UNIVERSIDADE FEDERAL DO RIO GRANDE DO SUL
INSTITUTO DE INFORMÁTICA
PROGRAMA DE PÓS-GRADUAÇÃO EM COMPUTAÇÃO

CRISTIANO ROBERTO CERVI

Rep-Index – Uma Abordagem Abrangente e Adaptável Para Identificar Reputação Acadêmica

Tese apresentada como requisito parcial para a obtenção do grau de Doutor em Ciência da Computação

Prof[a]. Dr[a]. Renata de Matos Galante
Orientadora

Prof. Dr. José Palazzo Moreira de Oliveira
Coorientador

Porto Alegre, dezembro de 2013.

CIP – CATALOGAÇÃO NA PUBLICAÇÃO

Cervi, Cristiano Roberto

Normas para Apresentação de Dissertações do Instituto de Informática e do PPGC [manuscrito] / Cristiano Roberto Cervi. – 2013.

121 f.:il.

Orientador: Renata Galante; Co-orientador: José Palazzo Moreira de Oliveira.

Tese (doutorado) – Universidade Federal do Rio Grande do Sul. Programa de Pós-Graduação em Computação. Porto Alegre, BR – RS, 2013.

1.Reputação Acadêmica. 2.Modelo de Perfil. 3.Métricas Científicas. 4.Adaptabilidade. I. Galante, Renata. II. Oliveira, José Palazzo Moreira de. III. Rep-Index – Uma Abordagem Abrangente e Adaptável Para Identificar Reputação Acadêmica.

A melhor maneira de nos prepararmos para o futuro é concentrar toda a imaginação e entusiasmo na execução perfeita do trabalho de hoje.

— DALE CARNEGIE

"A persistência é o menor caminho para o êxito."

— CHARLES SPENCER CHAPLIN

AGRADECIMENTOS

Gostaria de agradecer a todos que, de alguma forma, me ajudaram a chegar até aqui.

- Deus: Ser superior. Imensurável a quantidade de vezes que pedi ajuda e, em todas elas, o Sr. estava lá para me ouvir e estender a mão. Muito obrigado!
- Anderson e Alice: pais amados que me deram a oportunidade de vir a esse mundo. Os princípios de ética, respeito, honestidade e humildade, que sempre me ensinaram, foram fundamentais para minha caminhada até aqui. Amo vocês!
- Micheline: esposa, mãe, amiga, companheira e mulher com “M” maiúsculo. Me ensinou que paciência e persistência são premissas para a obtenção do êxito. Com você aprendi que a mulher é uma fortaleza quando comparada a um homem. Tua luta pela vida me inspirou a concluir esta tese. Obrigado por tudo. Te amo!
- Natália: filha querida e companheira inseparável. Mesmo diante de tantas ausências e dificuldades que assolaram nossas vidas, sempre se manteve forte e determinada. Os desenhos que fazia quando eu estava ausente e deixava na mesa do escritório, estarão guardados no meu coração para sempre. Obrigado por iluminar nossas vidas. Te amo minha princesinha!
- Renata: amiga e orientadora. Não tenho como traduzir em palavras a admiração, o respeito e o carinho que sinto por você. Obrigado por compartilhar comigo tua experiência acadêmica e de vida. Tua dedicação, teu conhecimento, tua visão, tua competência e teu jeito simples de mostrar as coisas foram fundamentais para o desenvolvimento desta tese. A experiência que adquiri a seu lado será levada por toda a vida. Obrigado, de coração!

- Palazzo: professor e orientador que motiva as pessoas. Sua bela trajetória acadêmica é um exemplo para todos. Seu conhecimento e visão estratégica foram suportes essenciais para o desenvolvimento desta tese. Agradeço pela oportunidade de poder trabalhar a seu lado, pois é uma referência na área de computação. Sua reputação ajudou a inspirar o tema desta tese. Muito obrigado!
- Família: irmãs, cunhadas, cunhados, sogro, sogra. Agradeço a compreensão da ausência e toda a força que deram para a finalização do trabalho.
- Colegas de doutorado: a todos os colegas e amigos que convivi ao longo do curso de doutorado, meus sinceros agradecimentos. Obrigado pelas discussões, pelos cafezinhos (fundamentais), pela parceria e também pelas críticas para o aperfeiçoamento do trabalho.
- Alunos: a todos os alunos que orientei em TCC, meu muito obrigado. Sintam-se parte desta tese, pois cada um teve contribuição em algum momento do trabalho.
- Colegas da UPF: agradeço aos colegas da Universidade de Passo Fundo, em especial à Reitoria e aos coordenadores do Conselho de Unidade do Instituto de Ciências Exatas e Geociências. Obrigado pelo apoio e pela compreensão durante os períodos de ausência. Aos colegas da área de informática, agradeço pelas discussões, dicas e ideias que contribuíram com a tese. Também pelos momentos de descontração para aliviar o estresse. Aos colegas do grupo de pesquisa Mosaico agradeço, principalmente, por terem conduzido os trabalhos de forma incansável, mesmo sabendo da minha restrita dedicação ao grupo. Agora volto com todo o gás. Me aguardem!

Por fim, gostaria de agradecer a Universidade Federal do Rio Grande do Sul, ao Instituto de Informática e ao Programa de Pós-graduação em Computação pela oportunidade de poder realizar o doutoramento em uma instituição pública de excelência e em um programa de referência nacional e internacional na área de Ciência da Computação.

SUMÁRIO

LISTA DE ABREVIATURAS E SIGLAS

ACM	Association for Computing Machinery
CAPES	Coordenação de Aperfeiçoamento de Pessoal de Nível Superior
CNPq	Conselho Nacional de Desenvolvimento Científico e Tecnológico
DBLP	Digital Bibliography & Library Project
WEKA	The Waikato Environment for Knowledge Analysis
SAC	Symposium on Applied Computing
CRF	Conditional Random Fields
OWL	Ontology Web Language

GLOSSÁRIO DE TERMOS

Métrica	A própria medida (métrica direta) ou um método de cálculo (métrica indireta) e a escala de medição (consideração no espaço métrico). Uma métrica determina padrões de medição pelos quais um determinado Indicador pode ser avaliado.

LISTA DE FIGURAS

LISTA DE TABELAS

RESUMO

A tarefa de avaliar a produção científica de um pesquisador é fortemente baseada na análise de seu currículo. É o que fazem, por exemplo, as agências de fomento à pesquisa e desenvolvimento ou comissões de avaliação, quando necessitam considerar a produção científica dos pesquisadores no processo de concessão de bolsas e auxílios, na seleção de consultores e membros de comitês, na aprovação de projetos ou simplesmente para avaliar o conceito de um programa de pós-graduação.

Nesse contexto, a modelagem de perfis de pesquisadores é tarefa fundamental, especialmente quando se quer avaliar a reputação dos pesquisadores. Isto pode ocorrer por meio de um processo de análise da trajetória de toda a carreira científica do pesquisador. Tal processo envolve não somente aspectos relacionados a artigos ou livros publicados, mas também por outros elementos inerentes à atividade de um pesquisador, como orientações de trabalhos de mestrado e de doutorado; participação em defesas de mestrado e de doutorado; trabalhos apresentados em conferências; participação em projetos de pesquisa, inserção internacional, dentre outros.

O objetivo deste trabalho é especificar um modelo de perfil de pesquisadores (Rep-Model) e uma métrica para medir reputação acadêmica (Rep-Index). O processo de modelagem do perfil envolve a definição de quais informações são relevantes para a especificação do perfil e as apresenta por meio de 18 elementos e 5 categorias. O processo para medir a reputação do pesquisador é definido por uma métrica que gera um índice. Esse índice é calculado mediante a utilização dos elementos constantes no perfil do pesquisador.

Para avaliar a abordagem proposta na tese, diversos experimentos foram realizados. Os experimentos envolveram a avaliação dos elementos do Rep-Model por meio de análise de correlação e por algoritmos de mineração de dados. O Rep-Index também foi avaliado e correlacionado com duas métricas amplamente utilizadas na comunidade científica, o h-index e o g-index. Como *baseline*, foram utilizados todos os pesquisadores do CNPq das áreas de Ciência da Computação, Economia e Odontologia.

O trabalho desenvolvido nesta tese está inserido no contexto da identificação da reputação de pesquisadores no âmbito acadêmico. A abordagem desta tese tem como premissa ser abrangente e adaptável, pois envolve a vida científica do pesquisador construída ao longo de sua carreira científica e pode ser utilizada em diferentes áreas e em diferentes contextos.

Palavras-Chave: Reputação de pesquisadores, modelo de perfil, métricas científicas, adaptabilidade.

Rep-Index – A Comprehensive and Adaptable Approach to Identify Academic Reputation

ABSTRACT

The task of evaluating the scientific production of a researcher is based strongly on the analysis of their curriculum. It's what makes the agencies for research support or evaluation committees, when they need to consider the scientific production of researchers in the process of awarding grants and aid in the selection of consultants and committee members in approving projects or simply to assess the concept of a program graduate.

In that context, the modeling of profiles of researchers is fundamental task especially when one wants to evaluate the reputation of the researchers. This can occur by means of a process of analysis of the trajectory of all the scientific career of the researcher. Such process involves not only aspects related to papers or books, but also other elements inherent in the activity of a researcher, as orientations of master's degree and doctorate; participation in defense of master's and doctoral degrees; papers presented in conferences, participation in research projects, international integration, among others.

This proposal specifies a profile template for researchers (Rep-Model) and a metric to measure academic reputation (Rep-Index). The profile modeling process involves define which information is relevant to the specification of the profile and shows through 18 elements and 5 categories. The process for measuring researcher's reputation is defined by a metric that generates an index. This index is calculated by using the information contained in the profile of the researcher.

To evaluate the approach proposed in the thesis, extensive experiments were conducted. The experiments involved the evaluation of Rep-Model by means of correlation analysis and data mining algorithms. The Rep-Index was also evaluated and correlated with two metrics widely used in the scientific community, the h-index and g-index. As a baseline, all of CNPq researchers in the areas of Computer Science, Economics and Dentistry were used.

The work in this thesis is set in the context of identifying the reputation of researchers within the academic sphere. The approach of this thesis is premised be comprehensive and adaptable, because it involves the life science researcher built throughout his scientific career and can be used in different research areas and in different contexts.

Keywords: Researcher reputation, profile model, scientific metrics, adaptability.

1 INTRODUÇÃO

Atualmente, a gestão da ciência, da tecnologia e da inovação passa por um processo de qualificação, uma vez que se busca cada vez mais conhecer os pesquisadores e suas atividades com vistas à correção de rumos em suas pesquisas, recomendação e orientação para uma correta aplicação dos recursos que são oportunizados. As agências de fomento, os centros de pesquisa e grande parte das universidades sentem a necessidade de obter informações sobre a produção científica e as atividades desenvolvidas por seus pesquisadores com vistas ao suporte na tomada de decisão. Dessa forma, as agências de fomento conseguiriam canalizar recursos para grupos de comprovada competência em áreas específicas do conhecimento, estimular o desenvolvimento da ciência e tecnologia, bem como buscar diferenciais competitivos e excelência acadêmica.

O procedimento para avaliar a produção científica de um pesquisador ainda hoje é baseado fortemente na análise de seu currículo. É o que fazem as instituições quando necessitam considerar a produção científica dos pesquisadores no processo de concessão de bolsas e auxílios (como as diversas modalidades de bolsas suportadas pelo CNPq[1]), na seleção de consultores e membros de comitês, na aprovação de projetos, na classificação de periódicos (como o *Qualis* da Capes[2]) ou simplesmente para avaliar o conceito de um programa de pós-graduação.

Nesse escopo, uma área de aplicação que tem obtido destaque mundial em computação é a colaboração científica entre pesquisadores e as métricas de avaliação associadas. Várias abordagens, métodos e sistemas têm sido propostos para modelar o perfil de um pesquisador ou avaliar seu prestígio na comunidade científica (CARMEL; JOSIFOVSKI; MAAREK, 2011; GAUCH et al., 2007; LOH et al., 2010; GASPARINI et al., 2012).

Aliado a essa busca por modelos de perfil de pesquisadores, a procura por critérios de avaliação de qualidade na academia está aumentando e tem sido o foco de estudos na última década. Esse crescimento surge devido à busca da excelência nas principais áreas de pesquisa. Ele também é motivado por outros fatores, como a concorrência para concessões de bolsas e disponibilização de recursos por agências de fomento. A alta competição em tal cenário requer que os critérios de qualidade sejam objetivamente definidos, de preferência utilizando uma abordagem que pode ser facilmente reproduzida. Esta última característica é importante, pois o resultado de qualquer

[1] Conselho Nacional de Desenvolvimento Científico e Tecnológico. Disponível em http://www.cnpq.br/

[2] Coordenação de Aperfeiçoamento de Pessoal de Nível Superior. Disponível em http://www.capes.gov.br/

estratégia de classificação pode ser questionada após a sua publicação. Por isso, é desejável permitir que qualquer pessoa possa reproduzir o procedimento a fim de checar seus resultados (LOPES et al., 2012).

A utilização das diversas abordagens, métodos e sistemas existentes se dá nas mais variadas áreas de aplicação, como por exemplo: (i) detecção de perfil de clientes em sites de vendas on-line; (ii) recomendação de produtos; (iii) recomendação de especialistas para participarem de defesas de mestrado e de doutorado; (iv) recomendação de pesquisadores para formação de rede de colaboração científica; (v) organização de comitê de programa de conferência científica; (vi) organização de corpo editorial e revisores de periódicos; (vii) medição do impacto de publicações, dentre outras.

No Brasil, a plataforma Lattes[3], apesar de ser o mais importante instrumento de armazenamento dos dados referentes à produção científica de pesquisadores brasileiros, possui um mecanismo atual de representação dos currículos que não permite consultas mais elaboradas sobre os mesmos. Dessa forma, todo o processo de análise é realizado de forma manual, demandando muito tempo, sendo extremamente cansativo e podendo levar a avaliações equivocadas, pois toda análise fica baseada na percepção humana. Essa deficiência é devido à forma como os dados são representados, o que torna inexistente a possibilidade de se fazer consultas diretas para recuperação de informações que permitam análise e cruzamento de dados.

No meio acadêmico não existe consenso ou padronização do que é necessário para se modelar o perfil de um pesquisador, e nem mesmo um perfil ideal. Atualmente, na maioria das abordagens, são consideradas quase que exclusivamente as publicações do pesquisador. A identificação de um perfil deve ocorrer por meio de um processo de análise da trajetória de toda a carreira de um pesquisador. Tal processo envolve não somente aspectos relacionados à produção científica, como artigos e livros publicados, mas também por outros elementos inerentes à atividade de um pesquisador, como por exemplo: (i) orientações de trabalhos de mestrado e de doutorado; (ii) participação em defesas de mestrado e de doutorado; (iii) trabalhos apresentados em conferências; (iv) participação em projetos de pesquisa, dentre outros. Uma proposta de modelo de perfil abrangente deve preencher ou minimizar esse *gap*, pois permitirá uma análise mais completa da trajetória científica do pesquisador.

Complementando o processo de modelagem do perfil de pesquisadores, a avaliação de sua reputação torna-se uma tarefa importante e tem sido objeto de estudo de diversos grupos na área de ciência da computação. Na última década uma série de trabalhos foram desenvolvidos e apresentados à comunidade acadêmica internacional. Todos eles envolvem a definição de métricas científicas. A área de métricas científicas, no âmbito de avaliação de pesquisadores, tem um papel importante na comunidade acadêmica, pois pode auxiliar no processo de medição da qualidade da produção científica e na identificação de especialistas em determinada área, tornando-se uma importante ferramenta para apoio à tomada de decisão.

Avaliação da Ciência, e em especial a avaliação de cientistas leva em consideração muitos aspectos, incluindo a produção, a produtividade e o impacto do trabalho do cientista. De uma forma geral as métricas de produção incluem o número total de artigos publicados, o número total de artigos indexados em bases bibliográficas, como a

[3] http://lattes.cnpq.br/

Web of Science[4] e mais alguns outros indicadores. Medidas de produtividade são as medidas de produção calculadas sobre um intervalo de tempo recente, como por exemplo, o número total de artigos publicados por um pesquisador nos últimos 5 anos. Já as medidas de impacto geralmente referem-se às citações recebidas pelos artigos publicados pelo pesquisador. Mas para Wainer e Vieira (2013) o padrão-ouro para a avaliação de pesquisadores é avaliação pelos pares, o que muitas vezes pode ser compreendido como algo não mensurável, como é o caso do prestígio do pesquisador perante à comunidade científica.

O trabalho de Krapivin et al. (2009) apresenta que a função de se utilizar métricas científicas para avaliar a produção de pesquisadores está fundamentada em dois contextos: medir a qualidade da produção científica e medir a contribuição e a reputação relacionada a pesquisadores. Nesse sentido, Small (1973), White e McCain (1998), Hirsch (2005), Egghe (2006a), Jin (2007), Jin et al. (2007), Rousseau e Jin (2008), Zhang (2009), Ye (2011), Yan, Zhai e Fan (2013), Zhai, Yan e Zhu (2013) e Zhang (2013-a) abordam temas como a citação para medir a reputação de pesquisadores.

A popularização do *PageRank* (PAGE et al., 1998) fez com que diversos trabalhos no âmbito acadêmico se utilizassem de seus algoritmos para a análise da reputação de pesquisadores. Trabalhos como os de Ding et al. (2009) e Chen et al. (2007) aplicam o *PageRank* a conjuntos de dados a fim de ordenar autores e artigos, respectivamente. O trabalho de Ding et al. (2009) ordena autores por meio de redes de cocitação. Já o trabalho de Chen et al. (2007) utiliza o conceito de *damping factor*, que se baseia no fato de que um leitor tende a procurar um caminho de ordem igual a dois links nas referências do artigo original antes de sair em busca de um artigo novo. O algoritmo *PaperRank* (KRAPIVIN et al., 2009) modifica o *PageRank* original. A alteração considera, além das citações entre artigos, o ranking do artigo citante e a densidade de citações feitas pelo artigo citante.

Antes das recentes métricas científicas serem apresentadas à comunidade científica, Makino (1998) publicou um estudo em importante periódico de abrangência internacional desenvolvido com dois grupos de pesquisa de astrofísicos teóricos japoneses. O trabalho envolveu a análise da reputação dos grupos estudados no âmbito da quantidade de publicações e média de citações a estas publicações. Os dois grupos apresentaram resultados semelhantes para esses índices macroscópicos. O resultado sugeriu que medidas quantitativas oriundas da produtividade não são medidas significativas para a contribuição real de um grupo de pesquisa para a Ciência.

Recentemente, uma nova métrica com um importante componente avaliado pelos pares foi desenvolvida e publicada pelo *ResearchGate*[5]. Em uma visão ampla, este processo de avaliação envolve não apenas os aspectos científicos, tais como periódicos e livros, mas também outros fatores inerentes às atividades de um pesquisador, como participação em bancas e conselhos, orientações, palestras e colaboração científica com outros pesquisadores. Além desses, outros elementos são utilizados atualmente. A ACM[6] (*Association for Computer Machinery*) reconhece a excelência por meio de uma série eminente de prêmios por conquistas e contribuições técnicas e profissionais de destaque nas áreas de Ciência da Computação e Tecnologia da Informação.

[4] http://thomsonreuters.com/web-of-science/
[5] https://www.researchgate.net
[6] http://www.acm.org/

Apesar do esforço da comunidade acadêmica em desenvolver estratégias para identificar a reputação de pesquisadores definindo métricas científicas, a grande maioria dos trabalhos fundamenta sua contribuição nas citações dos artigos dos pesquisadores. Entretanto, essas formas de avaliação não consideram a trajetória do pesquisador, nem mesmo o cenário onde ele está inserido. As citações a artigos deveriam ser apenas um dos elementos a ser analisado e não o único elemento.

Uma métrica que engloba toda a vida de um pesquisador e também fornece uma análise detalhada é necessária para alcançar uma avaliação justa. Essa solução pode envolver vários pontos de vista, quando os pesquisadores estão trabalhando em diferentes áreas de pesquisa. Uma abordagem mais ampla, envolvendo vários elementos da carreira de um pesquisador, aliada à possibilidade de adaptação dessa abordagem para contemplar a diversidade de áreas de pesquisa que pode ser utilizada em diferentes contextos, seria o ideal. Com uma abordagem assim, o problema de avaliar pesquisadores comparando todos eles e usando os mesmos critérios e os mesmos pesos seria vencido. Pesquisadores de áreas diferentes possuem características diferentes, sendo assim, os critérios de avaliação e suas ponderações devem ser diferentes.

Nesse sentido, Bollen et al. (2009) apresenta uma avaliação de diferentes classificações usando 39 indicadores de impacto e conclui que o conceito de impacto científico é multidimensional e não pode ser medida usando apenas um indicador. Da mesma forma, Lima et al. (2013) afirma que apenas a utilização de indicadores bibliométricos para avaliar pesquisadores proporciona uma análise relativa que precisa ser adaptada ou combinada com outros indicadores. Ainda, em muitos casos, é necessário fornecer informações adicionais para garantir as especificidades da avaliação desejada.

O trabalho desenvolvido nesta tese está inserido no contexto da identificação da reputação de pesquisadores no âmbito acadêmico. Para isso, propõe-se um modelo de perfil como eixo estruturante para o cálculo da reputação. Contrariamente aos trabalhos que fundamentaram esta tese, essencialmente voltados a dados bibliométricos de citações, o presente trabalho considera a estatística de citações como um dos indicadores da abordagem no momento da definição do perfil do pesquisador e não o único indicador. Com essa premissa, busca-se identificar outros elementos da trajetória científica de um pesquisador que são relevantes para estruturar seu perfil e medir sua reputação.

Na próxima seção são apresentados os objetivos da tese bem como as contribuições da mesma no âmbito de modelagem de perfis e métricas para identificação de reputação.

1.1 Objetivos e Contribuições

O objetivo desta tese é especificar um modelo de perfil de pesquisadores denominado Rep-Model e uma métrica para medir reputação acadêmica denominada Rep-Index. A modelagem do perfil envolve a identificação das informações relevantes da carreira do pesquisador. Tais informações são modeladas em categorias e elementos. As categorias, assim como os elementos, foram estruturados para suportarem ponderação por meio da definição de pesos. Já a reputação do pesquisador é definida por uma métrica que gera um índice calculado levando-se em consideração os elementos constantes no perfil do pesquisador que representam a produção científica desse pesquisador.

A tese tem como premissa ser abrangente e adaptável, pois engloba a vida científica do pesquisador construída ao longo de sua carreira científica. Tais premissas permitem a utilização da abordagem em diferentes áreas e em diferentes contextos, uma vez que áreas diferentes usam critérios diferentes. Dessa forma, dependendo do propósito de utilização, ou seja, o que se quer avaliar, o que se quer medir e para que se quer medir, os critérios podem ser adaptados para que contemplem os requisitos do usuário e suas ponderações.

Essas características não foram encontradas em outras abordagens, pois não foi localizado na literatura um modelo abrangente que permitisse modelar o perfil de um pesquisador de forma a considerar toda sua produção científica, nem métricas que identificassem a reputação desse pesquisador levando-se em consideração os elementos modelados em seu perfil.

A questão de pesquisa a ser avaliada na tese é de que a reputação de um pesquisador se dá pelo equilíbrio de sua produção científica desenvolvida ao longo de sua carreira. A hipótese é formulada considerando a premissa de que quanto mais tempo um pesquisador atuar e mais equilibrada for sua produção, maior será sua reputação.

Acredita-se que a questão de pesquisa elaborada para a tese representa a grande maioria dos pesquisadores das mais diferentes áreas do conhecimento. Mas como em todas as populações existem exceções, e no âmbito de identificação de reputação acadêmica não deve ser diferente, podem existir pesquisadores que se distanciam dessa questão de pesquisa. Como exemplo, pode-se citar o caso de pesquisadores iniciantes que fazem uma descoberta que causa grande impacto na comunidade científica ou que desenvolvem um produto ou processo que trará grandes benefícios à vida das pessoas. Isso não justifica que estes pesquisadores tenham reputação elevada, pois sua trajetória ainda é incipiente.

