução ótima impraticável mesmo automaticamente. Assim, para tratar esse problema, empregamos vários métodos de aprendizado de máquina supervisionados. Esses métodos realizam buscas heurísticas no espaço de soluções, o que permite encontrar eficientemente soluções próximas à ótima (Mitchell, 1997). Conforme já mencionado, em nossos experimentos consideramos os seguintes métodos de aprenzidado de máquina: árvores de decisão, métodos Bayesianos, regras de associação e *Support Vector Machines*. Apenas os resultados utilizando Floresta Aleatória (Breiman, 2001) são apresentados, pois foi a técnica que obteve os melhores resultados. Floresta Aleatória é um classificador que consiste de várias árvores de decisão. Uma coleção de árvores de decisão é construída seguindo determinadas retrições e cada árvore fornece a sua decisão relativa à categoria da conferência, ou seja, cada árvore dá o seu voto. A categoria com maior número de votos é escolhida. Além disso, utilizamos a técnica de *bagging* para tentar melhorar os resultados.

O desempenho dos classificadores foi medido utilizando as métricas de avaliação acurácia, precisão, revocação e F1, apresentadas na Seção 2.3. A divisão do conjunto de treino e teste utiliza validação cruzada com três partições (*three-fold cross-validation*). Validação cruzada é um processo padrão para avaliar classificadores baseados em técnicas de aprendizado de

máquina (Mitchell, 1997). Consiste em particionar o conjunto de dados original (Φ) em k subconjuntos disjuntos, $\Omega_1, \Omega_2, ..., \Omega_k$. Entre os k subconjuntos, é escolhido o subconjunto Ω_i ($1 \leq i \leq k$) para avaliar o desempenho do classificador e os subconjuntos restantes ($\Phi - \Omega_i$) são usados para treinar o classificador. O processo de validação cruzada é repetido k vezes, variando o valor de i. O resultado de cada execução do classificador para cada valor de i é avaliado, normalmente, utilizando acurácia, precisão, revocação e F1, e o valor médio das k medidas é considerado o valor final. O valor de k é escolhido de acordo com o problema e o tamanho do conjunto de dados, sendo comumente usado o valor 10. Utilizamos $k = 3$, pois a amostra do nosso conjunto de dados é muito pequena (apenas 194 conferências).

Todos os resultados são valores médios de 33 execuções. Em alguns dos experimentos o número de conferências consideradas é muito pequeno e, nesses casos, escolhemos aleatoriamente 50% das conferências para teste e 50% para treino. Os experimentos nos quais utilizamos essa divisão serão especificados explicitamente. Em todos os resultados, mostramos o valor médio das 33 execuções e o intervalo de confiança (IC) de 95% para a média.

5.1 Avaliação das Características

Considerando todas características descristas na Seção 3.2, criamos um conjunto de dados com 21 características sobre as conferências, que foram agrupadas de acordo com o aspecto a que elas estão relacionadas (contagem de citações, taxas de submissão e aceitação, tradição, e características do comitê de programa). Aplicamos dois métodos de seleção de atributos muito populares, denominados Ganho de Informação (*information gain*) e *Chi Quadrado* (χ^2) (Yang e Pedersen, 1997), para avaliar individualmente o poder de discriminação de cada uma dessas características. A Tabela 5.1 mostra o número de atributos mais discriminativos de acordo com o *ranking* produzido pelo Ganho de Informação. Os resultados para o χ^2 são muito similares e, portanto, foram omitidos. Note que, entre as 10 características mais importantes, 8 são baseadas em citações. Note também que, as características baseadas no comitê de programa não aparecem no conjunto Top 10. Essas são as características mais difíceis de serem obtidas, visto que, além da lista dos membros do comitê de programa, também, são necessários dados bibliográficos de alta qualidade. Como podemos ver na Tabela 5.2, as cinco características mais importantes, de acordo com o *ranking* gerado pelo Ganho de Informação são: CCF, C-Factor, CCI, CC, e o Número de Edições.

	Top 5	Top 10	Top 15	Top 20	Número de Características
Citações	4	8	9	10	10
Sub/Aceitação	0	1	2	2	2
Tradição	1	1	1	1	1
Comitê de Programa	0	0	3	7	8

Tabela 5.1: Número de atributos nas posições de topo do *ranking* gerado pelo Ganho de Informação.

Pos	Característica	Pos	Característica	Pos	Característica
1	CCF	8	WPR	15	Betweenness Comitê de Programa
2	C-Factor	9	CIF	16	Closeness Comitê de Programa
3	CCI	10	Taxa de Submissão	17	Tamanho Comitê de Programa
4	CC	11	Taxa de Aceitação	18	FI
5	Número de Edições	12	AVGCC	19	Citações Comitê de Programa
6	Y-Factor	13	H-index Comitê de Programa	20	Coautores Comitê de Programa
7	H-index	14	AVGCC Comitê de Programa	21	Artigos Comitê de Programa

Tabela 5.2: Posição de cada característica no *ranking* gerado pelo Ganho de Informação.

5.2 Classificação em Três Níveis de Qualidade

Nesta seção, estudamos o impacto de cada grupo de características no processo de classificar automaticamente a qualidade das conferências. A Tabela 5.3 mostra os resultados para a nossa amostra de 194 conferências divididas em três categorias (A, B e C), de acordo com o *Perfil-CC Ranking*. Como pode ser observado, os melhores resultados foram obtidos quando utilizamos as características baseadas na contagem de citações, que obteve 56,03% de acurácia e 67,62% de F1. Entretanto, de modo geral, os resultados apresentados na Tabela 5.3 podem ser considerados insatisfatórios. Apenas o classificador que utiliza as características baseadas em citações apresenta acurácia acima de 50%. Isso pode ser conseqüência de se utilizar separadamente cada grupo de características.

Grupo	Acurácia (CI)	Precisão (CI)	Revocação (CI)	F1 (CI)
Citações	56,03(±1,77)	65,06(±1,76)	67,62(±3,00)	67,62(±3,00)
Sub/Aceitação	42,31(±1,62)	47,57(±1,44)	56,24(±2,75)	56,24(±2,75)
Tradição	47,52(±1,79)	58,03(±2,45)	62,40(±3,77)	62,40(±2,37)
Comitê de Programa	47,84(±1,78)	51,27(±1,75)	64,07(±2,98)	56,78(±2,01)
Todos	56,49(±1,68)	67,12(±1,92)	69,35(±3,23)	67,74(±1,83)

Tabela 5.3: Resultados obtidos para os grupos de características quando as conferências são classificadas de acordo com o *Perfil-CC Ranking*. CI é o intervalo de confiança de 95% para a média.

Em seguida, combinamos os quatro grupos de características (mais a forma que a conferência é designada, sua periodicidade e o grupo ao qual a conferência pertence) e reportamos os resultados na última linha da Tabela 5.3. Os resultados foram estatisticamente os mesmos quando utilizamos apenas as características baseadas em citações, com 56,49% de acurácia e 67,74% de F1.

A Figura 5.1 mostra a matriz de confusão para todos os grupos de características. Note que a maioria dos classificadores classifica corretamente as conferências da categoria A. Essas, normalmente, são conferências tradicionais, com os valores muito diferentes para as características em relação às conferências das categorias B e C (como foi mostrado ao longo da Seção 3.2). No entanto, diversas conferências da categoria C são classificadas como B, mostrando que existe grande similaridade entre as conferências dessas categorias.

Um número significativo de conferências da categoria B foi classificado como sendo da categoria A. A grande maioria dessas conferências está muito próxima do limiar que divide

Citação		Predição A	Predição B	Predição C
Rótulo Verdadeiro	A	**62,50%**	32,95%	4,55%
	B	40,85%	**49,30%**	9,85%
	C	28,57%	48,57%	**22,86%**

Comitê de Programa		Predição A	Predição B	Predição C
Rótulo Verdadeiro	A	**62,50%**	35,23%	2,27%
	B	47,89%	**38,03%**	14,08%
	C	48,57%	40,00%	**11,43%**

Tradição		Predição A	Predição B	Predição C
Rótulo Verdadeiro	A	**62,50%**	32,96%	4,54%
	B	46,48%	**45,07%**	8,45%
	C	45,71%	45,71%	**8,58%**

Sub/Aceitação		Predição A	Predição B	Predição C
Rótulo Verdadeiro	A	**51,14%**	29,54%	19,32%
	B	49,30%	**42,25%**	8,45%
	C	51,43%	31,42%	**17,15%**

Todos		Predição A	Predição B	Predição C
Rótulo Verdadeiro	A	**60,23%**	36,36%	3,41%
	B	33,80%	**56,34%**	9,86%
	C	25,71%	51,43%	**22,86%**

Figura 5.1: Matriz de confusão obtida para cada grupo de características quando as conferências são classificadas de acordo com Perfil-CC.

as conferências das categorias A e B no *Perfil-CC Ranking*. Esse problema será discutido em detalhes na Seção 5.2.2.