Para validar o modelo de perfil de pesquisadores e a métrica para identificar reputação acadêmica, diversos experimentos foram realizados envolvendo análise de trajetória científica. Nos experimentos, dados dos bolsistas de produtividade em pesquisa e tecnologia do CNPq foram utilizados. Dentro desse espectro, como o volume de dados envolvidos nas bases do CNPq é muito elevado, definiu-se o escopo da avaliação experimental de modo a suportar estatisticamente os resultados obtidos. Assim, foram realizados experimentos com bolsistas do CNPq de três áreas de pesquisa envolvendo pesquisadores dos níveis 1A, 1B, 1C, 1D e 2.

A avaliação experimental foi estruturada para que apresentasse a correlação entre a abordagem proposta na tese, de identificar a reputação de pesquisadores usando o modelo de perfil como premissa, e duas métricas conhecidas e consolidadas na comunidade acadêmica internacional, o h-index e o g-index. Para isso, o resultado da abordagem foi comparado com o resultado do h-index e do g-index em todos os experimentos realizados. Para analisar a existência de correlação entre os índices utilizou-se o Coeficiente de Correlação de Postos de Spearman (ρ). O Coeficiente de Correlação de Spearman avalia como a relação entre duas variáveis pode ser descrita. Foi utilizado esse método devido à heterogeneidade dos pesquisadores, pois apesar de representarem um grupo de excelência no país, existe muita variação de produção entre eles, o que faz com que método estatístico mais apropriado seja o de Spearman. O método também é menos sensível a *outliers*, em que pesquisadores se distanciam muito dos demais, em nível de produção científica. Aliado a isso, o Coeficiente de Spearman tem sido utilizado em diversos trabalhos envolvendo avaliação ou ranking de

pesquisadores, como apresentado em Wainer e Vieira (2013), Lopes et al. (2011), Franceschet e Costantini (2011), Waltman et al. (2011), Li et al. (2010), Korn, Schubert e Telcs (2009), Van Raan (2006) e Rinia et al. (1998).

A abordagem descrita nesta tese apresenta as seguintes contribuições: (i) especificação do modelo de perfil de pesquisadores (Rep-Model); e (ii) métrica para identificar reputação acadêmica (Rep-Index). Tanto para modelagem do perfil como para identificação de reputação, a abordagem desta tese se difere da literatura atual por especificar um modelo de perfil de pesquisadores abrangente e adaptável, que pode ser utilizado em diferentes áreas e em diferentes contextos. A adaptabilidade do modelo envolve desde a inserção e a eliminação de categorias e elementos, bem como o ajuste de seus pesos. Outra questão relevante que a tese visa suprir é que o modelo e a métrica se propõem a identificar a reputação de forma abrangente, buscando um equilíbrio entre diferentes elementos que fazem parte da trajetória científica de um pesquisador ao invés de levar em consideração apenas citações bibliográficas, como ocorre em outras abordagens, no caso do h-index e do g-index, por exemplo. Ainda, a abordagem estruturada na tese possibilita a avaliação de pesquisadores de forma individual e de grupos. Isso mostra a flexibilidade da proposta frente a diversas possibilidades de utilização.

Vislumbra-se como possíveis aplicações decorrentes dos resultados da tese por meio do modelo de perfil de pesquisadores abrangente e adaptável, bem como da métrica para identificar reputação acadêmica, as seguintes utilizações pela comunidade cientifica: (i) análise de redes de colaboração científica de citação e de coautoria; (ii) identificação de similaridade de perfis entre pesquisadores; (iii) uso em sistemas de recomendação baseados no perfil dos pesquisadores; (iv) utilização por agências de fomento, universidades e instituições de pesquisa para auxiliar no processo de tomada de decisão; (v) avaliação individual de pesquisadores, como bolsistas de produtividade em pesquisa e tecnologia do CNPq; (vi) avaliação de grupos de pesquisadores, como programas de pós-graduação recomendados pela Capes; dentre outras formas de utilização.

1.2 Organização do Texto

O restante do texto está organizado como segue.

- No Capítulo 2 são apresentados os principais conceitos, definições e posicionamento dos temas que fundamentaram esta tese no que tange a revisão bibliográfica. Os assuntos expostos são concernentes às áreas de modelagem de perfil de usuários e métricas científicas para identificar reputação de pessoas no contexto acadêmico, sendo amplamente aceitos pela comunidade científica internacional.

- No Capítulo 3 são apresentados em detalhes os trabalhos que mais possuem relação com os temas discutidos na tese e sua comparação entre si, evidenciando os pontos em aberto que buscou-se suprir com a tese. Primeiro, apresenta-se os trabalhos relacionados no âmbito de modelagem de perfis de usuários. Após, detalha-se as métricas científicas para medir reputação acadêmica. Na sequência, mostra-se um comparativo entre os trabalhos relacionados envolvendo alguns critérios de comparação. Por fim, apresenta-se o enquadramento da tese em comparação aos trabalhos relacionados, evidenciando o que está em aberto.

- No Capítulo 4 é apresentada a abordagem proposta na tese. Na primeira seção é apresentada uma visão geral da tese. Na sequência é especificado o modelo de perfil de pesquisadores. Na terceira seção especifica-se a métrica para identificar reputação acadêmica com base no modelo de perfil.
- No Capítulo 5 apresenta-se a avaliação experimental da tese, composta por experimentos que envolveram análise da reputação de pesquisadores. Também neste capítulo são apresentados os resultados dos experimentos e as análises que fundamentaram a tese.
- No Capítulo 6 são apresentadas as conclusões referentes ao trabalho desenvolvido na tese, destacando os resultados obtidos e as contribuições alcançadas. Também são apresentados alguns indicativos de trabalhos futuros, bem como as publicações e orientações proporcionadas pelo desenvolvimento da tese.

2 FUNDAMENTAÇÃO TEÓRICA

Este capítulo tem por objetivo definir os principais conceitos que envolvem as áreas de modelagem de perfil de usuários e métricas científicas para identificar reputação de pessoas no contexto acadêmico, com vistas a fundamentar a tese.

2.1 Modelagem de Perfis

A área de modelagem de usuários teve seu início no final da década de 1970, sendo amplamente apresentada no trabalho de Elaine Rich (RICH, 1979). Por um tempo após as pesquisas iniciais, diversos sistemas foram desenvolvidos levando-se em consideração os interesses dos usuários. Muitos destes sistemas davam ênfase em adaptações de acordo com o contexto e com o perfil dos usuários (KOBSA; WAHLSTER, 1989; MCTEAR, 1993). Com o passar do tempo, para dar suporte às adaptações que eram necessárias nos sistemas, passou-se a adotar a modelagem de perfis de usuários como base para a adaptabilidade. Assim, a área de modelagem de perfil de usuários teve como premissa a descoberta de conhecimento sobre o usuário e de que forma representar esse conhecimento (TRAJKOVA; GAUCH, 2003).

A busca por informações para descoberta de conhecimento sobre o usuário pode ocorrer de forma explícita, implícita ou adotando-se ambas abordagens. A modelagem explícita visa modelar o perfil de uma pessoa utilizando-se da interação direta com o usuário. Um exemplo pode ser por meio de respostas a um questionário aplicado ao usuário, de forma a guiá-lo a fornecer informações pessoais sobre o contexto que se deseja definir o seu perfil. A modelagem implícita tem como base a análise do comportamento da pessoa. Como exemplo, pode-se citar a navegação do usuário em páginas web, suas compras on-line, publicações científicas e número de referências a publicações. A adoção das duas abordagens em conjunto, também denominada modelagem híbrida, tem como propósito modelar o perfil de um usuário dependendo do domínio de aplicação e do próprio usuário, uma vez que tem como premissa que o sistema deve ter a possibilidade de se adaptar diante do cenário de utilização.

A modelagem do perfil de um usuário pode ser baseada em seu conhecimento (*knowledge-based*) ou em seu comportamento (*behavior-based*) (MIDDLETON; DE ROURE; SHADBOLT, 2004). Na modelagem baseada em conhecimento, geralmente os usuários são associados a modelos estáticos de usuários e o processo de definição do perfil é guiado por entrevistas ou questionários. Nesse caso, o usuário deve participar de forma ativa do processo de modelagem, uma vez que as informações relevantes para a definição do perfil devem ser fornecidas por ele. Já na modelagem baseada em comportamento, parte-se do princípio que o próprio comportamento do usuário define o modelo. Dessa forma ele não necessita participar do processo de modelagem de forma explícita, pois a definição do perfil pode ser dar por meio de técnicas de mineração de

dados ou aprendizagem de máquina para descobrir padrões úteis de comportamento acerca do usuário.

Alguns autores abordam questões que vão além da definição do perfil. Montaner (2003) apresenta que além da geração do perfil é necessário pensar em atualização desse perfil. Em seu trabalho especifica cinco etapas para a modelagem do perfil: (i) uma técnica de representação do perfil; (ii) uma técnica usada para gerar o perfil inicial; (iii) *relevance feedback* que represente os interesses do usuário; (iv) uma técnica de aprendizagem de perfil; e (v) uma técnica de adaptação de perfil.

Para Trajkova e Gauch (2003) é necessário construir um modelo de usuário que represente exatamente os seus interesses, independente de ambiente, e que possua três objetivos principais:

- Descobrir o conhecimento ou interesse de uma pessoa em determinado assunto;
- Representar e armazenar este conhecimento ou interesse internamente em um sistema;
- Gerenciar possíveis alterações no conhecimento ou no interesse.

O perfil do usuário representado em Trajkova e Gauch (2003) se utiliza de uma ontologia, onde é definido o peso total e o número de páginas associados com cada conceito existente na ontologia. A entrada principal desse sistema para construir um perfil são páginas da Web que um usuário permaneceu por pelo menos 5 segundos. O sistema classifica cada página pelo conceito mais semelhante em uma hierarquia predefinida de conceitos na ontologia. Nesse sistema o processo de construir o perfil consiste em três fases:

- Treinar o classificador;
- Coletar dados do usuário;
- Classificar as páginas da Web.

Levando-se em consideração que os interesses de uma pessoa tendem a ser dinâmicos, ou seja, podem mudar com o tempo, a representação do perfil pode refletir possíveis mudanças, uma vez que é necessário representar com o máximo de exatidão os interesses e preferências do usuário. Moukas (1997) define que um sistema de controle de perfil deve aprender os interesses do usuário de maneira implícita, ou seja, sem intervenção do mesmo. Dessa forma, apresenta um sistema que tem abordagem baseada na utilização de métodos de inteligência artificial para gerar e manter perfis de usuários. A construção do perfil se dá por meio de uma aplicação multiagente, denominada “Amalthaea”, que observa a interação atual e passada do usuário com o sistema e também o *feedback* do usuário para atualizar seu estado e melhor representar seus interesses. Amalthaea fornece a seus usuários filtragem personalizada e descoberta de informação. Seu ambiente é Web, onde visa assistir usuários a encontrarem informação de seus interesses.

Métodos para criação e atualização de perfis de usuário também são discutidos em Kuflik e Shoval (2000). Os autores sugerem especificar o perfil conforme descrito a seguir:

- Perfil criado pelo usuário: o usuário define a área de interesse utilizando termos e atribui pesos a cada um destes termos;

- Perfil criado pelo sistema por meio de definição automática: dados que o usuário julga como relevantes são analisados por uma ferramenta que extrai os termos que possuem maior frequência. Os termos recebem pesos, de acordo com a frequência nas quais aparecem, e são utilizados para definir o perfil;
- Perfil criado pelo usuário e pelo sistema: é a combinação das técnicas apresentadas anteriormente. Primeiramente, o perfil é definido automaticamente pelo sistema e, logo após, o usuário revisa os termos extraídos, bem como os pesos atribuídos a estes;
- Estereótipos: assume que o sistema já possui perfis armazenados. O estereótipo é representado como um perfil baseado em conteúdo, por meio de uma lista de usuários que tenham informações em comum e o mesmo comportamento.

O perfil de usuários, assim como o perfil de pesquisadores, pode ser representado de diversas maneiras. Uma das formas mais utilizadas é por meio de ontologias, como em Schoefegger (2011), Sriharee (2010), Punnarut e Sriharee (2010), Tang et al. (2008b), Tang, Zhang e Yao (2007), Zhang, Song e Song (2007), Yao, Tang e Li (2007), Trajkova e Gauch (2004), Li e Zhong (2003), Razmerita, Angehrn e Maedche (2003), Kurki et al. (1999) e Pretschner e Gauch (1999). Uma ontologia é uma especificação explícita de uma conceitualização (GRUBER, 1995). Ela define os termos usados para descrever e representar uma área do conhecimento. As ontologias são usadas por pessoas e aplicações para a troca de informações sobre um determinado domínio. Uma ontologia fornece definições de conceitos básicos de um domínio, apropriadas para o processamento automático.

Outra forma de representação de perfil de usuários é por meio de vetores de termos, como apresentado em Sugiyama e Kan (2010), Sugiyama e Kan (2011), Kim et al. (2008), Widyantoro et al. (1999), Chen e Sycara (1998), Joachims, Freitag e Mitchell (1997), Mladenic (1996), Pazzani, Muramatsu e Billsus (1996), Lang (1995), Yan e Garcia-Molina (1995) e Sheth (1994). Um vetor de termos geralmente é definido por palavras-chave que representam as características do usuário para a composição de seu perfil. A utilização de vetores de termos no âmbito de modelagem de perfil de pesquisadores tem sido proposta essencialmente para a descoberta de especialistas em determinada área do conhecimento. Um dos métodos utilizados é a extração das palavras-chave das publicações do pesquisador para a estruturação do perfil. Em alguns trabalhos, como em Moukas e Maes (1996), Stefani e Strapparava (1998), Chen e Sycara (1998), Mladenic (1996) e Chen at al. (2008) são utilizados pesos para a definição do perfil. Os pesos flexibilizam a elaboração do perfil, pois permitem adaptabilidade diante de critérios que sejam úteis para os objetivos do sistema ou do próprio usuário.

2.2 Métricas para Reputação Acadêmica

Métricas para reputação têm um papel importante na comunidade acadêmica, pois podem auxiliar no processo de medição da qualidade da produção científica e na identificação de especialistas em determinada área. A utilização de métricas científicas para avaliar a ciência teve seu início com Eugene Garfield. Ele apresentou um índice de citação unificada para a literatura da ciência em 1955. Seu principal objetivo com a

proposta era que as referências de artigos científicos deveriam ser utilizadas como indexação, pois isso poderia funcionar como uma ferramenta de recuperação de informação relevante (GARFIELD, 1955). Assim, as citações poderiam ser vistas como vínculos formais e explícitas entre os trabalhos dos pesquisadores.

As métricas recentes se baseiam fortemente nas citações de artigos de pesquisadores, conforme proposto por Garfield. Entretanto, essa forma de avaliação não considera a trajetória do pesquisador, nem mesmo o cenário onde ele está inserido. As citações das publicações deveriam ser apenas um dos elementos a serem analisados e não o elemento exclusivo quando se avalia a reputação de pesquisadores.

No contexto acadêmico, a identificação de um perfil pode ocorrer por meio de um processo de análise da trajetória de toda a carreira de um pesquisador. Tal processo envolve não somente aspectos relacionados a produção científica, como artigos e livros publicados, mas também por outros elementos inerentes à atividade de um pesquisador, como por exemplo: orientações de trabalhos de mestrado e de doutorado; participação em defesas de mestrado e de doutorado; trabalhos apresentados em conferências; participação em projetos de pesquisa, dentre outros.

O trabalho de Krapivin, Marchese e Cassati (2009) argumenta que a função de se utilizar métricas para avaliar a produção de pesquisadores está fundamentada em dois contextos: medir a qualidade da produção científica e medir a contribuição e a reputação relacionada a pesquisadores. Nesse sentido, os trabalhos de Small (1973), White e McCain (1998), Hirsch (2005), Jin (2007) e Jin et al. (2007) abordam temas como a citação para medir a reputação.

Outras métricas também foram desenvolvidas, como o h-index (HIRSCH, 2005), o g-index (EGGHE, 2006a), o AR-index (JIN, 2007; JIN et al., 2007), o e-index (ZHANG, 2009), o Hg-index (ALONSO et al., 2010), o *h*'-Index (ZHANG, 2013-b), o H_I-index (ZHAI; YAN; ZHU, 2013) e o ca-index (LIMA et al. 2013), todos calculando a importância individual de pesquisadores. Estas métricas foram originadas a partir do h-index (HIRSCH, 2005). Elas visam complementar ou melhorar o h-index em relação a alguns fatores, como a idade da publicação, a inclusão de publicações que tiveram baixas citações, ou ainda a possibilidade de um pesquisador ter seu h-index diminuído com o passar do tempo. Estas métricas são apresentadas e discutidas em detalhes na seção 3.2.

Além de métricas que se baseiam essencialmente nas citações de publicações, novas métricas foram desenvolvidas. Algumas delas agregam um mecanismo de busca para facilitar a identificação da reputação e permitir a visualização de informações acerca de pesquisadores. Nesse contexto algumas propostas ganham destaque, como é o caso do Google Scholar[7], do ArnetMiner[8] (TANG et al., 2008b) e do ResearchGate[9].

O Scholar Google, proposto em 2004, é um mecanismo de busca que disponibiliza publicações científicas e trabalhos acadêmicos. A classificação dos trabalhos é por ordem de relevância, levando-se em consideração o autor e o impacto das citações pela comunidade científica. Uma das vantagens do Google Scholar é a qualidade dos dados que são disponibilizados, proporcionando maior precisão na resposta. Além da disponibilização dos trabalhos acadêmicos e de pesquisa ele mantém um ambiente com

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

[8] http://arnetminer.com

[9] https://www.researchgate.net/

o perfil individual de pesquisadores. Esse ambiente, proporciona a visualização do h-index de cada pesquisador, o número de suas publicações, bem como um novo índice, denominado Índice i10, que representa o número de publicações do pesquisador com, no mínimo, 10 citações. Além disso, apresenta uma segunda opção de visualização da métrica i10, com o número de publicações que receberam, pelo menos, 10 novas citações nos últimos cinco anos. O trabalho de Noruzi (2005) apresenta em detalhes o Google Scholar, inclusive aponta sugestões de melhorias que podem ser incorporadas no mecanismo, sendo algumas delas já implementadas.

Semelhante ao *Google Scholar*, um sistema denominado *Microsoft Academic Search*[10] foi proposto pela Microsoft. Ele disponibiliza um ambiente onde os pesquisadores podem visualizar seu perfil, bem como um mecanismo de busca para publicações e citações. Além disso, é possível visualizar diversas informações sobre a trajetória científica dos pesquisadores, como gráficos que apresentam o número de publicações da trajetória dos pesquisadores, suas áreas de publicação, o número de publicações e citações, o número de coautores, bem como o número de autores que já citaram suas publicações. Apesar de ser um mecanismo importante para visibilidade de informações acerca dos pesquisadores, o *Microsoft Academic Search* não disponibiliza mais o valor do h-index e o g-index desde janeiro de 2013.

Um sistema que vem sendo utilizado pela comunidade científica e que disponibiliza informações sobre a produção de pesquisadores é *Publish or Perish* (HARZING, 2007). Ele é um software que permite a busca, em tempo real, do h-index, do g-index, do e-index e do AR-index de pesquisadores, bem como apresenta outras informações, como o título das publicações, o ano, os autores, as citações e outros elementos que possibilitam a geração de estatísticas interessantes sobre a trajetória de pesquisadores. O *Publish or Perish* ainda permite ao usuário escolher a fonte dos dados para realizar a busca, se usando o *Google Scholar* ou o *Microsoft Academic Search*.

O ArnetMiner e o ResearchGate, assim como outros trabalhos mais diretamente relacionados com a tese, são apresentados em detalhes na seção 3.2.

2.3 Reputação Acadêmica

Na área de Ciência da Computação, o tema reputação vem ganhando destaque no campo da pesquisa aplicada. A utilização desse conceito pode se dar em diversos temas que necessitam considerar a reputação como fundamento para o objetivo da aplicação. Diversos estudos envolvendo sistemas, métodos e técnicas tem sido utilizados para fornecer suporte a diferentes aplicações. Dada a relevância do tema, uma das conferências científicas mais importantes na área de Ciência da Computação, a ACM SAC[11] (Symposium on Applied Computing), mantém uma trilha especial, desde 2005, chamada “Trust, Reputation, Evidence and other Collaboration Know-how”.

Reputação pode relacionar-se a um indivíduo ou a um grupo. A reputação de um grupo pode ser modelada como a média de todas as reputações individuais dos membros do grupo. Pode, ainda, ser definida como a média de como o grupo é percebido, como um todo, por terceiros (JØSANG; ISMAIL; BOYD, 2007). No caso de fazer parte de um grupo, um indivíduo pode herdar a reputação do grupo (TADELIS, 2003).

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

[11] http://www.acm.org/conferences/sac/

O conceito de reputação está relacionado com o conceito de confiabilidade (*trustworthiness*). Abaixo apresenta-se alguns conceitos de reputação, de acordo com alguns dicionários conceituados:

- Dicionário Oxford[12] (Oxford Dictionaries): Crenças ou opiniões que geralmente são realizadas sobre alguém.
- Dicionário Houaiss[13]: Renome, estima, fama, prestígio.
- Dicionário Aurélio[14]: Fama, celebridade, renome.

Os três dicionários apresentam reputação como sendo algo inerente à pessoa. Sua conduta, sua trajetória, seu prestígio, seu conceito junto à sociedade ou junto a seus pares. Dessa forma, assume-se, no contexto desta tese, que reputação acadêmica é o conceito atribuído a uma pessoa sobre o nível de conhecimento desenvolvido ao longo de sua carreira. Assim, entende-se que a reputação acadêmica de um pesquisador é diretamente proporcional ao tempo de atuação na carreira.

Nesta tese, entende-se que a melhor representação da reputação acadêmica de um pesquisador é o equilíbrio de sua trajetória científica, representada por sua produção, desenvolvida ao longo de sua carreira. Dessa forma, quanto mais tempo um pesquisador atuar e mais equilibrada for sua produção, maior será sua reputação. Isso se justifica pelo fato de que uma carreira científica é construída levando-se em consideração diversos elementos. Na abordagem definida na tese, apresentada em detalhes no capítulo 4, definiu-se 18 elementos para compor o modelo de perfil de um pesquisador. Esses elementos são o eixo estruturante para a definição da métrica que identifica a reputação do pesquisador.

---

[12] http://oxforddictionaries.com/
[13] http://www.iah.com.br/
[14] http://www.dicionariodoaurelio.com/

3 TRABALHOS RELACIONADOS

Este capítulo apresenta os trabalhos analisados que possuem relação com os temas abordados na tese e os compara entre si, evidenciando os pontos em aberto de cada um deles. O capítulo é finalizado com uma análise crítica e comparativa dos trabalhos relacionados com a abordagem proposta nesta tese com vistas a posicionar a abordagem proposta frente à literatura atual.

Os trabalhos apresentados nas seções deste capítulo estão relacionados com modelagem de perfis e métricas científicas. No âmbito de modelagem de perfis é dada ênfase na especificação de perfis de pesquisadores no contexto acadêmico. Em relação a métricas científicas aborda-se o tema enfatizando a determinação da reputação de pesquisadores por meio de análise de diversos índices que possuem afinidade com nossa proposta. A sequência de apresentação dos trabalhos relacionados segue a ordem cronológica decrescente de publicação.

3.1 Modelagem de Perfis

Esta seção tem por objetivo apresentar os trabalhos mais relevantes que possuem relação com a tese no âmbito de modelagem de perfis. A escolha dos trabalhos foi motivada pela forma como são definidos e representados os perfis, independente da aplicabilidade da proposta.

3.1.1 Uma Abordagem de Modelagem de Usuários para Apoiar o Trabalho de Conhecimento em Sistemas Sócio Computacionais

O trabalho de Schoefegger (2011) visa modelar perfis de pesquisadores para identificar habilidades ou especialidades. Para modelar o perfil, utiliza-se uma ontologia com termos relacionados ao conhecimento do pesquisador. Os termos são associados por meio de um método de agrupamento utilizando *tags*, que realiza a associação dos conceitos aos pesquisadores. O trabalho utiliza a navegação dos usuários em um sistema para realizar associação de interesses. Os eventos realizados em cada um dos termos de um agrupamento são considerados como sendo pertencentes à mesma *tag*. Um evento de usuário no sistema é modelado como *u*, *R*, *E*, *et* e *dt*, com base em cinco entidades diferentes: um usuário *u*; um recurso *r*; um tipo de evento *e*; um tópico de evento *et*; um *timestamp dt*. Isto permite gerar um modelo de rede de eventos do usuário, utilizado para análise estatística e mineração de dados.

Uma das técnicas apresentadas é a avaliação e a aplicação de abordagens teóricas dos processos de aprendizagem humana para modelar os níveis de conhecimento dos usuários para os diferentes tópicos. Isso é realizado por meio de mapeamento entre o conhecimento real dos usuários e os níveis de conhecimento como implementados no

sistema. Este mapeamento é baseado nas interações do usuário com o sistema, como por exemplo, a generalidade das buscas que são disparadas pelo usuário ou anotações realizadas nas *tags*.

A abordagem proposta pode ser utilizada em diferentes situações, como por exemplo, na confiabilidade das informações que são geradas por funcionários de uma empresa ou na personalização de buscas em sistemas de recomendação baseados no conhecimento do usuário. Uma das principais contribuições é que a abordagem diferencia usuários novatos de especialistas, pois a necessidade de informação e o comportamento dos indivíduos diferem entre novatos e especialistas ao utilizar o sistema ou executar tarefas.

Para validar a proposta, o trabalho apresenta um experimento com poucos dados, realizado com alunos, utilizando um sistema de *bookmarking* para coleta e compartilhamento de fontes de informação. Não há evidências na proposta de como são coletadas as informações dos usuários, se de forma implícita ou de forma explícita. Ainda, a diferenciação entre usuário novatos e especialistas não é explorada, pois não são apresentados os critérios para a definição do perfil, nem mesmo seus elementos.

3.1.2 Recomendação de Artigos Acadêmicos Via Interesses de Pesquisas Recentes do Usuário

A proposta de Sugiyama e Kan (2010) visa a recomendação de trabalhos acadêmicos a pesquisadores utilizando-se de artigos dos próprios pesquisadores. O conceito é de que os trabalhos publicados em determinada área são indicativos de interesse do pesquisador, assim como a citação de artigos também influencia o interesse naquela área. O trabalho diferencia pesquisadores iniciantes de experientes tendo em vista o número de publicação dos autores.

A modelagem do perfil se dá pela utilização das publicações e citações recentes do pesquisador. Para a construção do perfil são utilizados vetores de termos associados ao pesquisador, oriundos de suas publicações e de suas citações. São atribuídos pesos diferentes para artigos recentes e passados, onde os recentes são mais valorizados.

Para a definição do perfil de cada pesquisador é utilizada uma lista de suas publicações anteriores e, então, se recomenda artigos comparando os perfis com o conteúdo dos artigos candidatos. A abordagem inclui evidência contextual sobre cada artigo na forma de seus artigos próximos, ou seja, os artigos que citam o artigo-alvo e artigos referenciados pelo artigo-alvo. Os passos da abordagem são apresentados a seguir:

- define-se o perfil do usuário de uma lista de trabalhos publicados por um pesquisador;
- define-se vetores de características para artigos que são candidatos a serem recomendados;
- calcula-se a similaridade do cosseno entre os elementos do perfil do usuário com as características dos artigos representadas nos vetores de termos (para recomendar artigos com alta similaridade).

Uma das contribuições da proposta é que na modelagem do perfil dos pesquisadores são criadas duas categorias: pesquisador júnior e pesquisador sênior. Isto é importante devido aos dois tipos de pesquisadores possuírem características diferentes do ponto de

vista de produção acadêmica. O pesquisador júnior é definido como aquele que possui apenas um artigo publicado recentemente e que ainda não possui citações. O pesquisador sênior é definido como aquele que possui várias publicações e pode ter citações.

Fica evidenciado que pesquisadores iniciantes, ou seja, que possuem poucos artigos publicados ou citados, não possuem dados históricos para a definição do perfil. Dependendo do propósito de utilização isto pode ser um problema, uma vez que o critério para considerar uma pessoa como sendo pesquisador é possuir apenas um artigo publicado. Essa característica demonstra que a proposta não atende o perfil de pesquisadores de forma abrangente, pois considera apenas o elemento publicação de artigo na definição do perfil.

3.1.3 Recomendação Inesperada para Artigos Acadêmicos Considerando Relações Entre Pesquisadores

A proposta de Sugiyama e Kan (2011) é uma evolução trabalhos prévios dos autores (SUGIYAMA; KAN, 2010), apresentada na seção anterior. O diferencial da abordagem é fornecer recomendação inesperada a usuários quando estes estão fazendo uma busca utilizando usuários diferentes, ou seja, usuários que não fazem parte da mesma rede social acadêmica. O trabalho envolve os seguintes passos:

- construção de um perfil básico para cada pesquisador que se deseja gerar recomendações. Usando o perfil básico define-se o perfil de recomendação inesperada;
- define-se vetores de características para artigos que são candidatos a serem recomendados usando citações de artigos e referências de artigos-alvo;
- calcula-se a similaridade do cosseno do perfil de recomendação inesperada com os termos dos vetores de características para gerar as recomendações, retornando em ordem de alta similaridade.

Na abordagem, mesmo que pesquisadores tenham interesses diferentes, é possível gerar recomendações interessantes e surpreendentes. Assim, a abordagem utiliza perfis de usuários que são diferentes, aplicando o inverso da similaridade entre o usuário-alvo e outro usuário no qual será baseada a recomendação.

A definição do perfil é realizada sem a intervenção do usuário. Aspectos temporais são levados em consideração para a distinção de um pesquisador júnior e de um sênior. Na proposta, o pesquisador sênior possui uma ponderação maior devido ao número de artigos publicados. O pesquisador júnior é assim classificado quando tem apenas um artigo e o sênior quando tem mais de um.

3.1.4 Um Sistema de Busca de Pesquisadores Especialistas Usando Mineração de Dados Baseada em Ontologia

O trabalho de Punnarut e Sriharee (2010) propõe uma abordagem para descobrir a especialidade de pesquisadores baseada em uma ontologia que modela o conhecimento do pesquisador. A ontologia representa as relações semânticas por meio de termos extraídos de títulos de artigos e projetos dos pesquisadores. O processo de modelagem do perfil envolve a coleta de dados da web referente aos artigos e projetos dos pesquisadores, a extração e tratamento destes dados para a definição da ontologia e a

associação dos termos extraídos dos artigos e dos projetos para a definição da especialidade do pesquisador.

A Figura 3.1 representa o processo de desenvolvimento do sistema de busca de especialistas.

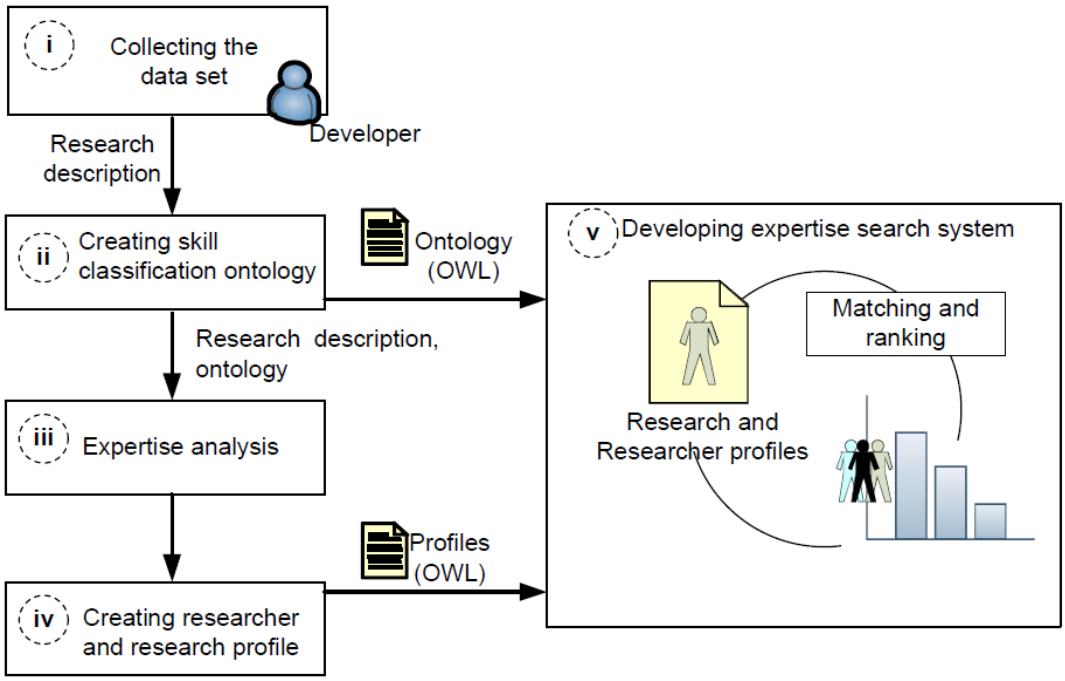


Figura 3.1: O processo para identificar especialistas (PUNNARUT; SRIHAREE, 2010).

Há várias etapas envolvidas, conforme especificado a seguir:

1. *Coleta do conjunto de dados*: o conjunto de dados consiste na descrição dos artigos e projetos de pesquisa.
2. *Criação da ontologia para classificação de conhecimentos*: a ontologia utiliza o Sistema de Classificação da Computação da ACM[15] (Association for Computing Machinery) e é utilizada para encontrar especialistas.
3. *Análise da especialidade do pesquisador*: utiliza a ontologia para analisar a especialidade relacionada com o pesquisador.
4. *Criação de perfil do pesquisador e perfil de pesquisa*: fornece um mecanismo para geração dos perfis na linguagem OWL[16] (Ontology Web Language).
5. Desenvolvimento do sistema e inclusão do processo de ranqueamento e casamento para busca de especialistas.

Usando a ontologia de classificação de conhecimentos, o sistema é capaz de determinar a competência do pesquisador. Além disso, o sistema define uma pontuação para representar o grau de conhecimento que um pesquisador possui. Todo o processo

[15] Disponível em http://www.acm.org/about/class/2012

[16] Disponível em http://www.w3.org/TR/owl-features/

de definição do perfil é realizado sem a intervenção do usuário, por meio de uma abordagem implícita.

3.1.5 Uma Abordagem Colaborativa de Modelagem de Usuários para Recomendação de Conteúdo Personalizado

A proposta de Kim et al. (2008) para definição do perfil de usuários utiliza uma abordagem para descobrir padrões úteis e relevantes acerca do próprio usuário que são identificados pelo monitoramento de suas ações. Diante da obtenção de dados extraídos deste monitoramento, o perfil do usuário é construído, tendo como base a associação dos termos capturados ao longo do processo. Após a definição inicial do perfil, é utilizada uma abordagem colaborativa para identificar interesses similares de outros usuários. Esta estratégia visa resolver o problema do perfil vazio, que ocorre nos casos de usuários novos ou com baixa interação.

O trabalho é estruturado em 3 abordagens: (i) formação de vizinhança baseada em conteúdo; (ii) enriquecimento colaborativo de preferências do usuário; e (ii) geração de recomendação de conteúdo. A formação de vizinhança baseada em conteúdo visa a definição inicial do perfil. O principal objetivo é identificar um conjunto de vizinhos de usuário que é definido como um grupo de usuários que apresenta interesse de termos próximos aos dos usuários-alvo. Para a seleção dos melhores usuários vizinhos são utilizados termos personalizados de cada usuário. Para encontrar os vizinhos mais próximos, a similaridade do cosseno, que quantifica a semelhança de dois vetores de acordo com o seu ângulo, é utilizada para medir os valores de similaridade entre o usuário alvo e todos os outros usuários. Os termos personalizados de dois usuários são representados como vetores. O índice de similaridade entre dois usuários está no intervalo entre [0,1] e quanto maior a pontuação que um usuário tem, mais ele é semelhante ao usuário alvo. Após calcular a similaridade entre todos os usuários, define-se um conjunto de vizinhos mais próximos de cada usuário como uma lista ordenada de usuários.

Em relação ao enriquecimento colaborativo de preferência do usuário, o processo utiliza uma lista ordenada de usuários-alvo que são próximos, um conjunto de padrões de termos personalizados para o usuário e um conjunto de padrões de termos personalizados para usuários próximos do usuário-alvo. Para cada padrão identificado, padrões específicos do usuário-alvo são identificados. Para fazer o enriquecimento eficiente, consideram-se somente padrões específicos que contam com o maior padrão de suporte que o padrão geral. Finalmente, um conjunto de padrões colaborativos é identificado a partir de vizinhos mais próximos em relação ao usuário-alvo.

A geração de recomendação de conteúdo, depois que o modelo é enriquecido, está pronta para fornecer recomendações para novos conteúdos que o usuário ainda não tenha clicado ou lido. Baseado no modelo enriquecido para cada usuário, recomenda-se conteúdos ranqueados para o usuário, que ele poderia estar interessado. Para este fim, a tarefa mais importante na recomendação personalizada é gerar uma predição, isto é, a tentativa de especular sobre como um determinado usuário prefere conteúdos invisíveis. A proposta considera padrões de correspondência, ou seja, quantos padrões de interesse em um modelo de usuário estão contidos no conteúdo novo.

Uma vez que a predição do usuário-alvo ao seu conteúdo, que ele ainda não tenha lido, são computadas, os conteúdos são classificados em ordem decrescente de valor

previsto. Após, um conjunto de *N* conteúdos ordenados que tiveram os valores mais elevados são identificados para o usuário-alvo. Assim, os conteúdos são recomendados.

Todo o processo de construção do perfil é realizado sem a intervenção do usuário. A proposta não faz uso de aspectos temporais para a definição do perfil e não prevê um método ou mecanismo de atualização ou manutenção.

3.1.6 Extração de Redes Sociais de Pesquisadores Acadêmicos

O trabalho de Tang, Zhang e Yao (2007) tem por objetivo construir um perfil semântico para pesquisadores por meio de identificação e anotações coletadas da web, utilizando-se de *tags* e da abordagem CRF (*Conditional Random Fields*). O trabalho utiliza o sistema Arnetminer e define o perfil por meio de uma ontologia que inclui informações básicas (foto, afiliação e posição), informações de contato (endereço, e-mail e telefone), histórico educacional (onde se graduou e onde cursou doutorado) e publicações. Para cada pesquisador, cria-se um perfil com base na ontologia, extraindo as informações de páginas web. A Figura 3.2 mostra a página inicial do pesquisador.

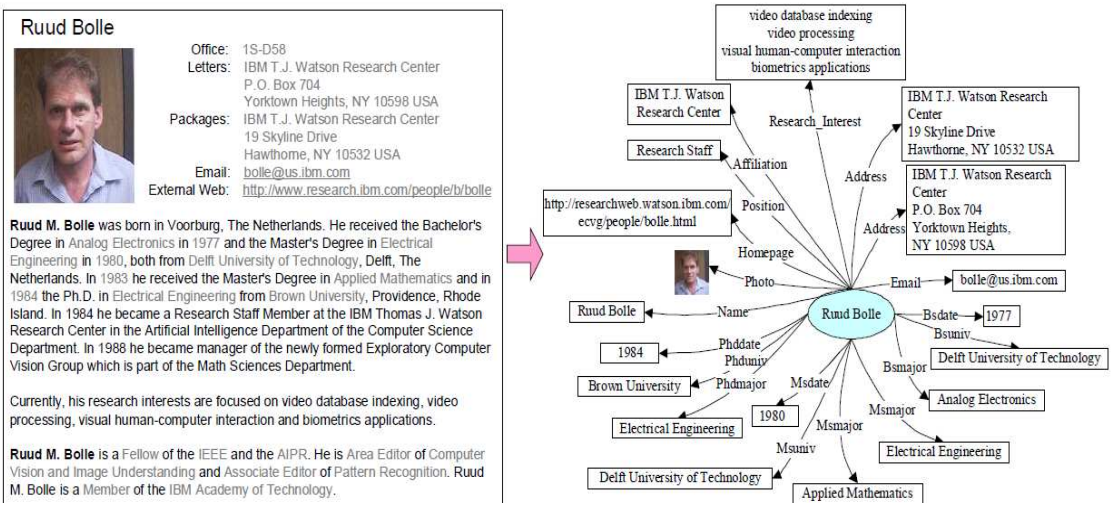


Figura 3.2: Exemplo do perfil do pesquisador (TANG; ZHANG; YAO, 2007).

A proposta do trabalho é dividida em três etapas: (i) encontrar páginas relevantes; (ii) pré-processamento; e (iii) marcação (*tagging*). Na primeira etapa, dado um nome de pesquisador, o sistema faz uma busca na web com informações sobre o pesquisador, em seguida, identifica as páginas usando um classificador e armazena no perfil uma página para o pesquisador alvo. Na segunda é realizado um pré-processamento onde o texto extraído da página do pesquisador é segmentado em *tokens* e cada *token* é associado as *tags*. Na última etapa, há o processo de marcação (*tagging*), onde dada uma sequência de *tokens* pode-se determinar a sequência mais provável correspondente de *tags*. A proposta emprega *Condition Random Fields*[17] (CRFs) como modelo de marcação (*tagging*).

Para cada unidade de *token*, três tipos de características são definidas: de conteúdo, padrão e de termo.

Para as *características de conteúdo* utiliza-se:

[17] http://en.wikipedia.org/wiki/Conditional_random_field

- características de palavras – se o *token* corrente é uma palavra padrão;
- características morfológicas – se o *token* corrente é maiúsculo.

Para uma imagem os recursos de conteúdo incluem:

- tamanho da imagem;
- relação altura/largura;
- formato de imagem (jpg, bmp, etc);
- cor da imagem;
- reconhecimento facial através da ferramenta *Open CV18*[19] para detectar a face de pessoas;
- nome do arquivo de imagem (se o nome do arquivo de imagem, mesmo que parcialmente, contém o nome do pesquisador);
- image Alt (se o atributo "alt" da imagem contém, mesmo que parcialmente, o nome do pesquisador); e
- palavras-chave da imagem.

Para as *características padrão* utiliza-se:

- palavras-chave – se o *token* corrente contém positivos de "*Fax/Phone*" ou "*Manager*"; e
- símbolos especiais – se o *token* corrente é uma palavra especial.

Para as *características de termo* utiliza-se:

- características de termo – se a unidade de *token* é um termo; e
- critérios de dicionário – se o termo é incluído em um dicionário.

O processo de extração dos dados, definição do perfil e identificação das redes de colaboração científica é conduzido sem a intervenção do usuário. A proposta não leva em consideração aspectos temporais.

3.1.7 Encontrando Especialistas em Redes Sociais

A proposta de Zhang, Tang e Li (2007) visa encontrar especialistas em uma rede social e se utiliza de uma abordagem baseada em propagação. A abordagem se divide em dois passos. Inicialmente, são utilizadas informações do usuário para gerar uma pontuação inicial dele mesmo dentro de sua especialidade. Os usuários são colocados em ordem decrescente de pontuação. Os usuários selecionados no topo são utilizadas para a construção de um sub grafo. Na sequência, vem a abordagem baseada em propagação, que utiliza os usuários do topo para fazer associações com as pessoas às quais possuem relacionamento, definindo uma pontuação de especialidade.

[18] Disponível em http://opencvlibrary.sf.net

[19] http://en.wikipedia.org/wiki/OpenCV

As associações são baseadas nas habilidades e no conhecimento do usuário. Os autores estabelecem a seguinte definição: uma rede social pode ser definida como um grafo $G = (V, E)$, onde $v \in V$ representa uma pessoa na rede social e $e^t_{ij} \in E$ representa um relacionamento com tipo t entre pessoas v_i e v_j, onde t pode ser, por exemplo, um coautor ou um colega (Figura 3.3).

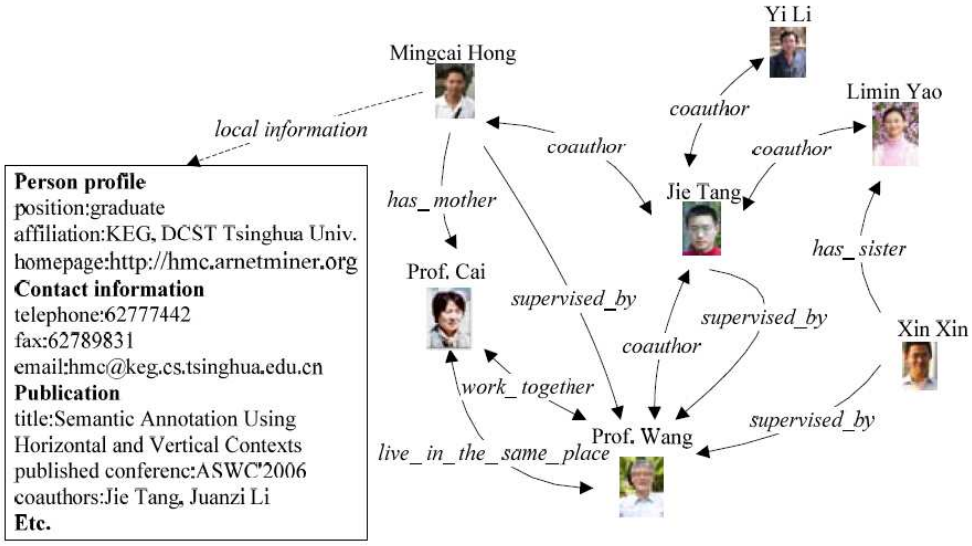


Figura 3.3: Exemplo de uma rede social (ZHANG; TANG; LI, 2007).

A Figura 3.3 apresenta parte de uma rede social acadêmica. Na rede, cada pessoa tem vários tipos de informação local, como, por exemplo, o perfil pessoal, informações de contato e publicações. Duas pessoas podem ter relacionamentos um com o outro e a relação pode ser direcional ou bi-direcional. Na mesma figura, "Jie Tang" tem um relacionamento denominado *supervised_by* com "Prof. Wang" e outros quatro relacionamentos bidirecionais, como *coauthor* com "Mingcai Hong".

A tarefa de encontrar especialistas é definida como: dado um tópico *q* de consulta, encontre um conjunto de pessoas da rede social e retorne uma lista ordenada destas pessoas. O trabalho é baseado em duas etapas: inicialização e propagação. Na inicialização, utilizam informação pessoal para calcular uma pontuação inicial de especialista para cada pessoa. A hipótese é que se uma pessoa é autora de muitos trabalhos sobre determinado tema, ou se ela é coautora de muitos trabalhos sobre determinado tema, é provável que esta pessoa seja um candidato a especialista nesse tema. A estratégia para a pontuação inicial de especialista é baseada no modelo probabilístico de recuperação de informação. Para uma pessoa, primeiro cria-se um documento *d* combinando informações pessoais. Após, estima-se um modelo probabilístico para cada documento e utiliza-se o modelo para calcular a relevância do documento a um determinado tópico. A pontuação é então estabelecida como a pontuação inicial de especialista daquela pessoa.

Na etapa de propagação, utilizam-se os relacionamentos entre as pessoas para melhorar a precisão da constatação de especialistas. A hipótese é que se uma pessoa conhece muitos especialistas de determinado tema ou se o nome da pessoa aparece como coautor muitas vezes com outro especialista, então é provável que esta pessoa seja um especialista no tema. A rede social é vista como um grafo onde é atribuído um peso em cada aresta para indicar o quão boa é a pontuação de especialista propagada de uma

pessoa para os seus vizinhos. Os assim chamados coeficientes de propagação variam de 0 a 1.

O trabalho define o perfil por meio de modelagem implícita, onde não há intervenção do usuário. Aspectos temporais não são considerados para a definição do perfil.

3.1.8 Encontrando Especialistas Usando Análise de Redes Sociais

O trabalho de Fu et al (2007) tem por objetivo identificar associações entre pessoas para construção de uma rede social e introduzir um algoritmo de propagação de especialistas para auxiliar a reclassificar outras pessoas como especialistas. De acordo com o algoritmo, uma pessoa pode adquirir mais expertise se possuir forte relação com outro especialista.

O problema de encontrar um especialista pode ser definido como a probabilidade de um candidato *c* ser um especialista, dado um tópico de consulta *q*. No entanto, não é possível estimar diretamente a probabilidade *P (c | q)*, pois o que se tem são apenas documentos, como páginas web e e-mails. Para isto, existe outro tipo de associação que está entre os próprios candidatos. As pessoas não estão isoladas, mas sim conectadas. A conexão entre os candidatos a especialistas pode ser identificada por meio da análise de documentos, tais como membros de um mesmo grupo de interesse que podem se comunicar com frequência por e-mails e pesquisadores que publicam um relatório técnico que figuram como coautores. As associações entre as pessoas podem ajudar na identificação de especialistas pelo fato que, dado um especialista x_c, o candidato y_c que tem forte associação $a\ (c_x, c_y)$ com ele, também é provável que seja um especialista.