O fato de que combinar todos os grupos de características não levou a resultados estatisticamente melhores (quando comparados com os resultados obtidos com as características baseadas em citações) foi intrigante. Portanto, realizamos uma análise detalhada dos dados e identificamos três problemas que justificam os resultados obtidos e mostram a dificuldade de se classificar conferências nas categorias A, B e C, conforme proposto pelo projeto Perfil-CC. O primeiro desses problemas é a quantidade de dados faltantes, visto que não conseguimos encontrar todos os dados necessários para algumas das conferências. Esse problema é estudado mais detalhadamente na Seção 5.2.1.

Os outros dois problemas estão relacionados com o nosso gabarito, o *Perfil-CC Ranking*. O primeiro deles é relacionado aos aspectos subjetivos do processo da enquente do Perfil-CC. Por exemplo, quando os especialistas classificam as conferências podem favorecer veículos nos quais possuem publicações ou citam freqüentemente. Ademais, pesquisadores dos diferentes subcampos da Ciência da Computação podem utilizar diferentes critérios para classificar uma conferência como pertencente às categorias A, B ou C; estudamos o impacto desse problema rodando diferentes classificadores para as conferências de cada grupo na Seção 5.2.3.

O último problema é referente aos limiares empregados para dividir as conferências em três categorias. A divisão das conferências em três categorias no *Perfil-CC Ranking* é feita através de limiares que foram estabelecidos previamente, como já mencionado na Seção 3.1.1.1. Como conseqüencia, quando comparamos conferências nas posições de "topo" de cada categoria, podemos certamente dizer que essas conferências pertencem a essa categoria. Entretanto, quando vamos descendo posições no *ranking* e aproximamos da próxima categoria (i.e., aproximamos da linha estabelecida pelo limiar), as características das conferências são equivalentes às conferências da categoria "inferior". Por exemplo, conferências da categoria A próximas ao limiar que divide as categorias A e B são mais similares às conferências do topo da categoria

B do que às conferências do topo da categoria A. Essas conferências são difíceis de serem classificadas na categoria correta, além de induzir o classificador a aprender padrões errados.

5.2.1 Remoção Conferências com Dados Faltantes e Outliers

Em nossa análise anterior, uma das dificuldades apontadas foram os dados faltantes, levando a baixas taxas de acurácia. Para os experimentos discutidos nesta seção, removemos de nossa amostra inicial conferências com dados faltantes para pelo menos um dos grupos de características. A nova amostra, com 156 conferências, foi usada pra treinar classificadores com todas as características. A Tabela 5.4 mostra os resultados obtidos, em que a acurácia e F1 alcançaram 62,88% e 73,08%, respectivamente. Esses resultados são estatisticamente melhores que os obtidos com as 194 conferências (primeira linha da Tabela 5.4). A Figura 5.2 mostra a matriz de confusão para esse experimento. Note que, nesse caso, nenhuma das conferências pertencentes à categoria A foi erroneamente classificada como C. Isso é uma evidência que, no experimento anterior, os dados faltantes estavam induzindo os classificadores a julgar erroneamente conferências da categoria A como C. E ainda, apenas 5,89% dos exemplos da categoria C foram corretamente classificadas. Esse resultado é devido à grande similaridade entre as conferências das categorias B e C, sendo uma tarefa ainda mais difícil diferenciar as conferências da categoria C, para as quais possuímos todos os dados, das conferências da categoria B.

	Acurácia (CI)	Precisão (CI)	Revocação (CI)	F1 (CI)
194 Conferências	56,49(±1,68)	67,12(±1,92)	69,35(±3,23)	67,74(±1,83)
156 Conferências Sem Dados Faltantes	62,88(±1,89)	69,55(±1,90)	77,53(±2,71)	73,08(±1,78)
143 Conferências Sem *Outliers*	67,90(±1,41)	74,96(±1,92)	83,53(±2,03)	78,81(±1,50)

Tabela 5.4: Resultados obtidos pelo classificador com a amostra original (194 Conferências), sem dados faltantes (156 Conferências Sem Dados Faltantes) e sem dados faltantes e *outliers* (143 Conferências Sem *Outliers*).

		Predição		
		A	B	C
Rótulo	A	**79,27%**	20,73%	0,00%
Verdadeiro	B	36,84%	**59,65%**	3,51%
	C	35,29%	58,82%	**5,89%**

Figura 5.2: Matriz de confusão para o classificador com conferências sem dados removidas.

Em um segundo experimento, além das conferências com dados faltantes, também removemos de nossa amostra conferências que consideramos *outliers*. *Outliers* são conferências que não receberam avaliações apropriadas pelos especialistas. Identificamos esses casos porque suas características são muito diferentes, em média, das outras conferências no mesmo grupo. Existem diversas razões para isso, como conferências com baixo impacto entre a comunidade

	Divisões	Acurácia (CI)	Precisão (CI)	Revocação (CI)	F1 (CI)
1	A e B	69,13(±1,68)	73,30(±1,92)	70,34(±2,61)	71,49(±1,68)
2	A e B (limiar A)	75,30(±1,37)	78,30(±0,87)	92,05(±1,68)	84,54(±0,93)
3	A e B (limiar B)	76,98(±1,50)	78,63(±2,79)	58,59(±3,93)	66,32(±2,85)
4	B e C	65,44(±1,95)	69,92(±1,27)	85,34(±3,08)	76,62(±1,59)
5	B e C (limiar B)	80,77(±0,71)	81,93(±0,42)	98,26(±0,85)	89,33(±0,44)
6	B e C (limiar C)	59,52(±2,25)	63,03(±1,78)	69,14(±3,73)	65,61(±2,31)
7	A e C	74,43(±1,83)	79,06(±1,47)	87,94(±2,49)	83,03(±1,32)

Tabela 5.5: Resultados obtidos quando classificamos as conferências em duas categorias com os limiares originais e modificados.

de especialistas consultados ou que cobrem temas muito específicos. Exemplos dessas conferências são a *International Conference on Genetic Algorithms (ICGA)* e a *IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*. No total, 13 conferências foram manualmente identificadas como *outliers* e removidas da amostra. A nova amostra, com 143 conferências, apresentou resultados estatisticamente melhores do que a original, com acurácia de 67,90% e F1 de 78,81% (veja a Tabela 5.4). Esse resultado mostra que essas 13 conferências removidas certamente estavam levando os classificadores a aprender padrões errados sobre os dados e, então, classificando erroneamente novas conferências.

5.2.2 Análise do Limiar das Categorias

Outro problema que dificulta o processo de classificação automática está relacionado aos limiares empregados para dividir as conferências do *Perfil-CC Ranking* em três categorias. Para entender melhor esse problema, nesta seção descrevemos três experimentos envolvendo a amostra original com 194 conferências. Utilizamos a amostra original para analisar isoladamente cada fator que influência no processo de classificação. Experimentos considerando os diversos problemas identificados simultaneamente serão deixados para trabalhos futuros. Primeiramente, o problema original de classificação foi subdividido em problemas menores e testamos o desempenho do classificador quando consideramos apenas duas das três categorias originais: A e B, B e C, e A e C. Depois, variamos os limiares originais e avaliamos o impacto dessa mudança nos resultados. Por último, retornamos ao problema com três categorias e avaliamos o efeito de mudar o limiar original que divide as categorias A e B, e B e C, simultaneamente.

A Tabela 5.5 sumariza os resultados. Como esperado, nos experimentos com o classificador que trabalha com apenas duas categorias e com os limiares originais (linhas 1, 4 e 7), os piores resultados foram obtidos pelo classificador que separa conferências B e C. Esse resultado enfatiza a similaridade das conferências dessas categorias. Os melhores resultados foram obtidos quando classificamos nas categorias A e C, o que confirma os resultados apresentandos na Seção 3.2, onde estudamos as diversas características dessas duas categorias.

No segundo experimento, variamos os valores dos limiares originais escolhidos para separar as categorias A e B, e B e C. Primeiro, criamos uma "zona cinza". Incluímos nessa zona todas

as conferências com pontuação, no *Perfil-CC Ranking*, 10% maior ou menor que a pontuação original da primeira conferência da categoria imediatamente inferior, por exemplo, entre as categorias A e B a "zona cinza" é formada por todas as conferências com valor 10% maior ou menor do que a conferência da categoria B com maior pontuação. Então, duas configurações foram testadas. Na primeira, todas as conferências na zona cinza foram consideradas como conferências pertencentes à categoria imediatamente superior (linhas 2 e 5). Na segunda, as conferências na zona cinza foram incluídas na categoria imediatamente inferior (linhas 3 e 6). Os resultados obtidos são apresentados na Tabela 5.5.

Com respeito aos experimentos variando o limiar para as categorias A e B, os valores de acurácia e precisão em ambos os experimentos são estatisticamente mellhores que os obtidos com o limiar original. Entretanto, os valores da revocação e, conseqüentemente, de F1, são significativamente maiores quando as conferências na zona cinza são consideradas como sendo da categoria B. Esses resultados indicam que as conferências nas posições de topo da categoria B são realmente similares às conferências da categoria A.

Para as conferências na zona cinza das categoria B e C, quando consideramos todas como sendo da categoria B (linha 5), encontramos resultados que são estatisticamente melhores que os obtidos com o limiar original. Isso mostra que ao reduzir o número de exemplos, removendo as conferências da zona cinza da categoria C, os conjuntos de características incluídos em nosso estudo podem facilmente distinguir os exemplos das duas categorias (B e C). Por outro lado, quando consideramos as conferências da zona cinza como sendo da categoria C (linha 6), o problema de classificação torna-se mais difícil e resultados piores são obtidos. Esses experimentos enfatizam ainda mais a similaridade das conferências das categorias B e C.

Por último, variamos simultaneamente ambos os limiares entre as categorias A e B, e B e C. No caso do limiar B/C, os experimentos anteriores mostraram que é melhor aumentar o limiar em 10%, incluindo os antigos exemplos da categoria C na categoria B. Todavia, para o limiar entre as categorias A e B, acreditamos que uma análise mais detalhada é necessária. Em nosso próximo experimento, fixamos o novo limiar B/C como discutido acima e propomos a mesma abordagem do experimento anterior para variar o limiar A/B. A Tabela 5.6 reporta os resultados obtidos quando consideramos as conferências na zona cinza (A/B) como sendo da categoria A e, a seguir, da categoria B. Note os resultados obtidos para acurácia e F1. Quando aumentamos o número de conferências na categoria A, temos uma classificação mais balanceada, com alto valor de F1 e baixa acurácia. Todavia, quando aumentamos o número de conferências na categoria B, temos uma situação oposta, alto valor de acurácia e baixo valor de F1. Esses resultados enfatizam a necessidade de uma análise individual das conferências presentes na zona cinza das categorias A e B.

	Acurácia (CI)	Precisão (CI)	Revocação (CI)	F1 (CI)
Limiar A	62,27(±1,46)	69,24(±1,34)	82,91(±2,25)	75,28(±1,25)
Limiar B	70,20(±1,34)	76,09(±3,49)	56,95(±4,04)	64,13(±2,91)

Tabela 5.6: Resultados obtidos variando simultaneamente os limiares das categorias A, B e C.

5.2.3 Classificação das Conferências por Grupo

Outra dificuldade encontrada no processo de classificação automática das conferências em diferentes níveis de qualidade está relacionada aos critérios subjetivos do processo de classificação manual. Por exemplo, pesquisadores trabalhando em diferentes grupos têm diferentes perfis. Campos de pesquisa maiores e mais antigos podem ter diferentes padrões de citação, critérios de autoria, número de conferências, etc. Essas diferenças podem ter tido grande impacto na construção do *Perfil-CC Ranking*. Nesta seção descrevemos alguns dos experimentos conduzidos para avaliar separadamente conferências de diferenes grupos. Embora essa informação esteja presente em nossos dados, pode não está sendo explorada apropriadamente pelas técnicas de aprendizado de máquina.

Os resultados são apresentados na Tabela 5.7. Como o número de conferências em alguns grupos é muito pequeno, utilizamos 50% dos exemplos para treino e 50% para teste em todos os experimentos com os grupos. Note que os melhores resultados foram obtidos para os grupos IHC e BD, com acurácia de 67,31% e 65,74%, e F1 de 77,22% e 70,93%, respectivamente. Esses dois grupos são muito homogêneos e as conferências de cada categoria bem definidas. Ademais, são os grupos que possuem a melhor cobertura de citações na Libra. Como mostrado anteriormente, as características baseadas em citações estão entre as mais importantes para classificação.

	Acurácia (CI)	Precisão (CI)	Revocação (CI)	F1 (CI)	Num de Conf
BD	65,74(±2,89)	79,65(±5,30)	67,64(±5,86)	70,93(±3,93)	40
BIO	43,64(±4,96)	53,84(±7,19)	67,68(±10,44)	57,58(±7,23)	10
IHC	67,31(±3,37)	82,36(±5,07)	76,46(±5,44)	77,22(±3,20)	24
AM	43,52(±3,30)	54,36(±3,96)	62,61(±5,47)	56,97(±3,35)	32
REDES	51,78(±2,05)	57,38(±1,81)	81,12(±4,03)	66,72(±1,79)	55
WEB	49,21(±3,29)	58,40(±5,38)	59,31(±6,91)	56,28(±4,67)	33
EXG	55,55(±1,61)	68,78(±2,19)	67,91(±3,02)	67,91(±1,90)	-
ING	56,49(±1,68)	67,12(±1,92)	69,35(±3,23)	67,74(±1,83)	-

Tabela 5.7: Resultados obtidos quando classificamos conferências de acordo com seus grupos. ING e EXG são os resultados para todas as conferências incluindo e excluindo a informação do grupo do conjunto de dados.

No entanto, os piores resultados foram obtidos pelos grupos AM e BIO, com acurácia de 43,52% e 43,64%, e F1 de 57,58% e 56,97%, respectivamente. Note que o grupo BIO possui apenas 10 conferências e, ainda, é o grupo com o menor número de especialistas na comunidade que foi consultada na enquete realizada pelo projeto Perfil-CC. Assim, essas conferências podem não ter sido avaliadas adequadamente. Em relação ao grupo AM, uma análise manual mostrou que um número significativo de conferências de alta qualidade, por exemplo, o *IEEE Congress on Evolutionary Computation (CEC)* e a *International Conference on Hybrid Intelligent Systems (HIS)* não possuem características suficientemente distintas para uma separação precisa das conferências de média e baixa qualidade, tornando o processo de classificação mais difícil. Note que os resultados são similares aos obtidos quando não

separamos as conferências em grupos, exceto para os grupos IHC e BD para os quais os resultados são melhores.

Por último, executamos um experimento no qual a informação do grupo foi excluída do conjunto dados. Como mostrado na Tabela 5.7, os resultados excluindo (EXG) e incluindo (ING) a informação do grupo são estatisticamente equivalentes.

Por fim, em decorrência de todos os resultados reportados nesta seção, podemos concluir que a informação do grupo não é essencial para o processo de classificação.

5.2.4 Impacto na Redução do Número de Atributos

As análises conduzidas nesta e na Seção 3.2 mostram os desafios para se classificar as conferências em três níveis de qualidade. Em especial, mostramos que utilizar todos os grupos de características ou apenas as citações levam a resultados similares. Esta seção analisa o impacto de remover características da amostra original. Na Seção 5.1 havíamos ordenado as características de acordo com o *ranking* gerado pelo Ganho de Informação. Aqui utilizamos essa informação para remover incrementalmente as características menos significativas. A Figura 5.3 apresenta os resultados da acurácia e F1. Note que reduzindo o número de características até seis não provoca grandes impactos nos resultados. Apenas depois desse valor a acurácia e o F1 diminuem.