O trabalho especifica um processo de propagação de conhecimentos para encontrar especialistas. São empregadas associações entre pessoas para propagar a probabilidade de especialistas altamente identificados gerarem possibilidade de se encontrar outros especialistas. O processo pode ser especificado como *P (c | q)*, onde, qual é a probabilidade do candidato y_c ser especialista, dado o especialista x_c? Tem-se um conjunto *S* de *N* candidatos que são mais propensos a serem especialistas em um primeiro momento. Assim, estima-se que candidatos que tenham fortes associações com os especialistas também sejam especialistas. Este fato sugere que a probabilidade de dividir experiência com especialistas se propaga para outros candidatos de acordo com as suas associações: se um especialista x_c tem uma probabilidade de conhecimento de P (c_x) e ele tem ω candidatos associados, cada um dos ω candidatos y_c tem a associação $a\ (c_x, c_y)$ e receberá um escore da fração de x_c.

A abordagem utilizada é híbrida, pois coleta informações do usuário de forma implícita e em determinada situação é necessária a intervenção do usuário. O trabalho não leva em consideração aspectos temporais.

3.2 Métricas Científicas

Esta seção tem por objetivo apresentar os trabalhos mais relevantes que possuem relação com a tese no âmbito de métricas científicas. A escolha dos trabalhos foi motivada pela aplicação das propostas estar alinhada com identificação de reputação de pesquisadores por meio da análise de diversos índices que possuem afinidade com esta tese.

3.2.1 Um Índice para Quantificar a Produção de Pesquisa Científica de uma Pessoa

A proposta de Hirsch (2005) é especificar um índice, denominado índice H (h-index), que tem por objetivo combinar critérios de qualidade e de quantidade, fornecendo uma métrica que mede o impacto de um pesquisador em um único número. Esse impacto é representado com base no número de artigos publicados pelo pesquisador e o número total de citações a seus artigos. Assim, um pesquisador tem um índice *H*, quando possuir *h* trabalhos com pelo menos *h* citações. Por exemplo, um pesquisador com $H = 9$ tem, pelo menos, 9 artigos que receberam 9 ou mais citações.

A Figura 3.4 apresenta graficamente a definição do h-index, onde o eixo (x) representa o número de artigos e o eixo (y) representa o número de citações. No gráfico, o ponto de intersecção entre as coordenadas (x) e (y) representa o valor do h-index do pesquisador.

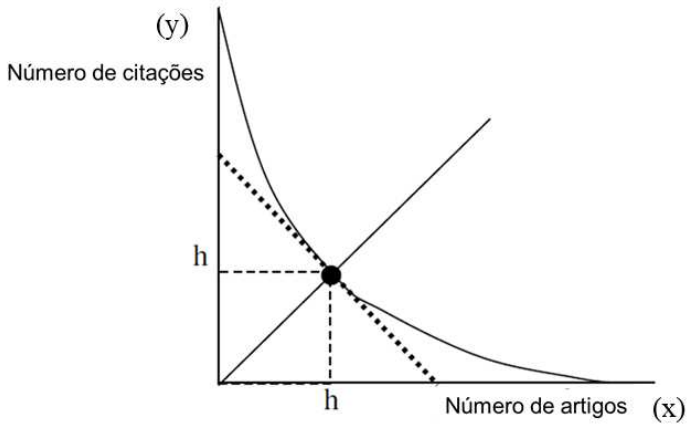


Figura 3.4: Gráfico que representa o índice h (adaptado de (HIRSCH, 2005)).

A Tabela 3.1 apresenta o exemplo da identificação do h-index de um pesquisador cujo valor *H* é 18. Para calcular o índice, deve-se ordenar os artigos do pesquisador em ordem decrescente de citações (do mais citado para o menos citado). Assim, um pesquisador tem um índice *H* se *h* de seus *Np* artigos tem, ao menos, *h* citações cada e os outros ($Np - h$) artigos tem não mais do que *h* citações cada.

Tabela 3.1: Exemplo de identificação do h-index de um pesquisador.

Número de Citações	Ranking
11338	1
695	2
178	3
167	4
144	5
130	6
121	7
114	8
113	9
113	10
91	11
75	12

67	13
61	14
49	15
33	16
26	17
19	18
17	19
17	20
9	21
7	22

No exemplo apresentado na Tabela 3.1, o pesquisador possui 11338 citações para um de seus artigos, no caso, o mais citado. Pode-se perceber que a diferença entre o artigo mais citado (11338 citações) e o segundo artigo mais citado (695 citações) é muito elevada. Neste caso, evidencia-se um dos problemas mais comuns quando se avalia a reputação de pesquisadores utilizando-se apenas o critério da citação, onde um trabalho altamente citado não influencia no cálculo da reputação.

Outra observação importante que pode-se identificar no exemplo da Tabela 3.1, é que o pesquisador possui outros artigos publicados, mas como não atingiram um mínimo de citação, ficam de fora do cálculo do h-index. Isso demonstra que, para se obter um h-index elevado, o trabalho do pesquisador deve repercutir na comunidade científica. Assim, se um pesquisador possui muitas publicações que não são citadas, bem como recebe muitas citações em poucos trabalhos, seu h-index será baixo.

O h-index, ao longo de seus oito anos de utilização pela comunidade científica, está sendo usado para avaliar não apenas a reputação de pesquisadores, mas também o impacto de conferências científicas, periódicos e instituições. Com isso, percebe-se que o índice proposto por Hirsch se tornou uma referência no meio científico, sendo utilizado em diversos trabalhos para avaliar a reputação individual de pesquisadores, como em (LI et. al, 2010; PATTERSON; HARRIS, 2009; van RAAN, 2006), apesar dos problemas evidentes que são discutidos na sequência. Mesmo assim, reacendeu a busca por métricas capazes de avaliar a ciência ou os pesquisadores, dando início a uma série de trabalhos que visam encontrar uma forma de avaliar a qualidade do que é produzido pelos cientistas.

Apesar do h-index ser uma das métricas mais utilizadas pela comunidade acadêmica e científica, seus problemas são evidentes. O índice pode não envolver trabalhos que são pouco citados e subavaliados por indicadores bibliométricos. Mesmo assim, isto não indica que o trabalho não seja bom ou não represente uma demanda específica da sociedade, da indústria ou da academia (ZHANG, 2009). Ainda, o valor do h-index de um pesquisador nunca diminui (JIN, 2007). Mesmo que o pesquisador não tenha mais atividade na comunidade científica ou que diminua sua produção, seu h-index vai, no mínimo, permanecer o mesmo.

Diante desse contexto, pode-se concluir que deve-se ter cuidado ao utilizar um indicador bibliométrico como o h-index de forma isolada. O ideal é encontrar um mecanismo que permita a avaliação do pesquisador de forma abrangente, levando em consideração os elementos de sua trajetória científica, construída ao longo de sua carreira, que é um dos objetivos desta tese. A utilização de um índice que utiliza o

critério de citação como garantia de qualidade do trabalho de um cientista pode não representar a realidade.

3.2.2 Um Aperfeiçoamento do h-index: o g-index

No trabalho de Egghe (2006a) é apresentado um novo índice, denominado g-index, que visa complementar o h-index (HIRSCH, 2005). O g-index melhora o h-index, dando mais peso aos artigos altamente citados. Ele pode ser definido da seguinte maneira: "para um conjunto de artigos, classificados em ordem decrescente das citações que receberam, o g-index é o maior número tal que os *top g* artigos receberam (somados), pelo menos g^2 citações" (EGGHE, 2006a).

Baseado na definição do g-index, pode-se concluir que o valor de *g* sempre será maior ou igual ao valor do h-index ($g \geq h$). O exemplo apresentado na Tabela 3.2 mostra essa relação. O exemplo se refere à produção de um pesquisador hipotético. Os elementos apresentados no exemplo são descritos a seguir:

- TC: representa o número total de citações de cada artigo no ranking *r*, em ordem decrescente de citações;
- r: refere-se a posição do artigo no ranking;
- ∑TC: indica o somatório das citações dos artigos do ranking *r*;
- r^2: representa o quadrado do ranking *r*.

Tabela 3.2: Identificação do h-index e do g-index de um pesquisador.

TC	r	∑TC	r^2
47	1	47	1
42	2	89	4
37	3	126	9
36	4	162	16
21	5	183	25
18	6	201	36
17	7	218	49
16	8	234	64
16	9	250	81
16	10	266	100
15	11	281	121
13	12	294	144
13	13	307	169
13	14	320	196
13	15	333	225
12	16	345	256
12	17	357	289
12	18	369	324
12	19	381	361
11	20	392	400

h-index

g-index

Fonte: adaptado de (EGGHE, 2006b).

O exemplo ilustra as diferenças entre o cálculo do h-index e o cálculo do g-index. Neste exemplo, o h-index do pesquisador é 13 e o g-index é 19. Observa-se que o h-index identificado para o pesquisador foi 13 porque ele possui 13 artigos que receberam, ao menos, 13 citações (o artigo 14 no *ranking* não tem mais de 13 citações). Já o valor obtido pelo g-index foi 19 porque os *g* artigos mais citados receberam, somados, pelo menos g^2 citações ($19^2 = 361$, onde o somatório dos 19 artigos é 381 e

381 > 361). No caso do artigo classificado em vigésimo no *ranking*, este não pode ser considerado, pois o somatório das citações somadas é equivalente a 392, sendo menor do que g^2 (20^2), ou seja, 392 < 400.

O g-index apresenta como principais vantagens em relação ao h-index o fato de considerar no cômputo o peso das citações recebidas pelos artigos mais citados e o número total de documentos que não limitam o valor do índice, como é o caso do h-index (COSTAS; BORDONS, 2008).

O trabalho de Costas e Bordons (2008) apresenta um estudo comparativo entre o h-index e o g-index. Os critérios utilizados foram comparar pesquisadores que possuem elevada produção científica (grupo A) com pesquisadores que possuem produção científica de moderada a baixa, mas tem artigos altamente citados (grupo B). Os resultados mostram que o g-index é mais sensível que o h-index para avaliar cientistas do grupo B, pois o resultado para os pesquisadores do grupo B foi um valor de g-index mais elevado. Em contrapartida, os pesquisadores do grupo B apresentaram um h-index menor do que os do grupo A, mesmo que eles tenham um maior número de citações por artigo e maior taxa de artigos altamente citados.

Embora o g-index apresente algumas vantagens em comparação ao h-index, algumas limitações permanecem, principalmente, a dificuldade de selecionar toda a produção dos pesquisadores e suas citações, a existência de diferentes tipos de documentos com diferentes impactos, o problema da autocitação e a incapacidade para comparar pesquisadores de diferentes áreas (VINKLER, 2007). Ainda, há uma limitação específica observada no g-index, a influência de um artigo altamente citado interferir no valor do índice, mesmo que isto não represente a média do desempenho de um pesquisador (COSTAS; BORDONS, 2008).

Da mesma forma que o h-index, o g-index utiliza apenas as publicações e o impacto dessas publicações para identificar a reputação de um pesquisador. Apesar de ser uma evolução do h-index, apresentando algumas melhorias como exposto anteriormente, o g-index também deve ser utilizado de forma cuidadosa, pois não representa a carreira científica de um pesquisador de forma abrangente. Aliado a isso, como a métrica visa suprir, mesmo que parcialmente, os problemas do h-index com relação a artigos altamente citados, o g-index não deve ser utilizado de forma isolada para avaliação de cientistas.

3.2.3 O AR-Index: Complementando o h-index

A abordagem proposta por Jin (2007) e Jin et al. (2007), denominada AR-index, visa complementar o h-index proposto por (HIRSCH, 2005). O AR-index pode ser definido como a raiz quadrada do somatório da média do número de citações por ano dos artigos incluídos no *h-core*. O *h-core* representa os artigos classificados entre 1 e h, onde h é o valor do h-index. O AR-index pode ser expresso pela Fórmula 1.

$$AR = \sqrt{\sum_{j=1}^{h} \frac{cit_j}{a_j}} \qquad (1)$$

Fonte: (JIN, 2007).

Onde h = h-index, cit = quantidade de citações e a = número de anos desde a publicação.

Além de empregar o número real de citações de artigos pertencentes ao *h-core* como parâmetro, o AR-index também considera a idade da publicação. Deste modo, o h-index é complementado por um índice que pode realmente diminuir. Segundo (JIN, 2007), este comportamento é uma condição necessária para um bom indicador de avaliação da pesquisa.

O exemplo a seguir, baseado em Jin et al. (2007), mostra o cálculo do o AR-index ao longo de alguns anos dos artigos de BC Brookes (Brookes foi premiado com a medalha Derek John de Solla Price, em 1989). Os resultados são apresentados na Tabela 3.3.

Tabela 3.3: Resultado do AR-index do pesquisador BC Brookes.

Ano	AR-index
2002	3,93
2003	3,89
2004	3,84
2005	3,79
2006	3,76
2007	3,73

Fonte: adaptado de (JIN et al., 2007).

Pode-se observar no exemplo apresentado que o AR-index do pesquisador BC Brookes diminuiu com o passar dos anos. Isto é compreensível pelo fato de que seu falecimento se deu no ano de 1991. Esta situação evidencia a relevância do AR-index em relação ao tempo, ou seja, se o pesquisador não continua publicando e não é citado, ou é pouco citado, sua reputação na comunidade científica tende a diminuir.

Existem outros dois trabalhos do mesmo autor que possuem relação com esta tese (JIN et al., 2007; ROUSSEAU; JIN, 2008), no entanto, ambos são extensões que também enfocam a idade da publicação incorporada à métrica.

Da mesma forma que o h-index e o g-index, o AR-index é dependente, exclusivamente, dos artigos publicados por um pesquisador e das citações a estes artigos. A abordagem desta tese se diferencia por possibilitar uma avaliação abrangente, que leva em consideração diversos elementos da carreira científica de um pesquisador, ao invés de focar apenas em trabalhos publicados e citações.

3.2.4 ArnetMiner – Extração e Mineração de Redes Sociais Acadêmicas

Tang et al. (2008-b) apresentam um sistema denominado ArnetMiner[20], que tem por objetivo ser uma fonte de informação referente a produção de pesquisadores, redes de coautoria, veículos de publicação, assim como outros elementos inerentes à atividades científicas.

A Figura 3.5 apresenta a arquitetura do Arnetminer. A arquitetura consiste de cinco componentes principais:

1. *Extração* – concentra-se em extrair perfis de pesquisadores da Web automaticamente. Primeiro, coleta e identifica uma página inicial a partir da Web. Em seguida, utiliza uma abordagem unificada para extrair as

[20] *Disponível em*: http://arnetminer.org/

propriedades do perfil de documentos identificado. O sistema extrai publicações on-line de bibliotecas digitais por meio de regras.

2. *Integração* – integra os perfis extraídos de pesquisadores e extrai as publicações usando o nome do pesquisador como o identificador. Uma estrutura probabilística foi proposta para lidar com o problema da ambiguidade de nomes na integração. Os dados integrados são armazenados em uma base denominada *Researcher Network Knowledge Base* (RNKB).
3. *Armazenamento e Acesso* – fornece armazenamento e índice para os dados extraídos/integrados no RNKB. Especificamente, para o armazenamento emprega MySQL e para o índice, utiliza o método de indexação de arquivo invertido (BAEZA-YATES; RIBEIRO-NETO, 2011).
4. *Modelagem* – utiliza um modelo probabilístico gerado para modelar simultaneamente diferentes tipos de informação e implementa uma distribuição por tópico para cada tipo de informação.
5. *Serviços de pesquisa* – baseado nos resultados da modelagem fornece vários serviços de pesquisa, como pesquisa por conhecimentos e busca por associação. Também oferece outros serviços, como, por exemplo, encontrar autores por interesse e sugestões acadêmica (artigos e citação de artigos).

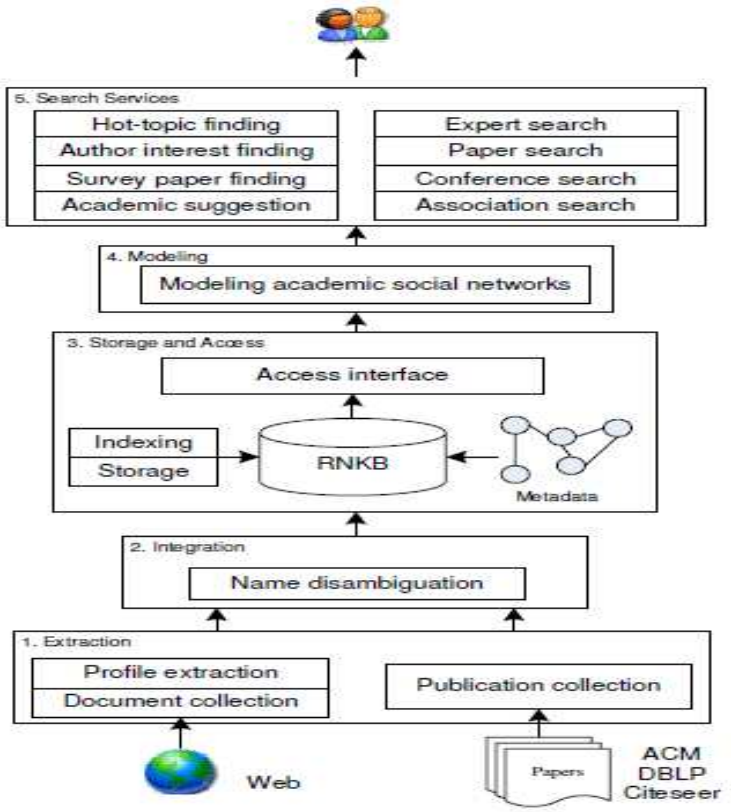


Figura 3.5: Arquitetura do Arnetminer (TANG et al., 2008b).

Atualmente, além de disponibilizar à comunidade científica as métricas h-index e g-index, a equipe do ArnetMiner apresentou seis novas métricas para classificar pesquisadores, a saber: *Longevity*, *Activity*, *Diversity*, *Sociability*, *Uptrend* e *New Star*. As métricas são detalhadas a seguir:

- *Longevity* – caracteriza a longevidade de um pesquisador, refletindo a duração de sua vida acadêmica. O início é considerado o ano de publicação do primeiro artigo e o final refere-se ao ano da última publicação. A longevidade de um pesquisador pode ser definida pela Fórmula 2.

$$\text{Longevity}(A) = Y_A(A\text{'s last paper}) - Y_A(A\text{'s first paper}) \quad (2)$$

Fonte: http://arnetminer.org/AcademicStatistics.

- *Activity* - refere-se às atividades de um pesquisador e é definida com base nos artigos publicados nos últimos anos. A métrica considera a importância de cada artigo para definir o *score*. Para isso, é necessário calcular o fator de impacto da publicação, conforme apresentado na Fórmula 3.

$$\text{IS}(G) = \sum_{each\ paper(P)\ in\ the\ G} IC(P) * weight(P) \quad (3)$$

Fonte: http://arnetminer.org/AcademicStatistics.

Na definição, *G* é um grupo de artigos. *IC (P)* é o impacto do veículo de publicação em que o artigo foi publicado. Na métrica do fator de impacto também é calculado o tamanho do artigo (em número de páginas). Se o artigo for menor que três páginas, então atribui-se 1/5 do fator de impacto da publicação. Se o artigo for menor ou igual a três páginas e menor do que cinco páginas, atribui-se 1/3. Dessa forma, *Weight (P)* refere-se a 1/5 ou a 1/3. Tendo o fator de impacto definido pode-se estabelecer a *Activity* de um pesquisador. A Fórmula 4 apresenta a definição.

$$\text{Activity}(A) = \sum_{each\ year(n)\ in\ recent\ N\ years} \text{IS}(G_n) * \text{weight}(n) \quad (4)$$

Fonte: http://arnetminer.org/AcademicStatistics.

Na definição, no ano *n* (*n* pertence a *N* últimos anos), *G* é um grupo de artigos publicados pelo pesquisador *A* no ano *n*. *Weight (n)* = $\alpha^{this\ year\ -\ n}$. Os valores de *N* e α são previamente definidos. *N = 4* e $\alpha = 0{,}75$ se o mês atual é na primeira metade do ano (mês <julho) e *N = 3* e $\alpha = 0{,}85$ se o mês atual é no segundo semestre.

- *Diversity* - refere-se à diversidade de áreas de atuação de um pesquisador e é definida pela extração de dados das publicações e dos veículos de publicação dos pesquisadores, conforme especificado em (TANG et al. 2008b). A Fórmula 5 apresenta a definição de como é realizado o cálculo dos artigos e suas áreas de abrangência.

$$P_A(t) = \frac{\#\text{papers of A belong to topic t}}{\#\text{all papers of A}} \quad (5)$$

Fonte: http://arnetminer.org/AcademicStatistics.

A diversidade do pesquisador é definida de acordo com a Fórmula 6.

$$\text{Diversity}(A) = -\sum_{t \in \text{ all topic of } A} P_A(t) \log P_A(t) \quad (6)$$

Fonte: http://arnetminer.org/AcademicStatistics.

- *Sociability* – caracteriza a quantidade de coautores que um pesquisador possui e é definida de acordo com a Fórmula 7.

$$\text{Sociability}(A) = 1 + \sum_{\text{each coauthor}(c)} \ln(\#\text{copaper}_c) \quad (7)$$

Fonte: http://arnetminer.org/AcademicStatistics.

O elemento *#copaperc* representa o número de artigos publicados entre o pesquisador e o coautor *c*.

- *Uptrend (Rising Star)* – é caracterizado para definir o grau crescente de um pesquisador. A informação de cada artigo inclui a data de publicação e o fator de impacto do veículo, assim, utiliza-se o método dos mínimos quadrados para ajustar uma curva a partir de artigos publicados em *N* anos recentes. A curva é usada para prever uma pontuação no ano seguinte, que é definida pela Fórmula 8:

$$\text{Slope}(A) = \frac{\sum_{i=1}^{N}\left(t_i * IM_{thisyear-N+i}(A)\right) - N\bar{t} * \overline{IM(A)}}{\sum_{i=1}^{N}\left(IM_{thisyear-N+i}(A)^2\right) - N\overline{IM(A)}^2} \quad (8)$$

Fonte: http://arnetminer.org/AcademicStatistics.

Tendo-se o valor da curva, o grau crescente de um pesquisador (*uptrend*) é definido pela Fórmula 9.

$$\text{Uptrend}(A) = Average\left(IM(A)\right) - Slope(A) * Average \quad (9)$$

Fonte: http://arnetminer.org/AcademicStatistics.

Para a definição do *uptrend* considera-se $t_i = i, IM_y(A) = \text{IS}(G_y)$. Se o cálculo for realizado na primeira metade do ano o valor de *N* não deve ser considerado. Se o cálculo for realizado para a segunda metade do ano, considera-se *N = 3* e $IM_y(A) = \frac{12}{this\ month} IM_y(A)$.

- *New Star* - identifica pesquisadores que possuem uma vida acadêmica menor ou igual a cinco anos e seu valor é baseado na métrica de atividade (*Activity*).

A Figura 3.6 apresenta um exemplo de utilização das métricas disponibilizadas pelo ArnetMiner referente ao pesquisador “Anil K. Jain”, da área de Ciência da Computação, que possui o mais elevado valor de h-index (121) dentre todos os pesquisadores vinculados a essa área.