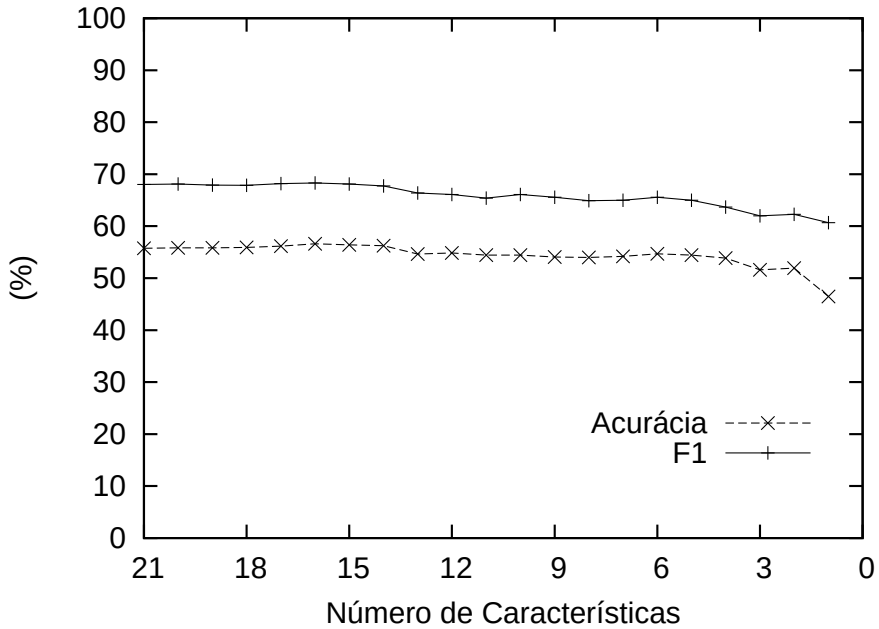

Figura 5.3: Acurácia e F1 obtidos com número reduzidos de atributos.

5.3 Classificação em Duas Categorias

Devido às dificuldades e desafios encontrados quando classificamos as conferências em três categorias de acordo com o Perfil-CC, esta seção simplifica esse problema. Transformamos o

problema anterior com três categorias em um problema com duas categorias: Alta Qualidade (AQ) e Outros (OU). Esse novo problema, similar ao abordado por (Zhuang et al., 2007), é mais simples e fácil de ser tratado, e a sua solução pode ser utilizada como um importante e útil serviço para bibliotecas digitais e sistemas similares. E ainda, analisar esse problema é uma contribuição importante, pois nossa amostra inclui conferências de diversos grupos temáticos e um grande conjunto de características, em contraste com (Zhuang et al., 2007) que inclui apenas conferências de Banco de Dados e utiliza características baseadas no comitê de programa.

Consideramos diferentes cenários para transformar as três categorias em duas. No primeiro, conferências das categorias B e C foram unidas para formar a categoria OU e a categoria A tornou-se a categoria AQ. Na segunda abordagem, as conferências nas últimas posições da categoria A (anteriormente incluídas na zona cinza) foram colocadas juntas com as conferências das categorias B e C na categoria OU, as conferências remanescentes da categoria A tornaram-se a categoria AQ. Por último, apenas as conferências nas 20 posições de topo (Top 20) e 30 posições de topo (Top 30) foram incluídas na categoria AQ e as demais foram incluídas na categoria OU.

Os resultados dos diferentes cenários propostos acima estão sumarizados na Tabela 5.8. Observe que em todos os cenários a acurácia é estatisticamente melhor do que a obtida no problema com três categorias, porém nesse caso a acurácia não é a melhor medida para avaliar o desempenho dos classificadores, visto que a distribuição das categorias para os exemplos é muito desbalanceada. No entanto, os valores de revocação são estatisticamente menores. As matrizes de confusão para esses experimentos são mostradas na Figura 5.4. Note que a dificuldade agora é classificar as conferências na categoria AQ, que possui poucos exemplos. Também, em alguns casos, as conferências da antiga categoria A estão agora na categoria OU e podem confundir os classificadores.

	Acurácia (CI)	Precisão (CI)	Revocação (CI)	F1 (CI)
A e B+C	72,30(±1,43)	76,73(±2,71)	57,11(±3,02)	64,89(±2,18)
A e Limiar A+B+C	79,88(±1,33)	80,50(±3,30)	51,23(±3,91)	61,72(±3,20)
Top 20	94,55(±0,64)	87,72(±4,55)	56,04(±5,24)	66,98(±4,61)
Top 30	90,83(±0,84)	87,31(±4,42)	48,48(±5,05)	60,91(±4,62)

Tabela 5.8: Resultados obtidos quando classificamos as conferências nas categorias AQ e OU.

Na tentativa de melhorar os resultados criamos uma matriz de custo em que os exemplos da categoria AQ erroneamente classificados é associado um custo maior. Essa abordagem força os classificadores a melhorar o seu desempenho nas categorias com menos exemplos, obtendo uma classificação mais balanceada. Nessa abordagem, às conferências corretamente classificadas é associado um custo 0 e às conferências da categoria OU incorretamente classificadas é associado um custo 1. Para conferências da categoria AQ incorretamente classificadas, uma razão de custo (R) variando de 1 até 5 (0% até 500% maior que o custo para os erros na classificação das conferências da categoria OU) foi testada. Note que para $R = 1$, o resultado é equivalente a não se utilizar uma matriz de custo.

A e B+C		Predição A	Predição B
Rótulo Verdadeiro	A	**54,55%**	45,45%
	B	13,21%	**86,79%**

Top 20		Predição A	Predição B
Rótulo Verdadeiro	A	**45,00%**	55,00%
	B	1,15%	**98,85%**

A e Limiar A+B+C		Predição A	Predição B
Rótulo Verdadeiro	A	**58,73%**	41,27%
	B	6,11%	**93,89%**

Top 30		Predição A	Predição B
Rótulo Verdadeiro	A	**50,00%**	50,00%
	B	1,22%	**98,78%**

Figura 5.4: Matrizes de confusão para os cenários com apenas duas categorias.

A Figura 5.5 mostra os resultados ao utilizarmos uma matriz de custo para o cenário com apenas as 20 conferências do topo (Top 20) sendo consideradas da categoria AQ. Note que quando R vale 3,75 o valor do F1 é igual 70,23%. Quando R é igual 1, o valor do F1 é 66,89%. Todavia, o resultado mais interessante pode ser observado ao analisarmos as curvas de precisão e revocação. Quando aumentamos o valor de R, os valores da revocação também aumentam e os valores da precisão diminuem. Nesse caso, altos valores de revocação significam uma classificação mais balanceada dos exemplos em ambas as categorias. Isso reflete no aumento do valor de F1, que mede precisamente a relação entre a precisão e a revocação.

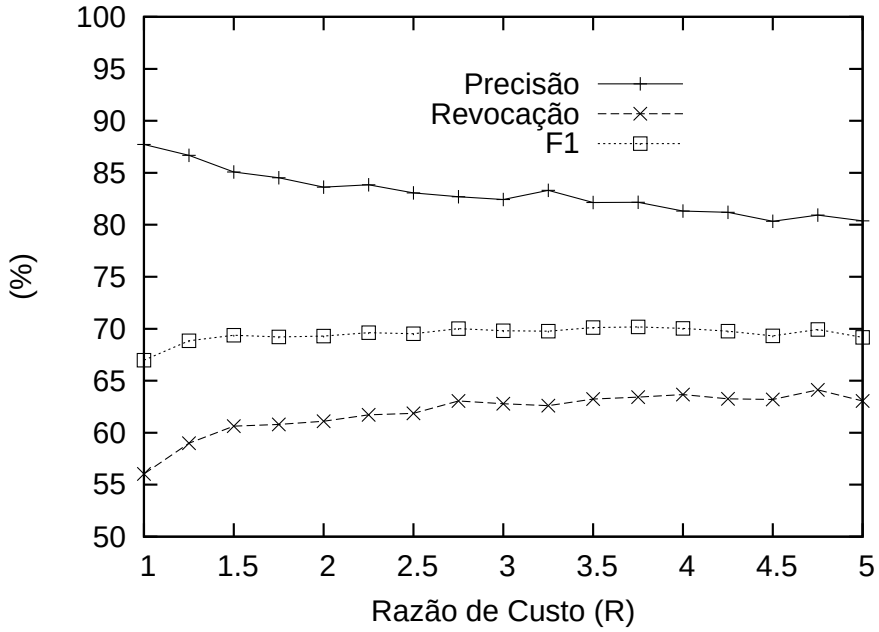

Figura 5.5: Precisão, Revocação e F1 obtidos variando a razão de custo (R).