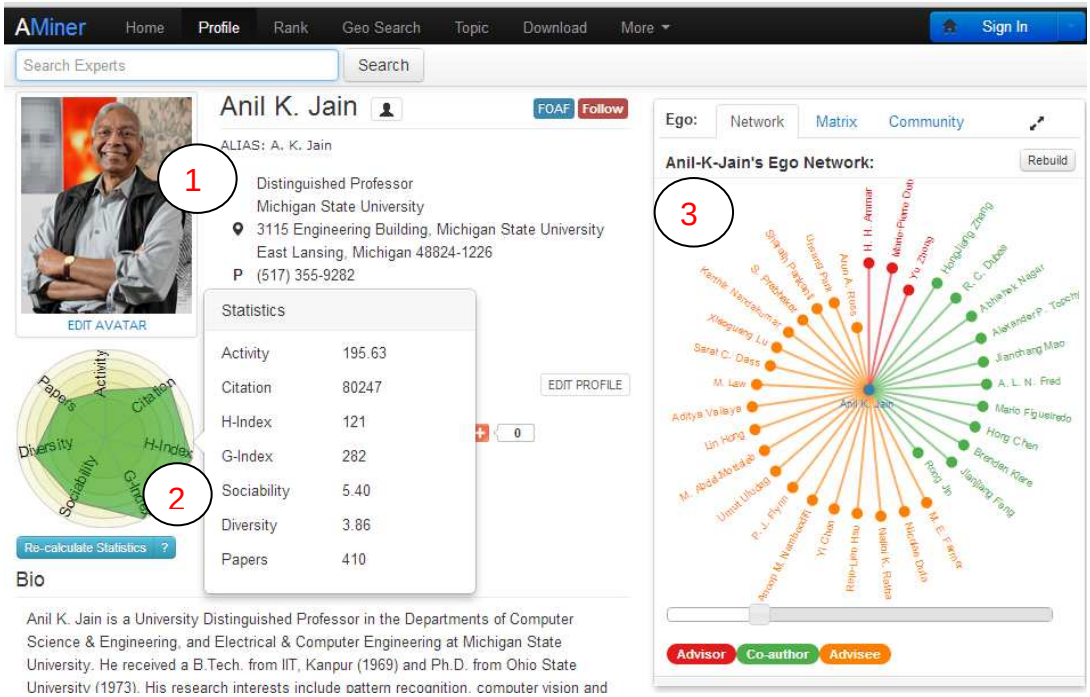


Figura 3.6: Dados do pesquisador Anil K. Jain. Adaptado de http://arnetminer.org/person/anil-k-jain-324696.html.

Na Figura 3.6, indicativo 1, são apresentados dados de identificação do pesquisador, como nome, nome em citações, foto, instituição de vínculo e endereço de contato. No indicativo 2 pode-se observar alguns valores e métricas do pesquisador, como atividade (*Activity*), número de citações (*Citation*), h-index (*h-index*), g-index (*g-index*), sociabilidade (*Sociability*), diversidade (*Diversity*) e quantidade de artigos publicados (*Papers*). Já o indicativo 3 refere-se a rede social acadêmica do pesquisador, apresentando dados como seus coautores e suas orientações. As demais métricas implementadas pelo ArnetMiner, como *New Star* e *Uptrend* (*Rising Star*) são apresentadas no ambiente individual somente quando o pesquisador possui dados associados, o que não é o caso do pesquisador Anil K. Jain. Em relação à métrica *Longevity*, esta não mais apresenta valor associado, sendo recentemente substituída pela métrica *Activity*.

Apesar de ser um mecanismo importante para a visibilidade das informações acerca do perfil de pesquisadores, suas métricas não incorporam um conceito de reputação mais abrangente. Isto pode ser percebido pelo fato de todas as métricas se basearem na produção de artigos ou no impacto dessa produção para a geração de seus índices. Estas características não desmerecem o ArnetMiner, mas proporcionam uma visão limitada, pois subentende-se que a reputação de um pesquisador é medida pelos artigos que publica ou pelas citações que estes artigos recebem.

3.2.5 ResearchGate – Scientific Network

ResearchGate[21] é um portal que tem por objetivo aproximar pesquisadores e dar visibilidade às pesquisas desenvolvidas por eles. Seu ambiente proporciona o compartilhamento de informações acerca de pesquisadores e cientistas que vão desde a

[21] https://www.researchgate.net

disponibilização de artigos para visualização e *download* até a formação de redes entre os pesquisadores.

O ResearchGate apresenta uma métrica para medir a reputação de pesquisadores denominada *RG Score*. A métrica é baseada nas interações entre os pesquisadores, tornando visível e quantificável o processo de compartilhamento de publicações, a disponibilização de dados das pesquisas e o envolvimento em discussões sobre assuntos dos trabalhos científicos. Embora o modelo de publicação científica tradicional tenha trazido inúmeras inovações e avanços, a velocidade da descoberta é muitas vezes dificultada pela falta de velocidade na publicação. O ResearchGate visa facilitar a divulgação dos resultados das pesquisas, dando um novo ritmo para a ciência. Os pesquisadores podem publicar seus resultados em tempo real, se beneficiar do *feedback* imediato de seus pares e, por meio do *RG Score*, transformar seu trabalho em um índice de reputação quantificável.

A pontuação do pesquisador é calculada com base em como a comunidade científica interage com o seu conteúdo, quantas vezes, e quem está interagindo. Um dos objetivos da métrica é garantir que a reputação seja definida pelos pares, pois as interações formam a base do *RG Score*. O algoritmo verifica não só como os pares recebem e avaliam a sua contribuição para a pesquisa, mas também verifica como é a reputação desses colegas. Dessa forma, quanto maior a pontuação daqueles que interagem com a pesquisa de um cientista, mais a pontuação deste cientista aumentará.

Apesar da métrica não estar disponível para visualização, a base de cálculo do *RG Score* está centrada em quatro elementos: publicações, perguntas, respostas e seguidores. O nível de interação entre os pesquisadores, envolvendo estes quatro elementos, bem como outros mais especializados, como visualizações das publicações e downloads, gera o índice de reputação de cada pessoa que pode ser visualizado em seu perfil.

O *RG Score*, métrica utilizada pelo ResearchGate, não mede a reputação de um pesquisador utilizando dados de toda sua produção científica, pois a base da reputação é a interação entre os pesquisadores. Isso tende a ser um ponto negativo da proposta, uma vez que para pesquisadores mais experientes o fator tempo pode ser um problema. Aliado a isso, existem muitos pesquisadores que são mais conservadores quando se trata de adoção de novas tecnologias, ainda mais quando há a necessidade de se dedicar a inserção ou confirmação de informações de suas pesquisas, bem como promover discussões entre membros da comunidade científica.

3.2.6 O e-index, complementando o h-index para o excesso de citações

O e-index foi proposto por Zhang (2009) e tem como objetivo suprir duas fragilidades do h-index, apresentado por Hirsch (2005). A primeira fragilidade é que no h-index existe perda de informação de citações, uma vez que as citações dos trabalhos além do h-core são ignoradas. O autor afirma que a avaliação de um pesquisador realizada como base no h-index podem ser enganosas porque os pesquisadores que tem um h-index inferior podem ter mais citações do que aqueles que tem um h-index superior. Outra desvantagem apresentada por Zhang (2009) em relação ao h-index é que ele possui baixa resolução, resultado de seu baixo potencial, uma vez que está incluído na categoria dos números naturais, muito mais baixo do que um conjunto de números reais. Além disso, o autor apresenta que o h-index tem um intervalo relativamente

estreito, o que aumenta significativamente a ocorrência de um grupo de cientistas ter um h-index idêntico.

O e-index pode ser definido como o número que representa o excesso ignorado de citações do h-index. As citações em excesso de todos os trabalhos do pesquisador ignoradas pelo h-index são representadas pela Fórmula 10.

$$e^2 = \sum_{j=1}^{h} \left(cit_j - h\right) = \sum_{j=1}^{h} cit_j - h^2 \qquad (10)$$

Fonte: (ZHANG, 2009).

onde cit_j são as citações recebidas por j^{th} artigos e e^2 denota as citações em excesso dentro do h-core.

Como o e-index representa o excesso de citações ignoradas pelo h-core, pode-se utilizar o valor de e-index para complementar o h-index, pois os valores são independentes. Outras métricas que são dependentes do h-index tem a redundância de informações e, por conseguinte, quando utilizadas em conjunto com o h-index, escondem as diferenças reais no excesso de citações de diferentes pesquisadores.

A Figura 3.7 apresenta graficamente o e-index. O e-index pode ser visualizado na região cinza escuro do gráfico.

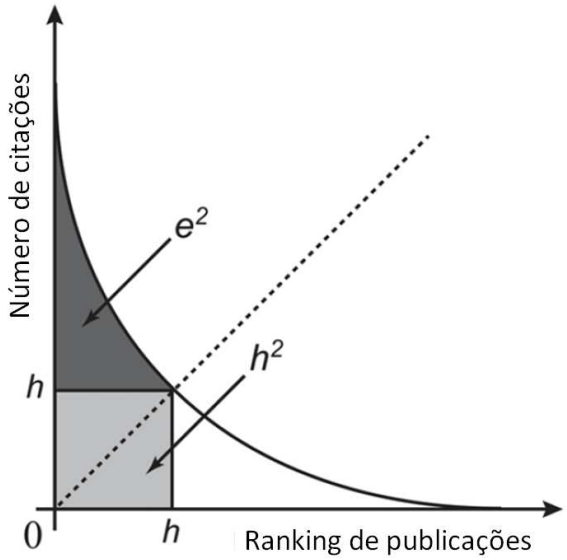


Figura 3.7: Explanação geométrica do e-index (e2) (adaptado de (ZHANG, 2009)).

De acordo com a Figura 3.7 percebe-se claramente que e^2 é independente de h^2, que representa a região de abrangência do h-index. Isto demonstra que os dois índices podem ser utilizados distintamente. Nota-se que quanto maior for a região representada pelo e^2, maior é o excesso de citações ignoradas, consequentemente, maior perda de informação ocorre quando se utiliza o h-index de forma isolada. Assim, quanto menor for o e^2, mais confiável é o h-index.

A Tabela 3.4 apresenta um exemplo de três renomados pesquisadores da área de química que possuem um elevado valor de h-index e analisa sua relação com o e-index. A ordem está pelo valor decrescente de h-index e, na sequência, pelo número total de citações.

Tabela 3.4: Comparação entre o h-index e o e-index de três cientistas da área de química.

Pesquisador	Número total de citações	h-index	e^2	e-index
A	5596	51	2995	54.73
B	3568	51	967	31.10
C	15496	50	12996	114.00

Fonte: adaptado de (ZHANG, 2009).

No exemplo apresentado na Tabela 3.4 pode-se observar que os pesquisadores A e B possuem o mesmo valor de h-index (51). No entanto, o Pesquisador A possui 2995 citações que não são consideradas no h-index. Como estas citações não são utilizadas no cálculo, existe uma compreensão de que os dois pesquisadores possuem o mesmo nível de reputação. Como o e-index considera o total de publicações que são ignoradas pelo h-index, fica evidenciado que o Pesquisador A possui reputação mais elevada, uma vez que possui e-index = 54.73, contra e-index = 31.10 do Pesquisador B. Essa diferença aponta que o valor do e-index do Pesquisador A é 1,76 vezes maior que a reputação do Pesquisador B, pois possui três vezes mais citações.

Mesmo suprindo algumas fragilidades do h-index, o e-index mantém a mesma base conceitual do h-index, de avaliar a reputação de pesquisadores utilizando como critério único o impacto da produção do pesquisador, representado pela quantidade de publicação e suas citações. A abordagem proposta nesta tese se diferencia por possibilitar uma avaliação abrangente, que leva em consideração diversos elementos da carreira científica de um pesquisador, ao invés de focar apenas em trabalhos publicados e citações.

3.2.7 Hg-index: Um Novo Índice para Caracterizar a Produção Científica de Pesquisadores Baseado no h-index e no g-index

A métrica proposta por Alonso et al. (2010), denominada hg-index, tem como objetivo caracterizar a produção científica de pesquisadores usando como base o h-index e o g-index, tentando manter as vantagens das duas propostas, bem como minimizar as suas desvantagens.

O hg-index é computado como a média geométrica entre o h-index e o g-index e pode ser visualizado pela Fórmula 11.

$$hg = \sqrt{h \times g} \tag{11}$$

Fonte: (ALONSO et al., 2010).

Isto demonstra que o h-index de um pesquisador é tal que $h \leq hg \leq g$ e que $hg - h \leq g - hg$, onde hg-index corresponde ao valor próximo entre h e g. A seguir, são apresentadas algumas vantagens do hg-index de acordo com (ALONSO et al., 2010):

- é simples de calcular, uma vez identificados os valores do h-index e do g-index;
- fornece uma maneira mais refinada de comparar cientistas, principalmente se comparado ao h-index, onde cientistas com número de publicações e citações diferentes possuem o mesmo h-index;

- é valorizado na mesma escala que o h-index e o g-index, pois representam os mesmos indicadores (produção e impacto), mas é mais fácil de entender e de comparar com os demais índices existentes;
- leva em consideração os artigos altamente citados (o h-index é insensível a isso), mas reduz o impacto dessa característica (um inconveniente no g-index), obtendo um melhor equilíbrio entre o impacto dos trabalhos.

A Tabela 3.5 apresenta um exemplo de alguns pesquisadores e seus respectivos índices a fim de evidenciar o comportamento destes cientistas em relação aos três índices analisados. A ordem dos dados é baseada na seguinte sequência: h-index, g-index e hg-index.

Tabela 3.5: Comparação entre o h-index, o g-index e o hg-index de pesquisadores reconhecidos com o Price Medal Awardees[22].

Pesquisador	h-index	g-index	hg-index
Garfield	27	59	39.91
Narin	27	40	32.86
Braun	25	38	30.82
Van Raan	19	27	22.65
Small	18	39	26.50
Schubert	18	30	23.24
Glänzel	18	27	22.05
Moed	18	27	22.05
Martin	16	27	20.78
Ingwersen	13	26	18.38
Egghe	13	19	15.72
Leydersdorff	13	19	15.72
Rousseau	13	15	13.96
White	12	25	17.32

Fonte: adaptado de (EGGHE, 2006b; ALONSO et al., 2010).

Pode-se observar em relação ao hg-index que ele fornece melhor granularidade que os outros índices comparados. Isto pode ser uma vantagem quando se necessita classificar pesquisadores. No caso de comparação entre os pesquisadores Van Raan e Small, a ordem pelo h-index coloca Van Raan na frente. Mas Small tem um valor de g-index mails elevado que Van Raan, consequência de seus trabalhos terem sido mais citados. Esta propriedade acaba por elevar o valor do hg-index de Small, fazendo com que fique a frente de Van Raan também nesta métrica. Este exemplo mostra que existe uma diferença significativa entre os cientistas se for utilizado como referência o valor de hg-index (22,65 de Van Raan contra 26,50 de Small, uma diferença de 3,85). Se fosse utilizado o g-index como meio de comparação, a diferença entre os pesquisadores seria de 12 (27 para Van Raan e 39 para Small). No caso de utilização do h-index como método comparativo, a diferença seria muito pequena, pois os pesquisadores possuem um valor de h-index muito próximo (19 para Van Raan e 18 para Small).

[22] Reconhecimento dado pelo *International Journal Scientometrics* a cientistas que se destacaram pela contribuição à Ciência.

Apesar do hg-index melhorar alguns pontos do h-index e do g-index, a métrica mantém a visão de que pesquisadores devem ser avaliados exclusivamente pela sua produção bibliográfica e o impacto dessa produção na comunidade científica.

3.2.8 O *h*'-Index, Efetivamente Melhorando o *h*-Index Baseado na Distribuição de Citações

A métrica proposta por Zhang (2013-b), denominada *h*'-Index, visa complementar o h-index, proposto por Hirsch (2005) e o e-index, proposto por Zhang (2009). A métrica busca agregar no cômputo da reputação a quantidade de publicações de um pesquisador que ficam abaixo do h-core, que alguns autores chamam de h-tail (YE; ROUSSEAU, 2010; ZHANG, 2013-a). Como o h-index, por si só, não leva em consideração o excesso de citações, nem mesmo o h-tail, estas características levam a perda de informação, o que pode comprometer a identificação da reputação, pois não reflete a realidade da produção de um pesquisador.

Para solucionar este problema, o *h*'-Index considera todas as citações que um pesquisador possui. Nos gráficos A, B e C, apresentados na Figura 3.8, pode-se observar as regiões onde as publicações não consideradas no cálculo do h-index estão localizadas. Os dados são referentes a três pesquisadores classificados como perfeccionista (gráfico A), prolífico (gráfico B) e produtor em massa (gráfico C) (BORNMANN; MUTZ; DANIEL, 2010). A região h^2 representa a produção computada pelo h-index. A região e^2 refere-se às publicações que são altamente citadas e ignoradas pelo h-index. Já a região t^2 representa as publicações que tiveram poucas citações e que ficam abaixo do h-core (o t^2 é apresentado por Zhang (2013-a) onde é discutido o t-index).

Pode-se observar que o pesquisador considerado perfeccionista (gráfico A) possui menos publicações, mas estas são altamente citadas. Já o pesquisador denominado prolífico (gráfico B) mantém um equilíbrio entre o número de publicações e número de citações. O pesquisador considerado produtor em massa apresenta um elevado número de publicações, no entanto, estas publicações são pouco citadas por outros pesquisadores.

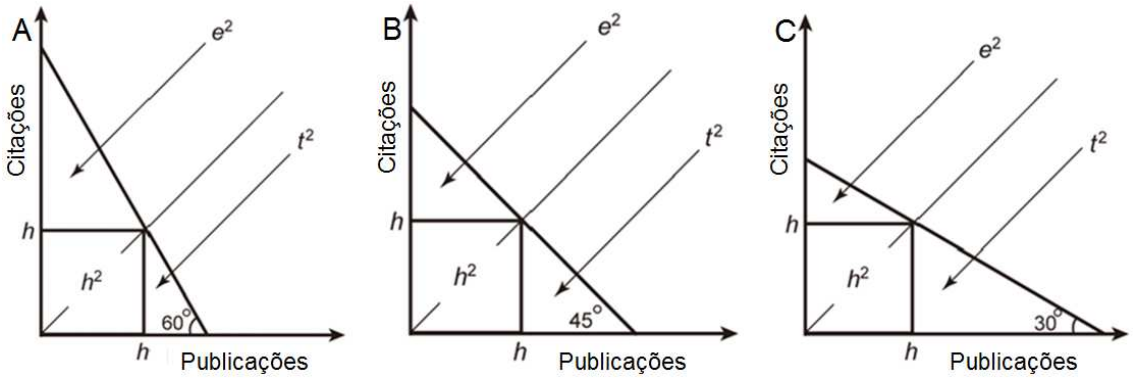


Figura 3.8: Curva de distribuição de citação, simplificada como uma linha reta (adaptado de (ZHANG, 2013-b)).

O cenário apresentado na Figura 3.8 evidencia uma das fragilidades do h-index, pois a métrica não considera dois importantes aspectos da produção de um pesquisador, os trabalhos altamente citados e os trabalhos com baixa citação. Mesmo sendo extremos

em relação ao equilíbrio da produção, tais elementos deveriam ser incorporados no cômputo da reputação de pesquisadores (ZHANG, 2013-b).

A definição do *h*'-Index é apresentada na Fórmula 12.

$$h' = rh = \frac{eh}{t} \tag{12}$$

Fonte: (ZHANG, 2013-b).

onde *e*, *h* e *t* são o e-index, o h-index e o t-index, respectivamente.

A Tabela 3.6 apresenta um exemplo de três cientistas que apresentam o mesmo valor de h-index (14). Além da identificação do h-index, a tabela mostra outros dados dos cientistas: o número de citações, os valores do e-index, do t-index e do r-index, bem como o resultado do *h*'-index. A ordem está pelo valor decrescente de h-index e, na sequência, pelo valor decrescente de *h*'-index.

Tabela 3.6: Comparação entre os índices e, t, r, h e h' de três pesquisadores que possuem h-index idênticos.

Pesquisador	Citações	e-index	t-index	r-index	h-index	h'-index
A	1321	32.91	6.23	5.23	14	73.19
B	408	12.61	7.28	1.73	14	24.25
C	592	7.69	18.37	0.42	14	5.86

Fonte: adaptado de (ZHANG, 2013-b).

No exemplo da Tabela 3.6 pode-se observar que mesmo os três pesquisadores apresentando valor igual de h-index (h = 14), o valor do *h*'-index é muito diferente, sendo 4,14 vezes maior do Pesquisador B para o Pesquisador C, 3,02 vezes maior do Pesquisador A para o Pesquisador B e 12,5 vezes maior do Pesquisador A para o Pesquisador C. Como as citações que não são envolvidas no cálculo do h-index, representadas no exemplo por e-index e por t-index, tem-se a impressão de que os três pesquisadores possuem a mesma reputação, uma vez que possuem o mesmo valor de h-index. No entanto, esta observação não é representativa, pois fica evidenciado que se forem utilizadas todas as citações dos pesquisadores, os resultados se modificam.

Apesar do *h*'-index melhorar alguns pontos do h-index, do g-index e das demais métricas relacionadas, a proposta mantém a visão de que pesquisadores devem ser avaliados exclusivamente pela sua produção bibliográfica e o impacto dessa produção.

3.2.9 O Hl-index: Melhoria do h-index Baseado em Qualidade de Citações de Artigos

No trabalho de Zhai, Yan e Zhu (2013) é proposto um novo índice denominado H_l-index que objetiva melhorar o h-index, proposto por Hirsch (2005). A proposta apresenta uma análise interessante do h-index e levanta uma questão pertinente, se todos os artigos que citam um artigo acadêmico contribuem igualmente para o impacto deste artigo na comunidade científica. O trabalho utiliza o h-index e o l-index[23] (KORN et al.,

[23] l-index é uma métrica que estima o impacto de uma publicação com base no impacto de outros artigos que citam a publicação.

2009) para melhorar a qualidade da reputação de pesquisadores. Na definição da métrica, o objetivo do H_l-index de um conjunto de artigos ser *h* significa que *h* é o maior número inteiro de tal modo que o conjunto de artigos tem, pelo menos, *h* artigos satisfazendo que o l-index não é menor que *h*.

A Figura 3.9 apresenta o exemplo de um pesquisador, alguns de seus artigos e algumas de suas citações. Como o H_i-index dá ênfase na importância das citações dos artigos, pois é fundamentado no l-index (KORN et al., 2009), a relevância da citação é importante para calcular o valor do índice do pesquisador.

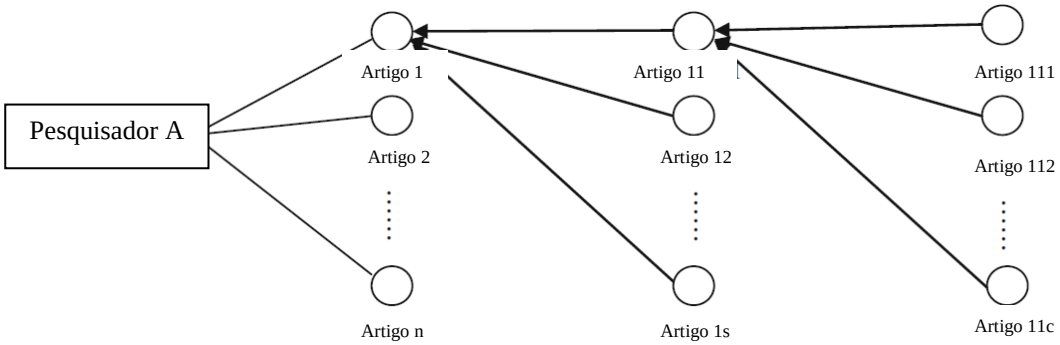


Figura 3.9: Exemplo dos artigos de um pesquisador e suas citações (adaptado de (ZHAI; YAN; ZHU, 2013)).

Na Tabela 3.7, apresenta-se um exemplo do cálculo do H_i-index de um pesquisador para um conjunto de artigos, suas citações e o valor do l-index de cada artigo.

Tabela 3.7: Exemplo do Hi-index de um pesquisador.

Número serial	Citações	l-index
1	20	10
2	23	8
3	15	7
4	10	5
5	10	4
6	5	2
7	2	2

h-index

h_i-index

Fonte: adaptado de (ZHAI; YAN; ZHU, 2013).

Na Tabela 3.7 são apresentados sete artigos. De acordo com o número de citações que cada artigo recebeu e o valor do l-index de cada artigo, o h-index para este pesquisador é igual a 5 (cinco), enquanto o valor do H_i-index é igual a 4 (quatro). Isso representa que apesar do valor do h-index ser mais elevado, a qualidade das citações dos artigos não foi confirmada, uma vez que o quarto artigo mais citado não atingiu o valor do l-index.

Apesar do H_i-index ser uma métrica interessante, pois está estruturada na qualidade das citações dos artigos, ela segue o mesmo padrão das demais métricas apresentadas, ou seja, entende que somente publicações de artigos e citações a esses artigos são suficientes para avaliar a reputação de um pesquisador.

3.2.10 Agregando Índices de Produtividade para Classificar Pesquisadores Através de Múltiplas Áreas

No trabalho de Lima et al. (2013) é apresentada uma abordagem para classificar pesquisadores de várias áreas de pesquisa. A métrica, denominada ca-index, tem como premissa que os índices de produtividade devem contabilizar as singularidades dos padrões de publicação de diferentes áreas de pesquisa, a fim de produzir uma avaliação imparcial do impacto da pesquisa científica. A métrica calcula o desempenho de um pesquisador em cada área levando em consideração o desempenho de outros pesquisadores da mesma área, de modo a garantir que a especificidade de cada área seja contabilizada em conformidade.