Capítulo 6

Conclusões e Trabalhos Futuros

6.1 Revisão do Trabalho

Nesta dissertação abordamos o problema de avaliação automática da qualidade de conferências utilizando duas abordagens diferentes. Na primeira, propomos um conjunto de novas métricas para avaliar a qualidade de conferências com base nas citações bibliográficas. As novas métricas foram inspiradas nas deficiências encontradas a partir de uma análise detalhada das métricas utilizadas para periódicos e buscam capturar importantes características como longevidade, popularidade, prestígio, e periodicidade. A grande vantagem dessa abordagem é a fácil aplicação a bibliotecas digitais visto que as citações bibliográficas são comumente encontradas nesses repositórios.

Nos experimentos relacionados a essa abordagem, comparamos a eficácia na avaliação automática de conferências das métricas utilizadas para avaliação de periódicos e do conjunto de novas métricas proposto. Os resultados mostraram que as novas métricas tiveram um desempenho significativamente superior em relação às métricas utilizadas para periódicos quando comparadas com um gabarito produzido por especialistas. Entre as métricas existentes, uma métrica muito simples, a Contagem de Citações (CC), obteve os melhores resultados no geral. O Fator de Impacto, métrica mais tradicional para periódicos, obteve resultados muito ruins, o que mostra que a rede de citações das conferências possui características distintas da rede de citações de periódicos. Portanto, são necessárias métricas específicas para conferências que se adequem a essas características. Entre as novas métricas propostas, aquelas que capturam um número maior de aspectos intrínsecos sobre as conferências, CCF e C-Factor, obtiveram os melhores resultados. Essas métricas obtiveram ganhos de até 8,4% na comparação de similaridade e 7,8% na acurácia na classificação em níveis de qualidade em relação ao gabarito (*Perfil-CC Ranking*) quando comparadas com as melhores métricas tradicionais para avaliação da qualidade de periódicos.

Na segunda abordagem, consideramos o problema de classificação automática de conferências em níveis de qualidade pré-definidos utilizando um grande número de características sobre as conferências, que foram extraídas de bibliotecas digitais, sítios de conferências e outras fontes da Web, e buscam atender os diversos aspectos considerados por um especialista quando

realiza manualmente o julgamento da qualidade de uma conferência. Essa abordagem é diferente de outros trabalhos relacionados (Bollen et al., 2005; Zhuang et al., 2007; Yan e Lee, 2007), visto que focamos na combinação efetiva de diversos tipos de característica ao invés de apenas um tipo específico. Assim, podemos capturar um número importante de aspectos que são de fato utilizados na prática para esse tipo de avaliação. Para tal, avaliamos o poder de discriminação de diversos tipos de característica e concluímos que as características mais efetivas são as baseadas em citações seguida pela tradição da conferência. Esse estudo pode ser um possível caminho para inspirar novos critérios e métricas para avaliação de conferências. Além disso, nosso estudo sobre essas características pode ser utilizado para padronizar os critérios usados por especialistas para avaliação da qualidade de conferências.

No entanto, obter um grande volume de dados sobre as conferências não é uma tarefa trivial, sendo essa a principal desvantagem dessa abordagem. Contudo, com a expansão de repositórios de alta qualidade, como as bibliotecas digitais, essa tarefa torna-se cada vez mais fácil.

Nos experimentos relacionados a essa abordagem, quando combinamos as diversas características através de técnicas de aprendizado de máquina, obtivemos sucesso em separar conferências de alta qualidade das conferências de média e baixa qualidade. Entretanto, mostramos que separar esses dois últimos tipos de conferência é uma tarefa muito mais difícil. A classificação em apenas duas categorias (alta e baixa qualidade) mostrou resultados mais precisos.

6.2 Trabalhos Futuros

Apesar dos bons resultados na avaliação da qualidade de conferências usando as métricas baseadas em citações propostas, existe muito espaço para melhorias, como mostrado pelas classificações geradas por nossas métricas e pelos especialistas que apresentam grandes diferenças. Assim, como possíveis extensões sugerimos uma análise profunda da rede de citações de conferências na busca de possíveis características intrínsecas que possam ser incorporadas a novas métricas. Por exemplo, o impacto de auto-citações no total de citações não é conhecido. As auto-citações podem ser consideradas uma resposta positiva sobre à qualidade dos artigos publicados nas conferências ou superestimam o total de citações? Outra possibilidade é estudar melhor o impacto de cada característica incorporada em nossas métricas e como cada uma delas influência o resultado final. Sugerimos, também, a investigação de uma técnica conhecida como Programação Genética (Banzhaf et al., 1998; Koza, 1992) para construção de métricas mais robustas e, através do processo evolutivo dessa técnica, ampliar o entendimento da importância de cada característica (longevidade, popularidade, prestígio, etc.). Além disso, investigações baseadas em dados de outras fontes de citações bibliográficas e com um número maior de conferências são necessárias, com o objetivo de melhorar ainda mais essas métricas.

Como possíveis extensões da abordagem que utiliza técnicas de aprendizado de máquina, sugerimos primeiro investigar formas de extrair automaticamente as características mais im-

portantes para avaliação de qualidade, visto que a falta de informação revelou-se um aspecto importante no desempenho dos métodos. Também, sugerimos investigar diferentes formas de combinar as características, por exemplo, através de regressão, se considerarmos o problema de classificação em níveis como um problema de ordenação de conferências. Além disso, sugerimos aumentar o número e a diversidade das características utilizadas, considerando, por exemplo, os patrocinadores e a qualidade dos artigos apresentados nas conferências.

Apêndice A

Divisão do Campo de Ciência da Computação em Grupos Temáticos

ID	Grupo
1	Algoritmos e Teoria da Computação
2	Bancos de Dados, Recuperação de Informação, Bibliotecas Digitais, Mineração de Dados
3	Biologia Computacional
4	Computação Aplicada
5	Computação Gráfica, Processamento de Imagens, Visão Computacional
6	Concepção de Circuitos Integrados
7	Engenharia de Software, Métodos Formais
8	Geoinformática
9	Informática na Educação
10	Inteligência Artificial
11	Interação Humano Computador, Sistemas Colaborativos
12	Linguagens de Programação
13	Multitemáticas
14	Pesquisa Operacional e Otimização Combinatória
15	Redes de Computadores, Sistemas Distribuídos, Sistemas P2P
16	Simulação e Modelagem
17	Web, Sistemas Multimídia e Hipermídia
19	Jogos e Entretenimento, Realidade Virtual
21	Sistemas de Informação
22	Aprendizado de Máquina
23	Robótica, Controle e Automação
25	Segurança
26	Arquitetura de Computadores, Processamento de Alto Desempenho, Sistemas Operacionais
27	Sistemas Embarcados, Sistemas de Tempo Real, Sistemas Tolerantes a Falhas
28	Computação Ubíqua
29	Formalismos, Lógica, Semântica da Computação
30	Processamento de Língua Natural

Tabela A.1: Lista dos 27 grupos existentes no Perfil-CC e seus respectivos identificadores (ID).

Apêndice B

Conferências da Amostra de Dados Usada

Sigla	Nome	SeqPos	Pont	Cat	Grupo
INFOCOM	Annual Joint Conference of the IEEE Computer and Communications Societies	1	1,00000	A	15
SIGCOMM	ACM Special Interest Group on Data Communications Conference	2	0,97980	A	15
CSCW	ACM Conference on Computer Supported Cooperative Work	3	0,96552	A	11
CHI	Conference on Human Factors in Computing Systems	4	0,96429	A	11
KDD	ACM SIGKDD International Conference on Knowledge Discovery and Data Mining	5	0,94340	A	2
WWW	International World Wide Web Conference	6	0,94118	A	18
ICML	International Conference on Machine Learning	7	0,93939	A	22
ACM-MM	ACM Multimedia Conference	8	0,93333	A	18
SIGMOD	ACM SIGMOD International Conference on Management of Data Conference	9	0,92683	A	2
ICC	IEEE International Conference on Communications	10	0,90991	A	15
Middleware	ACM/IFIP/USENIX International Middleware Conference	11	0,90909	A	15
VLDB	International Conference on Very Large Data Bases	12	0,89855	A	2
ICDE	IEEE International Conference on Data Engineering	13	0,89796	A	2
ICDCS	IEEE International Conference on Distributed Computing Systems	14	0,88542	A	15
IJCNN	IEEE International Joint Conference on Neural Networks	15	0,88542	A	22
Hypertext	ACM Conference on Hypertext and Hypermedia	16	0,88000	A	18
UIST	ACM Symposium on User Interface Software and Technology	17	0,86364	A	11
PODS	ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems	18	0,86325	A	2
ER	International Conference on Conceptual Modeling	19	0,85586	A	2
JCDL	ACM IEEE Joint Conference on Digital Libraries	20	0,85586	A	2
ICDM	IEEE International Conference on Data Mining	21	0,85417	A	2
ICWS	IEEE International Conference on Web Services	22	0,85185	A	18
ICME	IEEE International Conference on Multimedia and Expo	23	0,85000	A	18
NIPS	Neural Information Processing Systems	24	0,84314	A	22
PODC	ACM Symposium on Principles of Distributed Computing	25	0,83951	A	15
WI	IEEE/WIC/ACM International Conference on Web Intelligence	26	0,83333	A	18
FUZZ	IEEE International Conference on Fuzzy Systems	27	0,82716	A	22

Sigla	Nome	SeqPos	Pont	Cat	Grupo
ECML	European Conference on Machine Learning	28	0,82292	A	22
GECCO	Genetic and Evolutionary Computation Conference	29	0,82143	A	22
CIKM	International Conference on Information and Knowledge Management	30	0,81944	A	2
IMC	ACM/SIGCOMM Internet Measurement Conference	31	0,81111	A	15
SIGIR	Annual International ACM SIGIR Conference on Research and Development in Information Retrieval	32	0,81061	A	2
COLT	Annual Conference on Computational Learning Theory	33	0,80769	A	22
SIGDOC	ACM International Conference on Design of Communication	34	0,80702	A	18
HPDC	IEEE International Symposium on High Performance Distributed Computing	35	0,80247	A	15
Networking	IFIP International Conference on Networking	36	0,79570	A	15
IFSA	International Fuzzy Systems Association World Congress	37	0,79365	A	22
LCN	IEEE Conference on Local Computer Networks	38	0,78667	A	15
CEC	IEEE Congress on Evolutionary Computation	39	0,78495	A	22
ISCC	IEEE Symposium on Computers and Communications	40	0,77778	A	15
INTERACT	IFIP TC13 International Conference on Human-Computer Interaction	41	0,77778	A	11
DocEng	ACM Symposium on Document Engineering	42	0,76812	A	18
MASCOTS	IEEE International Symposium on Modeling, Analysis and Simulation of Computer and Telecommunication Systems	43	0,76190	A	15
Group	ACM Group Conference	44	0,76000	A	11
ICANN	International Conference on Artificial Neural Networks	45	0,75490	A	22
SENSYS	ACM Conference on Embedded Networked Sensor Systems	46	0,75309	A	15
ISMB	Intelligent Systems in Molecular Biology	47	0,75000	A	3
P2P	International Conference on Peer-to-Peer Computing	48	0,75000	A	15
PKDD	European Conference on Principles and Practice of Knowledge Discovery in Databases	49	0,74815	A	2
ICNP	IEEE International Conference on Network Protocols	50	0,74359	A	15
RECOMB	Annual International Conference on Research in Computational Molecular Biology	51	0,73611	A	3
ICPR	International Conference on Pattern Recognition	52	0,73077	A	22
ISM	IEEE International Symposium on Multimedia	53	0,73016	A	18
DSOM	IFIP/IEEE International Workshop on Distributed Systems: Operations and Management	54	0,72727	A	15
POLICY	IEEE International Workshop on Policies for Distributed Systems and Networks	55	0,71429	A	15
EDBT	International Conference on Extending Database Technology	56	0,71053	A	2
HCI	International Conference on Human-Computer Interaction	57	0,70667	A	11
BIBE	IEEE International Symposium on Bioinformatics and Bioengineering	58	0,70238	A	3
SDM	SIAM International Conference on Data Mining	59	0,69841	A	2
DIS	Designing Interative Systems Conference	60	0,69697	A	11
VL/HCC	IEEE Symposium on Visual Languages and Human-Centric Computing	61	0,69333	A	11
CIBCB	IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology	62	0,69231	A	3
CEC2	IEEE Conference on Electronic Commerce Technology	63	0,68421	A	18
ECOWS	IEEE European Conference on Web Services	64	0,68182	A	18
IUI	International Conference on Intelligent User Interfaces	65	0,67949	A	11
ICCC	International Conference on Computer Communication	66	0,67816	A	15
AINA	International Conference on Advanced Information Networking and Applications	67	0,67778	A	15
ECSCW	European Conference on Computer Supported Cooperative Work	68	0,66667	A	11
IPOM	IEEE International Workshop on IP Operations and Management	69	0,66667	A	15
ILP	International Conference on Inductive Logic Programming	70	0,66667	A	22
DAIS	IFIP International Conference on Distributed Applications and Interoperable Systems	71	0,66667	A	15
IPCCC	IEEE International Performance Computing and Communications Conference	72	0,66667	A	15

Sigla	Nome	SeqPos	Pont	Cat	Grupo
MMNS	IFIP/IEEE Conference on Management of Multimedia Networks and Services	73	0,66667	A	18
ICDT	International Conference on Database Theory	74	0,66667	A	2
FORTE	IFIP WG 6,1 International Conference on Formal Methods for Networked and Distributed Systems	75	0,66667	A	15
WISE	International Conference on Web Information Systems Engineering	76	0,66667	A	18
CIDM	IEEE Symposium on Computational Intelligence and Data Mining	77	0,65942	A	2
ICCCN	International Conference on Computer Communications and Networks	78	0,65714	A	15
WABI	Workshop on Algorithms in Bioinformatics	79	0,65432	A	3
ICONIP	International Conference on Neural Information Processing	80	0,65385	A	22
ESWC	European Semantic Web Conference	81	0,65217	A	18
HIS	International Conference on Hybrid Intelligent Systems	82	0,64368	A	22
IDA	International Symposium on Intelligent Data Analysis	83	0,64103	A	22
NOSSDAV	Network and Operating System Support for Digital A/V	84	0,63889	A	15
NCA	IEEE International Symposium on Network Computing and Applications	85	0,63333	A	15
ICDCN	IEEE International Conference on Distributed Computing and Networking	86	0,63218	A	15
CSB	IEEE Computational Systems Bioinformatics Conference	87	0,62821	A	3
EHCI-DSVIS	IFIP Working Conference on Engineering for Human-Computer Interaction	88	0,62667	A	11
ISWC	International Semantic Web Conference	89	0,62121	B	18
DaWaK	International Conference on Data Warehousing and Knowledge Discovery	90	0,62016	B	2
ECCB	European Conference on Computational Biology	91	0,61905	B	3
EuroGP	European Conference on Genetic Programming	92	0,61538	B	22
DISC	International Symposium on Distributed Computing	93	0,61538	B	15
CRIWG	International Workshop on Groupware	94	0,60920	B	11
DEXA	International Conference on Database and Expert Systems Applications	95	0,60897	B	2
DOA	International Symposium on Distributed Objects and Applications	96	0,60494	B	15
ADBIS	Symposium on Advances in Database and Information Systems	97	0,60417	B	2
WIDM	ACM International Workshop on Web Information and Data Management	98	0,60000	B	2
USITS	USENIX Symposium on Internet Technologies and Systems	99	0,60000	B	15
NSDI	Symposium on Networked Systems, Design and Implementation	100	0,59722	B	15
BIOCOMP	International Conference on Bioinformatics and Computational Biology	101	0,58974	B	3
OPODIS	International Conference on Principles of Distributed Systems	102	0,58730	B	15
CSCWD	International Conference on Computer Supported Cooperative Work in Design	103	0,58333	B	11
IDEAS	International Database Engineering and Applications Symposium	104	0,58140	B	2
ECDL	European Conference on Research and Advanced Technology for Digital Libraries	105	0,58095	B	2
SSDBM	International Conference on Scientific and Statistical Database Management	106	0,58095	B	2
SEAL	International Conference on Simulated Evolution and Learning	107	0,57971	B	22
ICMLA	International Conference on Machine Learning and Applications	108	0,57692	B	22
SPIRE	Symposium on String Processing and Information Retrieval	109	0,57576	B	2
KES	International Conference on Knowledge-Based and Intelligent Information and Engineering Systems	110	0,57576	B	22
ESANN	European Symposium on Artificial Neural Networks	111	0,56989	B	22