O trabalho apresenta que uma métrica para avaliar a produtividade de pesquisadores deve ter as seguintes propriedades:

- *Pluralidade* - a produtividade de um pesquisador deve ser avaliada em todas as áreas em que o pesquisador tem publicação;
- *Diversidade* - o perfil de cada área de pesquisa deve ser considerado quando se avalia a produtividade de um pesquisador;
- *Igualdade* - todas as áreas de pesquisa devem ser consideradas tão importantes e merecedoras de mérito científico.

A métrica ca-index apresenta que a pontuação atribuída a um pesquisador *i* é referente a uma área de pesquisa *a*. Assim, a pontuação é calculada somando-se a contribuição de cada uma das publicações do pesquisador *i* na área de pesquisa *a*. Para isto, cada publicação pode ser classificada em várias áreas, com base nas áreas de interesse do local da publicação. A definição do ca-index é apresentada na Fórmula 13.

$$s_i^a = \sum_{j=1}^{n_i} \mathbf{1}_j^a \frac{s_{i,j}}{m_{i,j}} \qquad (13)$$

Fonte: (LIMA et al., 2013).

onde n_i é o número total de publicações do pesquisador *i*, 1^a_j é uma função que atribui 1 se a publicação do pesquisador abrange uma área e 0 (zero) caso contrário, $S_{i,j}$ é a pontuação conferida por esta publicação para o pesquisador *i* e $M_{i,j}$ é o número total de áreas cobertas pela publicação.

A abordagem apresentada na proposta é interessante, pois transcende o impacto da produção científica de um pesquisador analisado de forma isolada. A inclusão do indicador de área na base de cálculo da reputação é um fator importante, pois permite uma análise mais detalhada do comportamento do pesquisador em todas as suas áreas de publicação. Apesar disso, a métrica também se baseia nas citações de artigos, como as demais métricas abordadas nesse trabalho e não incorpora outros elementos da carreira científica de um pesquisador, como orientações de mestrado e de doutorado, softwares desenvolvidos ou participações em comitês científicos.

3.3 Enquadramento da Tese em Comparação aos Trabalhos Relacionados

Foi apresentada nas seções anteriores deste capítulo uma visão geral das áreas de pesquisa em que esta tese está inserida e discutiu-se uma série de trabalhos relacionados. Esta seção mostra o enquadramento desta tese em comparação aos trabalhos relacionados, sendo reforçadas as novidades de suas contribuições

Os trabalhos relacionados abrangem as áreas de modelagem de perfil de pesquisadores e métricas para identificar reputação acadêmica. A análise dos trabalhos considerados relacionados envolve aspectos inerentes a produção científica de pesquisadores, a abrangência dessa produção, o impacto das citações da produção, se os trabalhos tem como premissa alguma métrica incorporada para gerar um índice de reputação, quais fontes de dados são utilizadas para a geração de informações sobre os pesquisadores, bem como que métodos estatísticos são usados para validação dos trabalhos.

Dentre os trabalhos relacionados no âmbito de perfil de pesquisadores, alguns trabalhos enfatizam a descoberta de especialistas em determinada área, como é o caso de Schoefegger (2011), Zhang, Tang e Li (2007), Punnarut e Sriharee (2010), Sugiyama e Kan (2010) e Fu et al. (2007). A descoberta de especialistas está muito relacionada com a reputação que um pesquisador possui em sua área de pesquisa. Assim, alguns trabalhos apresentam estudos que envolvem a área de manipulação de dados que formam o perfil de pesquisadores.

Para a representação do perfil, alguns trabalhos utilizam ontologias, como é o caso de Schoefegger (2011), Sriharee (2010), Punnarut e Sriharee (2010), Tang et al. (2008b), Tang, Zhang e Yao (2007), Zhang, Song e Song (2007). Uma ontologia é uma especificação explícita de uma conceitualização (GRUBER, 1995). Ela define os termos usados para descrever e representar uma área do conhecimento. Já outra forma de representação de perfil de usuários apresentada nos trabalhos relacionados é por meio de vetores de termos, como discutido em Sugiyama e Kan (2010), Sugiyama e Kan (2011), Kim et al. (2008) e Widyantoro et al. (1999). Um vetor de termos geralmente é definido por palavras-chave que representam as características do usuário para a composição de seu perfil.

Com o objetivo de proporcionar uma melhor visualização dos trabalhos relacionados no âmbito de modelagem de perfil de pesquisadores, a Tabela 3.8 apresenta uma classificação dos trabalhos analisados. Para essa classificação, foram definidos critérios de comparação julgados relevantes para o tema da tese, conforme descrito a seguir:

- Tipo de Detecção - Forma de detecção do perfil, podendo ser abordagem explícita, implícita ou híbrida. A abordagem explícita é quando existe intervenção do usuário. A abordagem implícita é quando o próprio sistema define o perfil, sem a participação direta do usuário. A abordagem híbrida é a união das duas formas;
- Tipo de Modelagem – critério de modelagem para a definição do perfil, podendo ser baseada em conhecimento (*knowledge-based*), em comportamento (*behavior-based*) ou híbrida, que é a utilização das duas abordagens;
- Forma de Representação – apresenta a forma como o perfil é representado ou armazenado;

- Métrica – identifica se o trabalho possui alguma métrica incorporada para identificar reputação ou qualquer outra característica avaliativa.

Tabela 3.8: Comparação entre os trabalhos relacionados no âmbito de modelagem de perfil de pesquisadores.

Trabalho	Tipo de Detecção	Tipo de Modelagem	Forma de Representação	Possui Métrica
(SCHOEFEGGER, 2011)	Implícita	Híbrida	Ontologia	Não
(SUGIYAMA; KAN, 2010)	Implícita	Comportamento	Vetor de termos	Não
(SUGIYAMA; KAN, 2011)	Implícita	Comportamento	Vetor de termos	Não
(PUNNARUT; SRIHAREE, 2010)	Implícita	Comportamento	Ontologia	Não
(KIM et al., 2008)	Implícita	Comportamento	Vetor de termos	Não
(TANG; ZHANG; YAO, 2007)	Implícita	Comportamento	Ontologia	Não
(ZHANG; TANG; LI, 2007)	Implícita	Conhecimento	Não apresenta	Não
(FU et al., 2007)	Híbrida	Comportamento	Não apresenta	Não
Tese	Implícita	Comportamento	Vetor de Termos	Sim

Dos 8 trabalhos relacionados estudados, 6 (seis) utilizam modelagem baseada em comportamento, 1 (um) modelagem baseada em conhecimento e 1 (um) modelagem híbrida. Ainda no aspecto de modelagem, 7 (sete) usam detecção implícita, ou seja, sem a intervenção do usuário e 1 (um) utiliza detecção híbrida, unindo a detecção implícita e a detecção explícita. No que tange o âmbito de representação do perfil, 3 (três) trabalhos utilizam ontologia, 3 (três) utilizam vetores de termos e 2 (dois) não apresentam qualquer forma de representação.

Levando-se em consideração o cenário apresentado nos trabalhos relacionados, nesta tese, o perfil de um pesquisador é modelado utilizando elementos inerentes a sua carreira científica. Esses elementos são definidos levando-se em consideração a abrangência de atuação de um pesquisador em diversos aspectos, como por exemplo: a produção em artigos, as orientações, a participação em bancas, a inserção em comitês científicos, dentre outros.

Outro elemento importante no estudo apresentado é com relação à incorporação de métricas. Nenhum trabalho estudado apresenta qualquer forma de avaliação, identificação de reputação ou análise de usuários. Assim, a abordagem proposta nesta tese vem apresentar alternativa mais completa, onde especifica um modelo de perfil de pesquisadores abrangente, adaptável, bem como define uma métrica para identificar reputação acadêmica.

Para o tema de métricas científicas no âmbito de identificação de reputação, os trabalhos analisados remetem a definição de um índice que determina o grau de reputação que um pesquisador possui. Grande parte dos trabalhos estudados geram esse índice com base na produção científica do pesquisador, mais especificamente no

impacto dessa produção. Assim, o elemento citação é um dos mais utilizados para medir reputação. A proposta de Hirsch (2005), denominada h-index, visa combinar critérios de qualidade e de quantidade, fornecendo uma métrica que mede o impacto de um pesquisador em um único número. Esse impacto é representado com base no número de artigos publicados pelo pesquisador e o número total de citações a seus artigos.

Após o surgimento do h-index, várias métricas foram propostas. Egghe (2006a) apresenta um novo índice, denominado g-index, que visa complementar o h-index. O g-index melhora o h-index, dando mais peso aos artigos altamente citados. Jin (2007) e Jin et al. (2007) propõe o AR-index que também visa complementar o h-index. A métrica incorpora aspectos temporais no cálculo da reputação, mais especificamente em relação ao ano de publicação dos artigos. O e-index, proposto por Zhang (2009) tem como objetivo suprir duas fragilidades do h-index: a perda de informação de citações e a baixa resolução. Além disso, afirma que o h-index tem um intervalo relativamente estreito, o que aumenta significativamente a ocorrência de um grupo de cientistas ter um h-index idêntico. Já a métrica proposta por Alonso et al. (2010), denominada hg-index, tem como objetivo caracterizar a produção científica de pesquisadores usando como base o h-index e o g-index, tentando manter as vantagens das duas propostas, bem como minimizar as suas desvantagens.

O *h*'-Index (ZHANG, 2013-b), também visa complementar o h-index, mas incorpora no cálculo da reputação outra métrica, o e-index (ZHANG, 2009). A nova métrica busca agregar no cômputo da reputação a quantidade de publicações de um pesquisador que ficam abaixo do h-core. Como o h-index, por si só, não leva em consideração o excesso de citações, estas características levam à perda de informação.

No trabalho de Lima et al. (2013) é apresentada uma abordagem para classificar pesquisadores de várias áreas de pesquisa. A métrica, denominada ca-index, tem como premissa que os índices de produtividade devem contabilizar as singularidades dos padrões de publicação de diferentes áreas de pesquisa, a fim de produzir uma avaliação imparcial do impacto da pesquisa científica. A inclusão do indicador de área na base de cálculo da reputação é um fator importante, pois permite uma análise mais detalhada do comportamento do pesquisador em todas as suas áreas de publicação. Apesar disso, a métrica também se baseia nas citações de artigos, como as demais métricas abordadas nesse trabalho e não incorpora outros elementos da carreira científica de um pesquisador, como orientações de mestrado e de doutorado, softwares desenvolvidos ou participações em comitês científicos.

Como o h-index foi a métrica que originou a maioria das demais métricas utilizadas atualmente, muitos dos pontos destacados nesta tese são referentes ao h-index. Apesar do h-index ser uma das métricas mais utilizadas pela comunidade acadêmica e científica, seus problemas são evidentes. O índice pode não envolver trabalhos que são pouco citados e subavaliados por indicadores bibliométricos. Mesmo assim, isto não indica que o trabalho não seja bom ou não represente uma demanda específica da sociedade, da indústria ou da academia (ZHANG, 2009). Ainda, o valor do h-index de um pesquisador nunca diminui (JIN, 2007). Mesmo que o pesquisador não tenha mais atividade na comunidade científica ou que diminua sua produção, seu h-index vai, no mínimo, permanecer o mesmo. O h-index também não deve ser utilizado para avaliar pesquisadores de áreas diferentes, pois cada área possui suas particularidades (BORNMANN; MUTZ; DANIEL, 2008; ALONSO et. al, 2009). Se forem comparados pesquisadores da área de Ciência da Computação com pesquisadores da área de Educação, as diferenças entre eles são muitas. A área de Computação tem por princípio

publicar trabalhos em conferências científicas, haja visto que até a Capes mantém um *Qualis* de conferências para atender as necessidades da comunidade. Já na área de Educação é comum o resultado das pesquisas serem publicados em livros ou capítulos de livros. Dessa forma, em uma avaliação de pesquisadores dessas duas áreas, os critérios de análise devem ser diferentes, sob pena de comprometer o processo e o resultado da avaliação.

O h-index e o g-index também podem ser influenciados por meio de autocitações, onde pesquisadores citam seus próprios trabalhos, muitas vezes para que possam repercutir na comunidade científica. Aliado a isso, não se pode usar o h-index e o g-index para avaliar pesquisadores em níveis diferentes na carreira, como um pesquisador júnior e um pesquisador sênior, pois eles são fortemente inclinados para pesquisadores mais experientes, com carreiras longas, onde os tempos de atuação são diferentes (MOED, 2009). Aliado a isso, com o avanço da ciência, o número de pesquisadores nas diversas áreas do conhecimento tem aumentado significativamente nos últimos anos. Isso ocasiona um grande número de trabalhos que necessitam encontrar um veículo de publicação para divulgação dos resultados obtidos. Muitas vezes, essa busca desenfreada por publicações pode colocar em dúvida a qualidade dos trabalhos que são publicados.

Embora o g-index apresente algumas vantagens em comparação ao h-index, algumas limitações também permanecem, principalmente a dificuldade de selecionar toda a produção dos pesquisadores e suas citações, a existência de diferentes tipos de documentos com diferentes impactos, o problema da autocitação e a incapacidade para comparar pesquisadores de diferentes áreas (VINKLER, 2007). Ainda, há uma limitação específica observada no g-index, a influência de um artigo altamente citado interferir no valor do índice, mesmo que isto não represente a média do desempenho de um pesquisador (COSTAS; BORDONS, 2008). O h-index e o g-index também utilizam apenas as publicações e o impacto dessas publicações para identificar a reputação de um pesquisador.

Da mesma forma que o h-index e o g-index, o AR-index, o e-index, o hg-index, o *h*'-index, o Hi-index são dependentes, exclusivamente, dos artigos publicados por um pesquisador e das citações a estes artigos. Essa característica faz com que as métricas sejam limitadas do ponto de vista da abrangência da trajetória de um pesquisador e da visão de especificidades das diversas áreas de pesquisa existentes.

Para possibilitar uma visualização mais qualificada dos trabalhos relacionados que envolvem métricas científicas, a Tabela 3.9 apresenta um estudo comparativo das principais características consideradas importantes quando se trabalha com reputação acadêmica. Para essa classificação, foram definidos critérios de comparação julgados relevantes para o tema da tese, conforme descrito a seguir:

- Ponderação – a métrica incorpora pesos nos indicadores que são utilizados para o cálculo da reputação;
- Tipo de Abrangência – refere-se a quantidade de indicadores que são utilizados para o cálculo da reputação. Métricas que se baseiam em publicações e citações são consideradas como de abrangência limitada. Já métricas que incorporam mais elementos da carreira de um pesquisador são consideradas como de abrangência ampla;

- Incorpora Impacto de Citações – se a métrica conta com algum elemento que representa o impacto de citações para o cálculo da reputação;
- Suporta Adaptabilidade – refere-se aos aspectos de poder oferecer o cálculo da reputação de forma seletiva, deixando a critério do usuário definir que indicadores deseja utilizar.

Tabela 3.9: Comparação entre os trabalhos relacionados no âmbito de métricas para reputação acadêmica.

Trabalho	Possui Ponderação	Tipo de Abrangência	Incorpora Impacto de Citações	Suporta Adaptabilidade
(LIMA et al., 2013)	Não	Limitada	Sim	Não
(ZHAI; YAN; ZHU, 2013)	Não	Limitada	Sim	Não
(ZHANG, 2013-b)	Não	Limitada	Sim	Não
(ALONSO et al., 2010)	Não	Limitada	Sim	Não
(ZHANG, 2009)	Não	Limitada	Sim	Não
ResearchGate	Não	Ampla	Sim	Não
(TANG et al., 2008b)	Não	Ampla	Sim	Não
(JIN, 2007; JIN et al., 2007)	Sim	Limitada	Sim	Não
(EGGHE, 2006a)	Não	Limitada	Sim	Não
(HIRSCH, 2005)	Não	Limitada	Sim	Não
Tese	Sim	Ampla	Sim	Sim

Dos 10 (dez) trabalhos relacionados com métricas para identificação de reputação acadêmica, 9 (nove) não utilizam pesos para a definição da reputação e apenas 1 (um) incorpora ponderação. Em relação ao tipo de abrangência, 8 (oito) são limitadas, pois utilizam quase que exclusivamente os indicadores de publicação e citação, ao passo que apenas 2 (duas) são consideradas métricas de abrangência ampla, pois envolvem outros indicadores. Referente a incorporação de impacto de citações, todas as métricas estudadas possuem esta característica implementada. Já em relação ao suporte a adaptabilidade, nenhuma métrica relacionada possui tal possibilidade.

Diante desse cenário, um dos pontos que foram analisados em vários dos trabalhos relacionados é que deve-se ter cuidado ao utilizar um indicador bibliométrico como o h-index ou qualquer outra métrica que descende dele de forma isolada. A utilização de um índice que utiliza o critério de citação como garantia de qualidade do trabalho de um cientista pode não representar a realidade. O ideal é encontrar um mecanismo que permita a avaliação do pesquisador de forma abrangente, levando em consideração os elementos de sua trajetória científica, construída ao longo de sua carreira, que é um dos objetivos desta tese.

Nesta tese, especifica-se um modelo de perfil de pesquisadores e uma métrica para medir reputação acadêmica. A modelagem do perfil envolve a identificação das informações relevantes da carreira do pesquisador. Tais informações são modeladas em categorias e elementos. As categorias, assim como os elementos, foram estruturados para suportarem ponderação por meio da definição de pesos. Já a reputação do pesquisador é definida por uma métrica que gera um índice calculado levando-se em consideração os elementos constantes no perfil do pesquisador que representam a produção científica desse pesquisador.

A abordagem proposta tem como premissa ser abrangente e adaptável, pois engloba a vida científica do pesquisador construída ao longo de sua carreira científica. Tais premissas permitem a utilização da abordagem em diferentes áreas e em diferentes contextos, uma vez que áreas diferentes usam critérios diferentes. Dessa forma, dependendo do propósito de utilização, ou seja, o que se quer avaliar, o que se quer medir e para que se quer medir, os critérios podem ser adaptados para que contemplem os requisitos do usuário e suas ponderações.

Essas características não foram encontradas em outras abordagens, pois não foi localizado na literatura um modelo abrangente que permite modelar o perfil de um pesquisador de forma a considerar toda sua produção científica, nem métricas que identifiquem a reputação desse pesquisador levando-se em consideração os elementos modelados em seu perfil.

4 ABORDAGEM PROPOSTA

Este capítulo apresenta a abordagem proposta nesta tese. O capítulo está dividido em três seções: a primeira apresenta uma visão geral da tese, a segunda especifica o modelo de perfil e a terceira apresenta a métrica proposta.

4.1 Visão Geral

Uma área de aplicação que vem crescendo significativamente é a colaboração científica entre pesquisadores. As redes de colaboração científica estão mudando a maneira de trabalhar com pesquisa e inovação. Cada vez mais pesquisadores expandem sua rede de coautoria em busca de melhorar a qualidade do seu trabalho. Assim, a identificação de pesquisadores com perfis semelhantes estimula o trabalho colaborativo e pode melhorar a qualidade das publicações científicas, proporcionar intercâmbio entre pesquisadores de diferentes países, ampliar a troca de experiências entre grupos de pesquisa, bem como promover a expansão das redes de colaboração científica.

No entanto, para que pesquisadores possam se aproximar e interagir, suas linhas de pesquisa e seus assuntos de interesse devem possuir afinidade. Assim, torna-se necessária a definição de um perfil para modelar a carreira científica de um pesquisador. Esse perfil deve ser representado pelos vários elementos que fazem parte da trajetória acadêmica de um pesquisador. Nesse sentido, a identificação de um perfil deve ser ampla e o processo deve envolver não apenas aspectos científicos, como publicações em periódicos ou livros publicados, mas também outros fatores inerentes à atividade de um pesquisador, como por exemplo: orientações que realizou, participação em bancas de defesa, publicações em conferências, participação em projetos de pesquisa e em comitês científicos, dentre outros.

Semelhante ao tema de perfil de pesquisadores, a área de métricas científicas tem um papel importante na comunidade acadêmica, pois pode auxiliar no processo de medição da qualidade da produção científica e na identificação de especialistas em determinada área de pesquisa. As métricas recentes se baseiam fortemente nas citações de artigos de pesquisadores, conforme detalhes apresentados nas Seções 2.3 e 3.2. Entretanto, essas formas de avaliação não consideram a trajetória do pesquisador, nem mesmo o cenário onde ele está inserido. As citações a artigos deveriam ser apenas um dos elementos a ser analisado e não o elemento único.

A abordagem proposta nesta tese consiste de um modelo de perfil, denominado Rep-Model e de uma métrica para identificar reputação acadêmica, denominada Rep-Index. As premissas para o desenvolvimento da proposta são que a abordagem deve ser abrangente e adaptável. O conceito de abrangência envolve a especificação de um modelo de perfil com diversos indicadores inerentes à atuação do pesquisador em sua

esfera de trabalho. Em relação à adaptabilidade, a ênfase é proporcionar uma utilização em diferentes áreas do conhecimento, bastando, para isso, que os usuários ajustem o modelo proposto de acordo com seus critérios e interesses.

Para a definição do Rep-Model buscou-se na literatura os principais elementos de trabalhos e sistemas que são utilizados no âmbito de perfil de pesquisadores. Assim, foi realizado um estudo exaustivo sobre os itens que caracterizam pesquisadores (conforme trabalhos relacionados) e, com base nisso, foram identificados 18 elementos mais relevantes. No capítulo dos experimentos foi realizada uma análise usando algoritmos de mineração para comprovar quais elementos de fato são relevantes.

A Figura 4.1 apresenta uma visão geral da tese, onde observa-se a sequência de sua estruturação, iniciando com a especificação do modelo de perfil (Rep-Model), a métrica para a geração da reputação (Rep-Index) e a avaliação experimental da abordagem proposta.

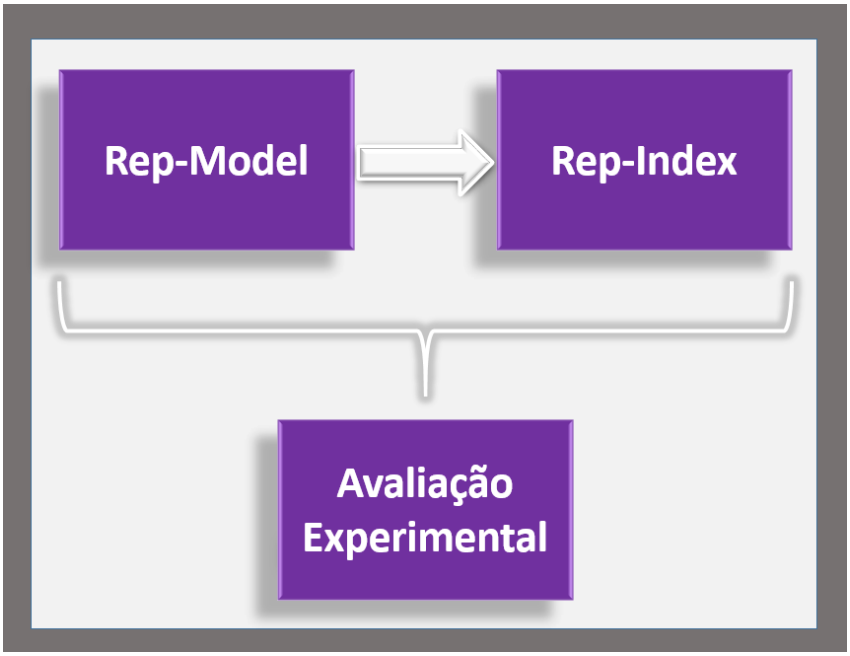


Figura 4.1: Visão geral da tese.

A seção 4.2 apresenta em detalhes a descrição do Rep-Model, especificando todos os elementos que compõem o modelo. A seção 4.3 apresenta a métrica para identificar a reputação de pesquisadores, envolvendo a ponderação entre os elementos do modelo, tendo como saída o valor do Rep-Index. Já na avaliação experimental são apresentados todos os testes realizados para avaliar o Rep-Model e o Rep-Index. A avaliação foi realizada utilizando-se de experimentos para comparar a correlação de Spearman entre os elementos do Rep-Model, a correlação de Spearman entre o Rep-Index, o h-index e o g-index, o Rep-Index com o ranking do CNPq, a aplicação de algoritmos de mineração para identificação dos elementos mais relevantes, bem como a utilização dos elementos mais relevantes para a geração do Rep-Index de pesquisadores.