Sigla	Nome	SeqPos	Pont	Cat	Grupo
WETICE	IEEE International Workshops on Enabling Technologies: Infrastructure for Collaborative Enterprises	112	0,56790	B	11
DOLAP	International Workshop on Data Warehousing and OLAP	113	0,56757	B	2
ICN	International Conference on Networking	114	0,56410	B	15
MIR	ACM International Workshop on Multimedia Information Retrieval	115	0,56190	B	2
AVI	International Working Conference on Advanced Visual Interfaces	116	0,56000	B	11
IWQoS	International Workshop on Quality of Service	117	0,55952	B	15
LA-WEB	Latin American Web Congress	118	0,55556	B	18
ECWeb	International Conference on Electronic Commerce and Web Technology	119	0,55556	B	18
CADUI	International Conference on Computer-Aided Design of User	120	0,55072	B	11
BC	IFIP International Conference on Broadband Communications	121	0,55072	B	15
DUX	ACM Designing for User Experience Conference	122	0,55072	B	11
EuroCOLT	European Conference on Computational Learning Theory	123	0,54762	B	22
UM	International Conference on User Modelling	124	0,54667	B	11
ESTIMedia	IEEE Workshop on Embedded Systems for Real-Time Multimedia	125	0,54545	B	18
AH	International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems	126	0,53968	B	18
NPC	IFIP International Conference on Network and Parallel Computing	127	0,53623	B	15
DBPL	International Symposium on Database Programming	128	0,53509	B	2
MMM	International Multi-Media Modelling Conference	129	0,53333	B	18
ICCI	IEEE International Conference on Cognitive Informatics	130	0,53333	B	22
ALT	International Conferences on Algorithmic Learning Theory	131	0,52632	B	22
MobileHCI	International Conference on Human-Computer Interaction with Mobile Devices and Services	132	0,52564	B	11
CC	Creativity and Cognition Conference	133	0,51515	B	11
DASFAA	International Conference on Database Systems for Advanced Applications	134	0,51515	B	2
WebDB	International Workshop on the Web and Databases	135	0,50833	B	2
ICCB	International Conference on Case-Based Reasoning	136	0,50725	B	22
SAINT	International Symposium on Applications and the Internet	137	0,50000	B	15
IIWAS	Information Integration and Web-based Applications and Services	138	0,50000	B	18
IPTPS	International Workshop on Peer-to-Peer Systems	139	0,49383	B	15
TestCom	IFIP International Conference on Testing Communicating Systems	140	0,49275	B	15
MMSP	International Workshop on Multimedia Signal Processing	141	0,49206	B	18
EuroITV	European Interactive TV Conference	142	0,49206	B	18
PAM	Passive and Active Measurement Conference	143	0,48718	B	15
TREC	Text Retrieval Conference	144	0,48649	B	2
MCS	International Workshop on Multiple Classifier Systems	145	0,48000	B	22
ICGA	International Conference on Genetic Algorithms	146	0,48000	B	22
ISADS	International Symposium on Autonomous Decentralized Systems	147	0,47619	B	15
Broadnets	International Conference on Broadband Communications, Networks and Systems	148	0,47222	B	15
MLDM	IAPR International Conference on Machine Learning and Data Mining	149	0,46667	B	22
WEBIST	International Conference on Web Information Systems and Technologies	150	0,46667	B	18
SWWS	Semantic Web and Web Services	151	0,46667	B	18
RIDE	International Workshop on Research Issues in Data Engineering	152	0,45946	B	2
DSV-IS	International Workshop on Design, Specification and Verification of Interactive Systems	153	0,45833	B	11
CIC	International Conference on Communications in Computing	154	0,44872	B	15

Sigla	Nome	SeqPos	Pont	Cat	Grupo
DAS	International Workshop on Document Analysis Systems	155	0,44444	B	18
XSym	International XML Database Symposium	156	0,44444	B	2
MIS	Workshop on Multimedia Information Systems	157	0,44444	B	18
Web3D	International Conference on 3D Web Technology	158	0,43939	B	18
ADMA	International Conference on Advanced Data Mining and Applications	159	0,43333	B	2
EDOCW	IEEE International Enterprise Distributed Object Computing Conference Workshops	160	0,42529	C	15
IWANN	International Work-Conference on Artificial and Natural Neural Networks	161	0,42308	C	22
IDEAL	International Conference on Intelligent Data Engineering and Automated Learning	162	0,42029	C	22
FQAS	International Conference on Flexible Query Answering Systems	163	0,41975	C	2
METMBS	International Conference on Mathematics and Engineering Techniques in Medicine and Biological Sciences	164	0,41667	C	3
NLDB	International Conference on Applications of Natural Language to Data Bases	165	0,41414	C	2
ICMI	International Conference on Multimodal Interfaces	166	0,41270	C	11
GIR	Workshop on Geographic Information Retrieval	167	0,41111	C	2
COMSWARE	ICST International Conference on COMmunication System softWAre and MiddlewaRE	168	0,41026	C	15
Qsine	International Conference on Quality of Service in Heterogeneous Wired/Wireless Networks	169	0,41026	C	15
TAMODIA	International Workshop on Task Models and Diagrams for User Interface Design	170	0,40278	C	11
ICMLC	International Conference on Machine Learning and Cybernetics	171	0,40278	C	22
ICETE	International Conference on E-Business and Telecommunication Networks	172	0,39130	C	15
DNA	International Meeting on DNA Computing	173	0,38596	C	3
W4A	International Cross-Disciplinary Conference on Web Accessibility	174	0,38596	C	18
PDC	Participatory Design Conference	175	0,38333	C	11
WebNet	World Conference on the WWW and Internet	176	0,38095	C	18
DS	International Conference on Discovery Science	177	0,37681	C	22
DIWeb	International Workshop on Data Integration over the Web	178	0,37500	C	2
ARES	International Conference on Availability, Reliability and Security	179	0,36000	C	15
DBISP2P	International Workshop on Databases, Information Systems and Peer-to-Peer Computing	180	0,36000	C	15
IKE	International Conference on Information and Knowledge Engineering	181	0,35185	C	2
NetGames	Workshop on Network and System Support for Games	182	0,34483	C	15
CIDR	Biennial Conference on Innovative Data Systems Research	183	0,33333	C	2
DCW	Workshop on Distributed Communities on the Web	184	0,33333	C	18
DCC	Data Compression Conference	185	0,33333	C	2
PDIS	International Conference on Parallel and Distributed Information Systems	186	0,33333	C	15
ISMIR	International Conference on Music Information Retrieval	187	0,32143	C	2
WCW	Web Caching Workshop	188	0,31884	C	18
ISSADS	IEEE International Symposium and School on Advance Distributed Systems	189	0,31746	C	15
WAIM	Interational Conference on Web Age Information Management	190	0,31579	C	18
MLMTA	International Conference on Machine Learning and Applications	191	0,28333	C	22
CANPC	Workshop on Communication, Architecture, and Applications for Network-Based Parallel Computing	192	0,27586	C	15
SAPIR	International Workshop on Service Assurance with Partial and Intermittent Resources	193	0,26667	C	15
SIROCCO	Colloquium on Structural Information and Communication Complexity	194	0,25758	C	15

Tabela B.1: Sigla ou nome curto, nome, posição, pontuação obtida no *Perfil-CC Ranking*, categoria da conferência (A, B ou C) e identificação (ID) do grupo das 194 conferências do Perfil-CC utilizadas em nossos experimentos.

Referências Bibliográficas

Amin, M. e Mabe, M. (2000). Impact factors: Use and abuse. *Perspectives in Publishing*, 1:1–6.

Banzhaf, W.; Francone, F. D.; Keller, R. E. e Nordin, P. (1998). *Genetic programming: an introduction: on the automatic evolution of computer programs and its applications*. Morgan Kaufmann Publishers Inc., San Francisco, CA.

Bollen, J. e de Sompel, H. V. (2008). Usage impact factor: The effects of sample characteristics on usage-based impact metrics. *Journal of the American Society for Information Science and Technology*, 59(1):136–149.

Bollen, J.; de Sompel, H. V. e Rodriguez, M. A. (2008). Towards usage-based impact metrics: first results from the MESUR project. In *Proceedings of the 8th ACM/IEEE Joint Conference on Digital Libraries*, pp. 231–240, Pittsburg, PA, USA.

Bollen, J.; de Sompel, H. V.; Smith, J. A. e Luce, R. (2005). Toward alternative metrics of journal impact: a comparison of download and citation data. *Information Processing and Management*, 41(6):1419–1440.

Bollen, J.; Rodriguez, M. A. e de Sompel, H. V. (2006). Journal Status. *Scientometrics*, 69(3):669–687.

Braun, T.; Glänzel, W. e Schubert, A. (2006). A hirsch-type index for journals. *Scientometrics*, 69(1):169–173.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

Brin, S. e Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, 30(1–7):107–117.

Giles, C. L.; Bollacker, K. D. e Lawrence, S. (1998). CiteSeer: An Automatic Citation Indexing System. In *Digital Libraries 98 - The Third ACM Conference on Digital Libraries*, pp. 89–98, Pittsburgh, PA, USA.

Harrison, G. A. e Worden, E. W. (2007). Genetically programmed learning classifier system description and results. In *Proceedings of the 2007 GECCO Conference Companion on Genetic and Evolutionary Computation*, pp. 2729–2736, London, United Kingdom.

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

Joachims, T. (1998). Text categorization with support vector machines: learning with many relevant features. In *Proceedings of 10th European Conference on Machine Learning*, number 1398, pp. 137–142.

Katsaros, D.; Sidiropoulos, A. e Manolopoulos, Y. (2007). Age-decaying h-index for Social Networks of Citations. In *Proceedings of the Workshop on Social Aspects of the Web*, pp. 27–43, Poznan, Poland.

Kendall, M. G. (1938). A New Measure of Rank Correlation. *Biometrika*, 30(1/2):81–93.

Kohavi, R. e Provost, F. (1998). Glossary of terms. *Machine Learning*, 30:271–274.

Koza, J. R. (1992). *Genetic programming: on the programming of computers by means of natural selection*. MIT Press, Cambridge, MA.

Laender, A. H. F.; de Lucena, C. J. P.; Maldonado, J. C.; Silva, E. S. e Ziviani, N. (2008). Assessing the Research and Education Quality of the Top Brazilian Computer Graduate Programs. *ACM SIGCSE Bulletin*, 40(2):135–145.

Larsen, B. e Ingwersen, P. (2006). Using citations for ranking in digital libraries. In *Proceedings of the 6th ACM/IEEE Joint Conference on Digital Libraries*, pp. 370–370, Chapel Hill, NC, USA.

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

MacRoberts, M. H. e MacRoberts, B. R. (1989). Problems of citation analyses: Critical Review. *Journal of American Society for Information Science*, 40(5):342–349.

Martins, W. S.; Gonçalves, M. A.; Laender, A. H. F. e Pappa, G. L. (2009a). Learning to Assess the Quality of Scientific Conferences: A Case Study in Computer Science. In *Proceedings of the 9th ACM/IEEE Joint Conference on Digital Libraries*, Austin, TX, USA (a ser publicado).

Martins, W. S.; Gonçalves, M. A.; Laender, A. H. F. e Ziviani, N. (2008). Avaliação da Qualidade de Conferências usando Citações Bibliográficas. *Simpósio Brasileiro de Banco de Dados*, Resumo.

Martins, W. S.; Gonçalves, M. A.; Laender, A. H. F. e Ziviani, N. (2009b). Assessing the Quality of Scientific Conferences Based on Bibliographic Citations. *Scientometrics*, (a ser publicado).

Mitchell, T. (1997). *Machine Learning*. McGraw-Hill, Hill, New York, NY, USA.

Nie, Z.; Ma, Y.; Shi, S.; Wen, J.-R. e Ma, W.-Y. (2007). Web Object Retrieval. In *Proceedings of the 16th International World Wide Web Conference*, pp. 81–90, Banff, Alberta, Canada.

Patterson, D. A. (2004). The Health of Research Conferences and the Dearth of Big Idea Papers. *Communications of the ACM*, 47(1):23–24.

Quinlan, J. R. (1993). *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Mateo, CA.

Quinlan, J. R. (1996). Bagging, Boosting, and C4.5. In *Proceedings of the 13th National Conference on Artificial Intelligence*, volume 1, pp. 725–730, Portland, OR, USA.

Rahm, E. e Thor, A. (2005). Citation Analysis of Database Publications. *SIGMOD Record*, 34(4):48–53.

Ribeiro-Neto, B. e Baeza-Yates, R. (1999). *Modern Information Retrieval*. Addison Weslay Longman Publishing Co., Inc., Boston, MA.

Saha, S.; Saint, S. e Christakis, D. A. (2003). Impact factor: a valid measure of journal quality? *Journal of the Medical Library Association*, 1(62):42–46.

Schölkopf, B.; Burges, C. J. e Smola, A. J. (1999). *Advances in kernel methods: support vector learning*. MIT Press, Cambridge, MA, USA.

Seglen, P. O. (1997). Why the impact factor of journals should not be used for evaluating research. *British Medical Journal*, 314(7079):498–502.

Sidiropoulos, A.; Katsaros, D. e Manolopoulos, Y. (2007). Generalized Hirsch h-index for disclosing latent facts in citation networks. *Scientometrics*, 72(2):253–280.

Sidiropoulos, A. e Manolopoulos, Y. (2006). Generalized Comparison of Graph-Based Ranking Algorithms for Publications and Authors. *Journal of Systems and Software*, 79(12):1679–1700.

Smith, R. E. e Jiang, M. K. (2007). MILCS: a mutual information learning classifier system. In *Proceedings of the 2007 GECCO Conference Companion on Genetic and Evolutionary Computation*, pp. 2945–2952, London, United Kingdom.

Souto, M. A. M.; Warpechowski, M. e de Oliveira, J. P. M. (2007). An Ontological Approach for the Quality Assessment of Computer Science Conferences. In *Proceedings of the 2007 Workshop on Quality of Information Systems QoIS*, volume LNCC 4802, pp. 202–212, Auckland, New Zealand.

Wasserman, S. e Faust, K. (1994). Social Networks Analysis: Methods and Application.

Yan, S. e Lee, D. (2007). Toward alternative measures for ranking venues: a case of database research community. In *Proceedings of the 7th ACM/IEEE Joint Conference on Digital Libraries*, pp. 235–244, Vancouver, BC, Canada.

Yang, Y. e Pedersen, J. O. (1997). A Comparative Study on Feature Selection in Text Categorization. In *Proceedings of the 14th International Conference on Machine Learning*, pp. 412–420, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Yu, G. e Wang, L. (2007). The self-cited rate of scientific journals and the manipulation of their impact factors. *Scientometrics*, 73(3):321–330.

Zhuang, Z.; Elmacioglu, E.; Lee, D. e Giles, C. L. (2007). Measuring conference quality by mining program committee characteristics. In *Proceedings of the 7th ACM/IEEE Joint Conference on Digital Libraries*, pp. 225–234, Vancouver, BC, Canada.