4.2 Rep-Model - Modelo de Perfil de Pesquisadores

Visando a estruturação de um perfil abrangente, tendo em vista a possibilidade de utilização em diferentes áreas de pesquisa, foi especificado o Rep-Model (*Reputation Model* - Modelo de Reputação). O Rep-Model está fundamentado em elementos e

categorias, sendo definido levando-se em consideração aspectos inerentes à trajetória científica de um pesquisador, construída ao longo de sua carreira.

A Tabela 4.1 apresenta as categorias, os elementos e as siglas que definem o Rep-Model.

Tabela 4.1: Categorias, elementos e siglas do Rep-Model.

Categoria	Elemento	Sigla
Identificação (ID)	Nome	NM
	Instituição	INST
	Grau de Instrução	GI
Orientação (ORI)	Orientação de Mestrado	OM
	Orientação de Doutorado	OD
	Orientação de Pós-doutorado	OP
Banca (BAN)	Participação em Banca de Mestrado	PBM
	Participação em Banca de Doutorado	PBD
Comitê (COM)	Coordenação de Comitê de Conferência	CCC
	Membro de Comitê de Conferência	MCC
	Membro de Corpo Editorial de Periódico	MCEP
	Revisão de Periódico	RP
Publicação (PUB)	Artigo em Periódico	AP
	Capítulo de Livro	CLIV
	Livro	LIV
	Trabalho Completo em Conferência	TCC
	H-Index	HI
	Rede de Coautoria	RC
	Projeto de Pesquisa	PP
	Software	SOFT

Fonte: adaptado de (CERVI; GALANTE; OLIVEIRA, 2012).

Para uma melhor compreensão do Rep-Model, apresentamos uma descrição de todas as categorias e seus elementos:

- Identificação (ID) – visa fornecer a identificação do pesquisador. Seus três elementos são:
  - Nome (NM) – representa o nome do pesquisador;
  - Instituição (INST) – representa a instituição de vínculo do pesquisador;
  - Grau de Instrução (GI) – apresenta o grau de instrução do pesquisador, podendo representar sua formação (tecnólogo, licenciado ou bacharel), bem como sua titulação ou aperfeiçoamento (especialização, mestrado, doutorado, estágio pós doutoral). As opções deste elemento são flexíveis de acordo com o contexto de utilização, assim como os critérios definidos pelo usuário.
- Orientação (ORI) – representa as orientações do pesquisador e podem ser referentes a mestrado, doutorado ou pós-doutorado. Seus três elementos são:

  - Orientação de Mestrado (OM) – refere-se aos trabalhos de mestrado que o pesquisador orientou;
  - Orientação de Doutorado (OD) – representa as orientações de doutorado realizadas pelo pesquisador;
  - Orientação de Pós-doutorado (OP) – refere-se aos trabalhos de pós-doutorado que o pesquisador orientou.

- Banca (BAN) – refere-se a participação do pesquisador em bancas de trabalhos de conclusão em nível de pós-graduação *stricto sensu*. Seus dois elementos são:
  - Participação em Banca de Mestrado (PBM) – representa a participação do pesquisador em bancas de defesa de mestrado;
  - Participação em Banca de Doutorado (PBD) – refere-se a participação do pesquisador em bancas de defesa de doutorado.
- Comitê (COM) – representa a participação do pesquisador como membro de comitês científico. Seus quatro elementos são:
  - Coordenação de Comitê de Conferência (CCC) – refere-se às conferências onde o pesquisador participou como coordenador;
  - Membro de Comitê de Conferência (MCC) – representa as conferências onde o pesquisador teve participação;
  - Membro de Corpo Editorial de Periódico (MCEP) – refere-se aos periódicos onde o pesquisador pertence ao quadro de editores;
  - Revisão de Periódico (RP) – representa os periódicos onde o pesquisador faz parte do quadro de revisores de artigos.
- Publicação (PUB) – especifica as publicações do pesquisador e outros elementos relacionados à produção científica. Seus oito elementos são:
  - Artigo em Periódico (AP) – representa os artigos publicados em periódicos pelo pesquisador;
  - Capítulo de Livro (CLIV) – refere-se às publicações em nível de capítulo de livro do pesquisador;
  - Livro (LIV) – representa os livros publicados pelo pesquisador;
  - Trabalho Completo em Conferência (TCC) – refere-se aos trabalhos/artigos publicados em conferências científicas;
  - H-Index (HI) – refere-se ao impacto da produção de um pesquisador, calculada de acordo com (HIRSCH, 2005), representando as citações que receberam os trabalhos publicados pelo pesquisador;
  - Rede de Coautoria (RC) – representa o número de pesquisadores que possuem trabalhos publicados em conjunto com o pesquisador em questão;
  - Projeto de Pesquisa (PP) – refere-se aos projetos de pesquisa desenvolvidos pelo pesquisador;

- Software (SOFT) – representa os softwares em que o pesquisador teve participação no desenvolvimento.

A abordagem proposta no Rep-Model visa englobar a trajetória científica de um pesquisador, pois fornece subsídios para uma análise detalhada. Devido a abrangência do Rep-Model, que não enfoca somente em publicações ou citações, é possível analisar a evolução da trajetória do pesquisador, pois o modelo permite uma visão de vários elementos da carreira de um pesquisador. Essa solução pode envolver vários pontos de vista, quando os pesquisadores trabalham em diferentes áreas de pesquisa. Desta forma, o Rep-Model pode ser utilizado em diferentes áreas e em diferentes contextos, pois pode ser adaptado aos critérios do usuário.

A adaptabilidade do Rep-Model frente a diferentes áreas de pesquisa e diferentes critérios de utilização pode ser realizada levando-se em consideração as particularidades de cada área. Como exemplo podem ser citadas as áreas de Ciência da Computação, Educação e Odontologia. Na área de Ciência da Computação os pesquisadores valorizam muito as publicações dos resultados de suas pesquisas em conferências científicas. A vantagem desta estratégia é que conseguem ter um *feedback* mais rápido sobre a qualidade do trabalho, pois os artigos aceitos para publicação são avaliados pelos pares em um curto espaço de tempo, se comparado a publicações em periódicos. Inclusive a Capes mantém um *Qualis* exclusivo para conferências da área. Além disso, outro elemento muito relevante na área de Ciência da Computação é a produção de softwares. Por ser um elemento inerente à área de tecnologia, os softwares acabam sendo produtos muito desenvolvidos por pesquisadores, onde muitas vezes acabam gerando patentes com propriedade intelectual registrada.

Na área de Educação os pesquisadores possuem uma tendência a divulgar os resultados de suas pesquisas em publicações como livros e capítulos de livros. Isto pode ser explicado pela necessidade existente na área de que os trabalhos publicados devem aprofundar conceitos, fundamentos e teorias educacionais, bem como pela característica dos pesquisadores pelo preferência de textos mais longos para apresentação de seus resultados. Diferentemente das áreas de Ciência da Computação e Educação, a área de Odontologia prioriza artigos publicados em periódicos, pois não possui tradição em organizar conferências científicas para apresentação de resultados de trabalhos de pesquisa. Geralmente, as conferências da área são promovidas com fins comerciais, tendo apresentação de produtos, equipamentos, materiais dentários e inovações inerentes à área odontológica.

Visando contemplar a adaptabilidade frente às especificidades das diversas áreas de pesquisa, o Rep-Model incorpora uma abordagem baseada em ponderação. Nessa abordagem cada elemento do Rep-Model possui um peso associado. Esse peso pode ser modificado tendo em vista as necessidades do usuário, bem como as particularidades das áreas envolvidas e dos pesquisadores que estão sendo avaliados. Da mesma forma, se o usuário quiser neutralizar alguma categoria ou elemento, pois pode entender que não são relevantes para sua utilização, basta não considerar valor para o peso nos elementos que não serão considerados.

4.3 Rep-Index - Métrica Para Identificar Reputação Acadêmica

Com o objetivo de oportunizar uma forma abrangente e adaptável de identificar a reputação de pesquisadores, foi especificada a métrica denominada Rep-Index (*Reputation Index* – Índice de Reputação). O Rep-Index utiliza o Rep-Model, com suas

categorias e elementos. O objetivo do Rep-Index é identificar a reputação dos pesquisadores em níveis de reputação. Os níveis são identificados por um índice de valor inteiro e positivo que pode variar de 1 a 5. Esta normalização foi proposta com o intuito de não distanciar com valores numéricos pesquisadores iniciantes, intermediários e experientes, bem como tornar mais compreensível a identificação da reputação. A métrica para identificar a reputação é definida por meio de um somatório entre os vários elementos que compõem o modelo de perfil do pesquisador, definidos no Rep-Model. Para isto, são definidos pesos para os elementos do Rep-Model. O somatório dos pesos dos elementos é limitado ao valor 100.

A seguir apresenta-se as informações importantes sobre o Rep-Index para calcular a reputação de pesquisadores. Na sequência, a métrica é apresentada na Fórmula 14.

- R : Refere-se ao pesquisador que se quer identificar a reputação.
- c : Representa o número total de categorias.
- i : Representa o intervalo de 1 até o número total de categorias (c).
- e_i : Representa o número total de elementos.
- j : Refere-se ao intervalo de 1 até o número total de elementos (e_i).
- v : Representa o valor do elemento.
- w_j : Refere-se ao peso do elemento.
- $max(v_j)$: Representa o maior valor do elemento.

$$\mathtt{Rep-Index}_{(R)} = \sum_{i=1}^{c} \left(\sum_{j=1}^{e_i} \frac{(v_j \cdot w_j)}{max(v_j)} \right) \qquad (14)$$

Fonte: adaptado de (CERVI; GALANTE; OLIVEIRA, 2012).

O Rep-Index especifica a reputação de cada pesquisador por meio do somatório das cinco categorias do modelo: `ID`, `ORI`, `BAN`, `COM` e `PUB`. O resultado de cada categoria (`c`) é obtido pela multiplicação do valor de cada elemento (v_j) pelo peso do próprio elemento (w_j), dividido pelo valor mais elevado do elemento ($max(v_j)$). A Tabela 4.2 apresenta um exemplo de categorias, elementos, pesos e o maior valor de cada elemento do Rep-Model. Esta tabela também é utilizada no Capítulo 5 para apresentar os experimentos realizados.

O valor mais elevado de cada elemento ($max(v_j)$) tem o objetivo de oportunizar a avaliação da reputação usando grupos de pesquisadores. Assim, quando um grupo estiver sendo avaliado os parâmetros de comparação, em termos de quantidades, são inerentes ao próprio grupo, não tendo influência externa. No caso da avaliação ser individualizada, o conjunto de valores mais elevados dos elementos são os valores do próprio pesquisador.

Tabela 4.2: Categorias, elementos, pesos e o maior valor do elemento.

Categoria	Elemento (Sigla)	Peso	Maior Valor do Elemento
Identificação (ID)	NM	-	-
	INST	-	-
	GI	15	15
Orientação (ORI)	OM	4	116
	OD	5	59
	OP	6	13
Banca (BAN)	PBM	4	159
	PBD	6	95
Comitê (COM)	CCC	1	23
	MCC	1	57
	MCEP	5	5
	RP	3	3
Publicação (PUB)	AP	15	237
	CLIV	5	84
	LIV	7	39
	TCC	8	299
	HI	8	31
	RC	3	262
	PP	1	130
	SOFT	1	19
Total	-	100	-

Fonte: adaptado de (CERVI; GALANTE; OLIVEIRA, 2013-b).

Dado o maior valor para cada elemento, define-se o número de intervalos para a classificação dos pesquisadores nos níveis de reputação. Nos experimentos foram definidos cinco intervalos, que resultaram em 5 níveis de reputação. Para gerar o cálculo final da reputação de cada pesquisador, utilizamos a Fórmula 15.

$$\text{Rep - Index}_{(R)} = \begin{cases} 1 \ni \text{Rep - Index}_{(R)} \geq 0 \wedge < 20 \\ 2 \ni \text{Rep - Index}_{(R)} \geq 20 \wedge < 40 \\ 3 \ni \text{Rep - Index}_{(R)} \geq 40 \wedge < 60 \\ 4 \ni \text{Rep - Index}_{(R)} \geq 60 \wedge < 80 \\ 5 \ni \text{Rep - Index}_{(R)} \geq 80 \wedge \leq 100 \end{cases} \quad (15)$$

Fonte: adaptado de (CERVI; GALANTE; OLIVEIRA, 2012).

O índice de reputação (Rep-Index) de um pesquisador, conforme apresentado na Fórmula 15, classifica o pesquisador em cinco níveis de reputação. Se o valor do Rep-Index do pesquisador avaliado estiver entre 0 e 20 (exclusive), este pesquisador será classificado no nível 1. Caso seu Rep-Index seja igual ou maior que 20 e menor que 40, será classificado no nível 2. Se o valor do Rep-Index for maior ou igual a 40 e menor que 60, será classificado no nível 3. Se o pesquisador apresentar o resultado do Rep-Index maior ou igual a 60 e menor que 80, seu nível de reputação será 4. Já se o valor

do Rep-Index for maior ou igual a 80 e menor ou igual a 100, o nível de reputação será 5.

A faixa de valores do índice de reputação varia de 0 a 100, onde um valor próximo a 0 representa uma baixa reputação (característica de pesquisadores iniciantes) e um valor próximo a 100 representa uma alta reputação (característica de pesquisadores experientes). Dessa forma, os níveis de reputação de um pesquisador representam o quanto um pesquisador produziu ao longo de sua trajetória científica. Ao tempo em que um nível de reputação com valor 1 representa um pesquisador iniciante, os níveis 4 e 5 representam os pesquisadores que possuem alta produção científica.

5 AVALIAÇÃO EXPERIMENTAL

Este capítulo apresenta os experimentos realizados nesta tese com o objetivo de avaliar o Rep-Model e o Rep-Index. Para isso, apresenta-se uma introdução acerca do *baseline* utilizado na tese, onde a avaliação experimental foi conduzida identificando-se reputação individual de pesquisadores, a comparação do resultado do Rep-Index com as métricas h-index e g-index, bem como experimentos envolvendo diferentes cenários com os elementos mais relevantes do Rep-Model. Além disso, este capítulo, com o objetivo de proporcionar uma avaliação mais qualificada, apresenta a discussão sobre os resultados obtidos.

5.1 Visão Geral dos Experimentos

Nesta seção é apresentada uma visão geral acerca dos experimentos, iniciando com a definição da base de dados utilizada. Na sequência, as fontes de dados e as métricas de avaliação são apresentadas. Por fim, apresenta-se como se deu o processo de extração dos dados para serem utilizados nos experimentos.

A Figura 5.1 apresenta uma visão geral dos experimentos, onde a avaliação é realizada com testes que envolvem o modelo de perfil (Rep-Model) e a métrica para identificar reputação (Rep-Index).

Figura 5.1: Grupos de experimentos realizados.

5.1.1 Base de Dados

Os experimentos para identificação da reputação de pesquisadores foram realizados utilizando-se como *baseline* todos[24] os bolsistas de produtividade em pesquisa e

[24] O CNPq possui 15.088 bolsistas em todas as áreas do conhecimento (dados de outubro/2013).

tecnologia do CNPq das áreas de Ciência da Computação, Economia e Odontologia. Isso envolveu pesquisadores dos níveis 2, 1D, 1C, 1B e 1A, perfazendo um total de 830 bolsistas, sendo estes 404 da área de Ciência da Computação, 210 da área de Economia e 216 da área de Odontologia. Os níveis das bolsas seguem uma ordem de classificação e são diferenciados pelo valor que cada bolsista recebe, onde a entrada se dá pelo nível 2 e o topo são os pesquisadores do nível 1A.

O critério para a utilização destas áreas é que as mesmas não possuem relação ou características comuns. Tal critério é importante para validar as premissas da tese, de que a abordagem deve ser abrangente e adaptável. A abrangência se dá pelas diversidade de elementos para a definição da reputação, estruturada no Rep-Model, e a adaptabilidade representa a capacidade da abordagem suportar diferentes áreas com seus critérios específicos.

Os bolsistas do CNPq recebem bolsas de produtividade em pesquisa de acordo com sua produção, após análise de seu currículo pelos comitês de área, formados por especialistas. A classificação dos bolsistas se dá em cinco níveis, onde cada área possui critérios próprios para a classificação dos pesquisadores em cada nível. Os critérios das áreas de Ciência da Computação[25], Economia[26] e Odontologia[27] estão disponíveis para consulta no site do CNPq. Para apresentar uma visão de alguns critérios utilizados pelos comitês de área, descreve-se parte dos requisitos mínimos utilizados pela área de Ciência da Computação em relação a produção dos pesquisadores:

- Pesquisador Nível 1A: apresenta produção científica regular nos últimos 12 (doze) anos; tem um alto número de publicações qualificadas em periódicos e em conferências internacionais; contribui para o desenvolvimento de sua área em seu país; contribui para a articulação de grupos de pesquisa e formação de novos cientistas; tem liderança nacional e reconhecimento internacional com claros indicadores de contribuição às comunidades científicas nacional e internacional;
- Pesquisador Nível 1B: apresenta regular produção científica nos últimos 10 (dez) anos; tem publicações qualificadas em periódicos e em conferências internacionais; contribui para a formação de grupos de competência e possui reconhecimento nacional e internacional;
- Pesquisador Nível 1C: apresenta produção científica regular há pelo menos 8 (oito) anos; possui produção regular, notadamente em periódicos internacionais de bom nível; possui independência científica e inserção internacional, comprovada através de participação em comitês de programa internacionais e em programas de cooperação internacional; demonstra capacidade de captar recursos para pesquisa; tem orientado um número de dissertações de mestrado ou teses de doutorado compatível com seu tempo

[25] Critérios completos da área disponíveis em http://cnpq.br/web/guest/view/-/journal_content/56_INSTANCE_0oED/10157/49290

[26] Critério completos da área disponíveis em http://cnpq.br/web/guest/view/-/journal_content/56_INSTANCE_0oED/10157/50527

[27] Critérios completos da área disponíveis em http://cnpq.br/web/guest/view/-/journal_content/56_INSTANCE_0oED/10157/49701

de doutorado, quando vinculado a instituição que possua programas de pós-graduação;

- Pesquisador Nível 1D: apresenta produção científica regular há pelo menos 6 (seis) anos; possui publicações de nível internacional, várias em periódicos, com resultados obtidos após o trabalho de doutorado; tem orientado dissertações de mestrado ou teses de doutorado, quando vinculado a instituição que possua programas de pós-graduação;
- Pesquisador Nível 2: possui um histórico de publicações de nível internacional; demonstra independência, com resultados obtidos após o trabalho de doutorado; possui envolvimento em atividades de orientação de alunos de iniciação científica e de pós-graduação.

A avaliação experimental foi estruturada para que apresentasse a correlação entre a abordagem proposta na tese, de identificar a reputação de pesquisadores usando o modelo de perfil (Rep-Model) como premissa, e duas métricas conhecidas e consolidadas na comunidade acadêmica internacional, o h-index e o g-index. Para isso, comparou-se o resultado do Rep-Index com o resultado do h-index e do g-index em todos os experimentos realizados. A avaliação foi realizada com pesquisadores do CNPq, todos classificados como bolsistas de produtividade em pesquisa e tecnologia. A avaliação de bolsistas do CNPq tem sido utilizada em alguns trabalhos recentes, como apresentado em Oliveira et al. (2012) e Spilki (2013).

A Figura 5.2 apresenta a distribuição de quantidades dos bolsistas por níveis do CNPq em cada uma das três áreas de avaliação.

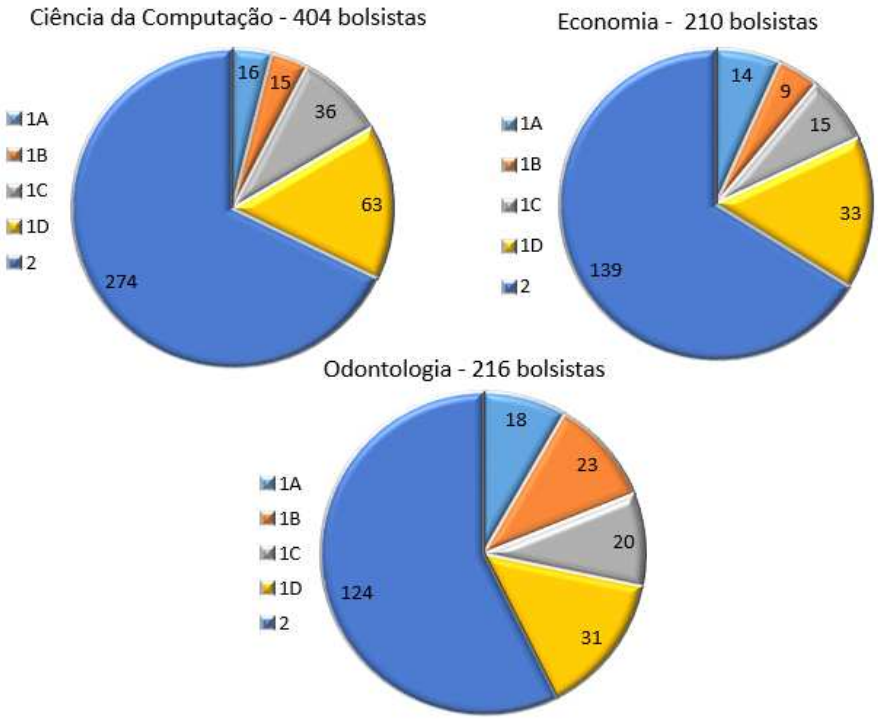


Figura 5.2: Distribuição dos bolsistas do CNPq, por níveis, nas áreas de Ciência da Computação, Economia e Odontologia[28].

[28] Dados coletados nos meses de janeiro, fevereiro e março de 2013.

Pode-se observar na Figura 5.2 que mais de 60% dos pesquisadores dessas três áreas do CNPq são classificados no nível 2. Isso demonstra que os critérios para concessão de bolsa são rígidos, pois poucos pesquisadores são classificados nas categorias do nível 1.

O objetivo de selecionar esse domínio de dados como objeto de comparação é que o CNPq possui rigorosos critérios para avaliar e definir quais pesquisadores possuem os requisitos para o recebimento da bolsa. Além disso, os seguintes aspectos foram importantes para a utilização dos dados de bolsistas do CNPq: (i) os dados permitem análise individual de pesquisadores; (ii) a análise atual dos dados é realizada de forma manual, assim, nossa abordagem pode apoiar a automatização do processo; (iii) todos os pesquisadores de uma área tem formação na mesma área; (iv) possibilidade de analisar a evolução dos bolsistas brasileiros; e (v) possibilidade de cruzar dados para suportar tomada de decisão.

5.1.2 Características das Fontes de Dados

O conjunto de dados dos 830 pesquisadores foram coletados de diversas fontes, todas com alto grau de confiabilidade: DBLP[29], Microsoft Academic Search[30], Arnetminer, Google Scholar[31], Publish or Perish (HARZING, 2007) e Plataforma Lattes[32]. Como algumas dessas fontes adotam mecanismos próprios para desambiguação de nomes, isto facilitou o processo de extração dos dados.

Destaca-se que das fontes de dados utilizadas nesta tese, as que mais apresentaram resultados de qualidade foram o Scholar Google e a DBLP. Apesar da limitação da DBLP, de reunir informações de somente pesquisadores da área de Ciência da Computação, ela foi muito útil pela confiabilidade dos dados que foram coletados. Além disso, quase 50% dos pesquisadores envolvidos nos experimentos são da área de Ciência da Computação. Em relação ao Scholar Google, este apresentou melhor resultado em comparação ao Microsoft Academic Search, pois proporcionou mais precisão na resposta, ao contrário do Microsoft Academic Search, que retornava dados duplicados e incompletos. O trabalho de Li et al. (2010) apresenta um estudo interessante onde compara algumas fontes de dados envolvendo publicações e citações de trabalhos científicos.

Em relação às áreas do CNPq envolvidas, foi implementado uma divisão nos experimentos para contemplar os testes individualizados por área, bem como para se ter uma avaliação geral. A separação por área teve como objetivo proporcionar uma avaliação individual dos grupos de pesquisadores.

5.1.3 Métricas e Ferramentas para Avaliação

A avaliação experimental foi estruturada para que apresentasse a correlação entre a abordagem proposta na tese, de identificar a reputação de pesquisadores usando o modelo de perfil como premissa, e duas métricas conhecidas e consolidadas na comunidade acadêmica internacional, o h-index e o g-index. Para isso, comparou-se o resultado da nossa abordagem com o resultado do h-index e do g-index em todos os

[29] http://www.informatik.uni-trier.de/~ley/db/

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

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

[32] http://lattes.cnpq.br/

experimentos realizados. Para analisar a existência de correlação entre os índices utilizamos o Coeficiente de Correlação de Postos de Spearman (ρ). Ele avalia como a relação entre duas variáveis pode ser descrita. Foi utilizado este método devido à heterogeneidade dos pesquisadores, pois apesar de representarem um grupo de excelência no país, existe muita variação de produção entre eles, o que faz com que método estatístico mais apropriado seja o de Spearman. O método também é menos sensível a *outliers*, em que pesquisadores se distanciam muito dos demais, em nível de produção científica. Aliado a isso, o Coeficiente de Spearman tem sido utilizado em diversos trabalhos envolvendo avaliação ou ranking de pesquisadores, como apresentado em Wainer e Vieira (2013), Zhai, Yan e Zhu (2013), Lopes et al. (2011), Franceschet e Costantini (2011), Waltman et al. (2011), Li et al. (2007), Li et al. (2010), Korn, Schubert e Telcs (2009), Van Raan (2006) e Rinia et al. (1998).

Com o objetivo de ampliar a avaliação experimental, foi realizado um conjunto de testes para identificar quais elementos do Rep-Model apresentaram resultados mais relevantes. Para a consecução dessa etapa foi utilizado o conceito de árvores de decisão por meio do algoritmo C4.5 (QUINLAN, 1993). Esse algoritmo possibilita encontrar relações entre elementos em um determinado conjunto de dados. Os experimentos utilizando o algoritmo C4.5 foram realizados na ferramenta Weka (HALL et al., 2009). A ferramenta Weka possui uma implementação de código aberto (*open source*) denominada J48 para o algoritmo C4.5. Na seção 5.2 são apresentados mais detalhes sobre o processo de realização dos experimentos e os cenários definidos para a identificação de quais elementos do Rep-Model apresentaram mais relevância.

5.1.4 Extração de Dados

Para a coleta de dados foi desenvolvida uma ferramenta com o objetivo de extrair dados da web do currículo Lattes dos pesquisadores. A ferramenta tem como entrada o nome do pesquisador e, na sequência, acessa o site da Plataforma Lattes desse pesquisador, retornando os dados do Lattes que estão representados no Rep-Model. Para a coleta dos dados do h-index e do g-index dos pesquisadores, o processo de extração se deu utilizando-se a ferramenta *Publish or Perish*.

A metodologia da extração dos dados se deu por área do CNPq, onde os pesquisadores foram agrupados nas áreas de Ciência da Computação, Economia e Odontologia. Primeiro foram coletados os dados de todos os pesquisadores e, após, esses pesquisadores foram classificados de acordo com cada nível do CNPq.

5.2 Experimentos do Rep-Model

Com o objetivo de analisar se todos os elementos do Rep-Model são relevantes para a identificação da reputação de pesquisadores, experimentos usando árvores de decisão foram realizados. Uma árvore de decisão é uma representação simples de um classificador e é utilizada por diversos sistemas de aprendizado de máquina, como o C4.5 (QUINLAN, 1993). O C4.5 é um algoritmo que constrói uma árvore de decisão necessitando um atributo que tem o maior poder de discriminação entre as classes envolvidas no processo. Essa característica possibilita ao algoritmo ser aplicado para redução de dimensionalidade, conforme apresentado por Blum e Langley (1997) e Guyon e Elisseeff (2003). Nesse sentido, o C4.5 é o mais recomendado para realizar os experimentos usando os bolsistas de produtividade em pesquisa e tecnologia do CNPq,

uma vez que é utilizado para verificar se os 18 (dezoito) elementos do Rep-Model são realmente relevantes.

A Figura 5.3 apresenta o grupo de experimentos do Rep-Model, que está definido da seguinte forma:

- Experimento 1 – análise dos 830 pesquisadores das três áreas de pesquisa (Ciência da Computação, Economia e Odontologia) usando todos para a construção de uma árvore de decisão;
- Experimento 2 – análise dos 830 pesquisadores das três áreas de pesquisa usando cada área individualmente para a construção das árvores de decisão, a saber:
    - Experimento 2.1 – executado com pesquisadores da Ciência da Computação;
    - Experimento 2.2 – executado com pesquisadores da Economia;
    - Experimento 2.3 – executado com pesquisadores da Odontologia.
- Experimento 3 – utilização dos elementos mais relevantes para a geração do Rep-Index dos 830 pesquisadores das três áreas de pesquisa.

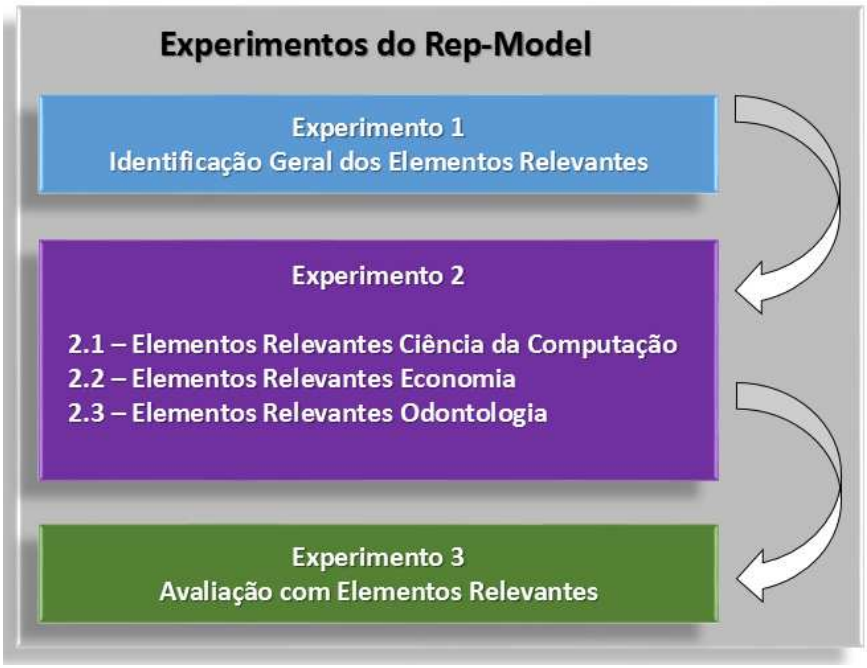


Figura 5.3: Distribuição dos bolsistas do CNPq, por níveis, nas áreas de Ciência da Computação, Economia e Odontologia.

Nos experimentos usando o algoritmo C4.5 foi utilizada uma implementação incorporada na ferramenta Weka (HALL et al., 2009), denominada J48. O objetivo é demonstrar, por meio dos experimentos, quais elementos do Rep-Model são mais indicados para a identificação da reputação de pesquisadores.

Para a execução dos experimentos foi necessário definir qual atributo deveria ser utilizado para rotular o nó atual da árvore. Tal atributo deve ser definido levando-se em consideração aquele que possui maior poder de discriminação entre todas as classes envolvidas no processo de construção da árvore de decisão. Definiu-se que o nível de

classificação do CNPq seria o atributo mais significativo para contemplar essa premissa, uma vez que possui classificação com critérios bem definidos.

5.2.1 Experimento 01 - Análise dos 830 Pesquisadores das Três Áreas de Pesquisa Usando Todos para a Construção da Árvore de Decisão

Para realizar os experimentos com os pesquisadores das três áreas de pesquisa foram utilizados diversos testes para verificar o maior grau de exatidão levando-se em consideração o número de classes utilizadas, bem como as características dos elementos do Rep-Model. Os experimentos foram executados usando a seguinte quantidade de instâncias por folha: M2 (padrão da ferramenta Weka), M3, M4, M5, M6, M7, M8, M9, M10, M11, M12, M13, M14, M15, M16, M17, M18, M19 e M20. Após análise de todos os testes, a quantidade de instâncias por folha que apresentou melhor resultado foi utilizando M10, onde obteve-se 65.1807% (numa escala de 1 a 100) de grau de exatidão.

A configuração utilizada no experimento é apresentada a seguir:

- Esquema: weka.classifiers.trees.J48 -C 0.25 -M10;
- Instâncias: 830 pesquisadores;
- Atributos: 19 (18 atributos do Rep-Model + 1 do nível do CNPq);
- Modo do teste: 10-fold cross-validation;

A Figura 5.4 apresenta a árvore construída pelo algoritmo J48 (C4.5) levando-se em consideração a configuração anteriormente descrita.

```
HI <= 11
|   TCC <= 147
|   |   AP <= 21: 5 (280.0/23.0)
|   |   AP > 21
|   |   |   CCC <= 3
|   |   |   |   CCC <= 0
|   |   |   |   |   TCC <= 18: 5 (88.0/15.0)
|   |   |   |   |   TCC > 18
|   |   |   |   |   |   HI <= 9: 5 (62.0/24.0)
|   |   |   |   |   |   HI > 9: 4 (17.0/6.0)
|   |   |   |   CCC > 0: 5 (97.0/24.0)
|   |   |   CCC > 3: 4 (17.0/9.0)
|   TCC > 147: 4 (14.0/7.0)
HI > 11
|   AP <= 31
|   |   AP <= 18: 5 (32.0/2.0)
|   |   AP > 18
|   |   |   RP <= 10
|   |   |   |   RC <= 70: 3 (19.0/11.0)
|   |   |   |   RC > 70: 4 (11.0/3.0)
|   |   |   RP > 10: 5 (20.0/7.0)
|   AP > 31
|   |   AP <= 198
|   |   |   RP <= 0: 1 (29.0/13.0)
|   |   |   RP > 0
|   |   |   |   HI <= 21
|   |   |   |   |   CCC <= 3
|   |   |   |   |   |   OD <= 13
|   |   |   |   |   |   |   TCC <= 0: 5 (24.0/8.0)
|   |   |   |   |   |   |   TCC > 0
|   |   |   |   |   |   |   |   OD <= 9
|   |   |   |   |   |   |   |   |   PP <= 18: 4 (18.0/7.0)
|   |   |   |   |   |   |   |   |   PP > 18: 5 (10.0/5.0)
|   |   |   |   |   |   |   |   OD > 9: 4 (11.0/5.0)
|   |   |   |   |   |   OD > 13
|   |   |   |   |   |   |   RC <= 107: 3 (18.0/9.0)
|   |   |   |   |   |   |   RC > 107: 2 (11.0/6.0)
|   |   |   |   |   CCC > 3: 4 (13.0/6.0)
|   |   |   |   HI > 21: 2 (16.0/10.0)
|   |   AP > 198
|   |   |   AP <= 251: 2 (10.0/4.0)
|   |   |   AP > 251: 1 (13.0/7.0)

Number of Leaves  :     22
Size of the tree :      43
```

Figura 5.4: Árvore gerada pelo algoritmo J48 (C4.5) com os elementos mais relevantes do Rep-Model dos 830 pesquisadores do CNPq das três áreas de pesquisa.

Na árvore gerada pelo algoritmo J48 (C4.5) pode-se observar que a classificação produzida manteve somente os elementos do Rep-Model que foram considerados relevantes e alinhados com a classificação nos níveis do CNPq, dentre eles: HI (H-Index), TCC (Trabalho Completo em Conferência), AP (Artigo em Periódico), CCC (Coordenação de Comitê de Conferência), RP (Revisão de Periódico), RC (Rede de Coautoria), OD (Orientação de Doutorado) e PP (Projeto de Pesquisa). Dos 18 (dezoito) elementos do Rep-Model, 8 (oito) foram considerados relevantes, tendo-se como cenário os dados dos 830 pesquisadores das três área de pesquisa

Levando-se em consideração o cenário apresentado no resultado do experimento, os elementos GI (Grau de Instrução), OP (Orientação de Pós-doutorado), PBM (Participação em Banca de Mestrado), PBD (Participação em Banca de Doutorado), MCC (Membro de Comitê de Conferência), MCEP (Membro de Corpo Editorial de Periódico), CLIV (Capítulo de Livro), LIV (Livro), OM (Orientação de Mestrado) e SOFT (Software) não foram considerados relevantes no processo de classificação dos pesquisadores nos níveis do CNPq. Isso representa 10 (dez) em um total de 18 (dezoito) elementos do Rep-Model.

Outra observação importante em relação à Figura 5.4 é que os números que aparecem entre parênteses (com destaque em círculo), após os nós folha, indicam o número de casos atribuídos a esse nó, seguido de quantos desses casos são incorretamente classificadas como resultado. Como exemplo, nesse experimento, em relação ao elemento AP (Artigo em Periódico), o algoritmo classificou corretamente 280 pesquisadores e 23 incorretamente.

A Figura 5.5 apresenta, de forma gráfica, a representação da árvore de decisão do experimento.

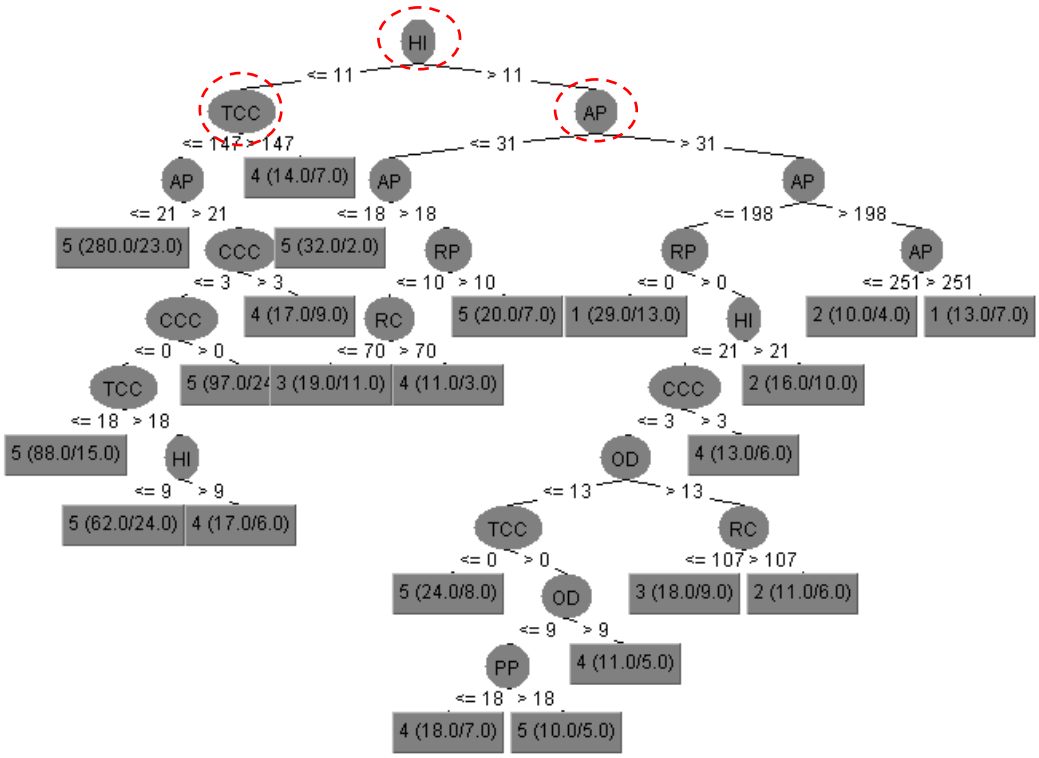
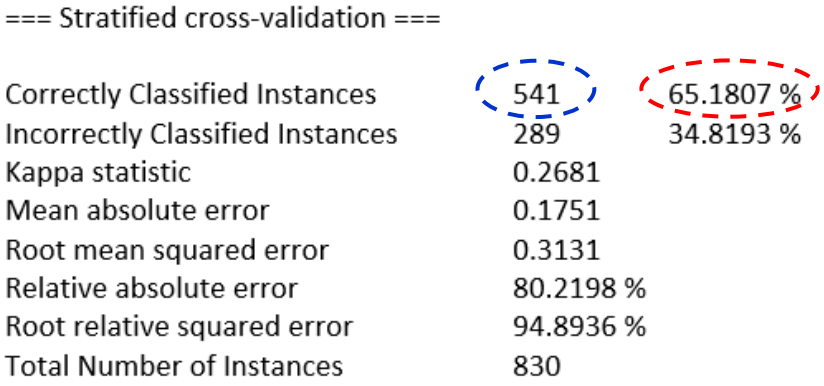


Figura 5.5: Visualização gráfica da árvore gerada para os elementos do Rep-Model de todos os 830 pesquisadores do CNPq das três áreas de pesquisa.

Pode-se observar na Figura 5.5 que o elemento mais relevante, levando-se em consideração todos os dados dos 830 pesquisadores, é o elemento HI (H-Index). Esse resultado é devido ao algoritmo J48 (C4.5) construir a árvore de decisão usando uma abordagem *top-down*, onde o atributo mais significativo, quando comparado aos outros atributos, é considerado a raiz da árvore. Na sequência do processo, o próximo nó da árvore é o segundo elemento mais significativo, e assim por diante. Nesse caso, o segundo e o terceiro elemento mais significativo são, respectivamente, TCC (Trabalho Completo em Conferência) e AP (Artigo em Periódico).

A Figura 5.6 apresenta o resultado da validação cruzada (*Stratified cross-validation*), onde pode ser visualizado alguns dados relevantes sobre o experimento.



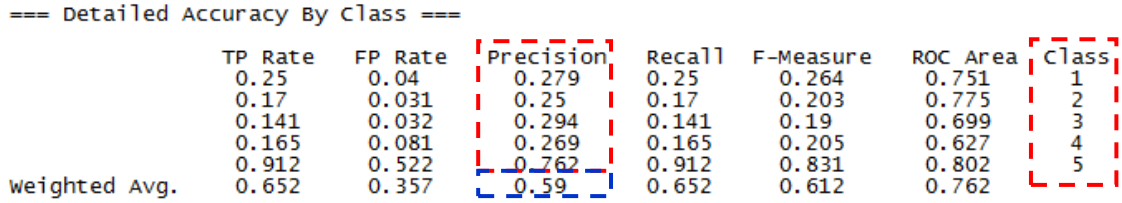
```
=== Stratified cross-validation ===

Correctly Classified Instances         541               65.1807 %
Incorrectly Classified Instances       289               34.8193 %
Kappa statistic                          0.2681
Mean absolute error                      0.1751
Root mean squared error                  0.3131
Relative absolute error                 80.2198 %
Root relative squared error             94.8936 %
Total Number of Instances              830
```

Figura 5.6: Resultado da validação cruzada do experimento com dados do Rep-Model dos 830 pesquisadores do CNPq das três áreas de pesquisa.

A Figura 5.6 apresenta o percentual de precisão do experimento realizado, atingindo 65.1807% (destaque em vermelho). Esse total representa que o algoritmo J48 (C4.5) conseguiu classificar corretamente 541 pesquisadores (destaque em azul), dos 830 considerados no teste. Isto representa que 289 pesquisadores não foram corretamente classificados.

Na Figura 5.7 pode ser visualizada a precisão (*Precision*) em relação aos resultados encontrados para a classificação dos pesquisadores de acordo com os níveis do CNPq. O resultado de cada nível é apresentado na coluna *Class* e corresponde ao nível 1A (*Class* 1), nível 1B (*Class* 2), nível 1C (*Class* 3), nível 1D (*Class* 4) e nível 2 (*Class* 5).



```
=== Detailed Accuracy By Class ===

                 TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                   0.25      0.04       0.279     0.25      0.264      0.751     1
                   0.17      0.031      0.25      0.17      0.203      0.775     2
                   0.141     0.032      0.294     0.141     0.19       0.699     3
                   0.165     0.081      0.269     0.165     0.205      0.627     4
                   0.912     0.522      0.762     0.912     0.831      0.802     5
Weighted Avg.      0.652     0.357      0.59      0.652     0.612      0.762
```

Figura 5.7: Resultado da precisão (Precision) encontrada para cada nível do CNPq (Class) dos 830 pesquisadores do CNPq das três áreas de pesquisa.

Os resultados apresentados na Figura 5.7 indicam que a classe (nível do CNPq) onde houve maior precisão foi a de número 5 (nível 2 do CNPq), alcançando 0.762. Já a classe onde houve menor precisão foi a de número 2 (nível 1B do CNPq), alcançando 0.25. Em relação a todas as classes do experimento, a média de precisão encontrada foi igual a 0.59.

Na Figura 5.8 está representada a matriz de confusão gerada pelos resultados do experimento. A matriz de confusão contém informações para o melhor entendimento do resultado do algoritmo, como a quantidade de instâncias classificadas corretamente e incorretamente, bem como a quantidade de instâncias que o algoritmo acreditava ser de um tipo, mas que foram classificadas como sendo de outro.

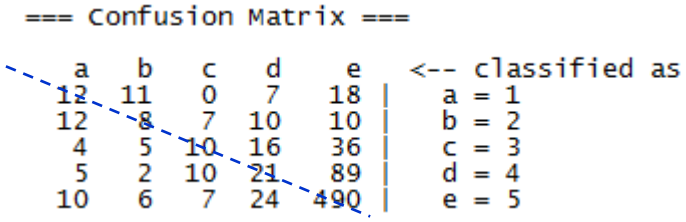


Figura 5.8: Matriz de confusão com resultados da classificação das instâncias dos 830 pesquisadores do CNPq das três áreas de pesquisa pelo algoritmo J48 (C4.5).

De acordo com os dados de saída apresentados na matriz de confusão da Figura 5.8, levando-se em consideração a classificação do CNPq dos 830 bolsistas, foram identificados os seguintes resultados:

- Dos 48 bolsistas do nível 1A – o algoritmo classificou 12 como sendo do nível 1A, 11 do nível 1B, nenhum do nível 1C, 7 do nível 1D e 18 do nível 2;
- Dos 47 bolsistas do nível 1B – o algoritmo classificou 8 como sendo do nível 1B, 12 do nível 1A, 7 do nível 1C, 10 do nível 1D e 10 do nível 2;
- Dos 71 bolsistas do nível 1C – o algoritmo classificou 10 como sendo do nível 1C, 4 do nível 1A, 5 do nível 1B, 16 do nível 1D e 36 do nível 2;
- Dos 127 bolsistas do nível 1D – o algoritmo classificou 24 como sendo do nível 1D, 5 do nível 1A, 2 do nível 1B, 10 do nível 1C e 89 do nível 2;
- Dos 537 bolsistas do nível 2 – o algoritmo classificou 490 como sendo do nível 2, 10 do nível 1A, 6 do nível 1B, 7 do nível 1C e 24 do nível D.

5.2.2 Experimento 2 – Análise dos 830 Pesquisadores das Três Áreas de Pesquisa Usando Cada Área Individualmente para a Construção da Árvore de Decisão

Esta seção apresenta o grupo composto por três experimentos distintos. O objetivo dos experimentos é analisar a relevância dos elementos do Rep-Model para as três áreas envolvidas. O Experimento 2.1 envolve pesquisadores da área de Ciência da Computação e, na sequência, os Experimentos 2.2 e 2.3, envolvem pesquisadores das áreas de Economia e Odontologia, respectivamente.

5.2.2.1 Experimento 2.1 – Pesquisadores da área de Ciência da Computação

Para realizar os experimentos com os 404 pesquisadores da área de Ciência da Computação foram utilizados diversos testes para verificar o maior grau de exatidão levando-se em consideração o número de classes utilizadas, bem como as características dos elementos do Rep-Model. Os experimentos foram executados usando a seguinte quantidade de instâncias por folha: M2 (padrão da ferramenta Weka), M3, M4, M5, M6, M7, M8, M9, M10, M11, M12, M13, M14, M15, M16, M17, M18, M19 e M20. Após análise de todos os testes, a quantidade de instâncias por folha que apresentou melhor

resultado foi utilizando M6, onde obteve-se 69.3069% (numa escala de 1 a 100) de grau de exatidão.

A configuração utilizada no experimento é apresentada a seguir:

- Esquema: weka.classifiers.trees.J48 -C 0.25 -M6;
- Instâncias: 404 pesquisadores;
- Atributos: 19 (18 atributos do Rep-Model + 1 do nível do CNPq);
- Modo do teste: 10-fold cross-validation.

A Figura 5.9 apresenta a árvore construída pelo algoritmo J48 (C4.5) levando-se em consideração a configuração anteriormente descrita.

```
AP <= 21: 5 (241.0/20.0)
AP > 21
|   MCEP <= 5
|   |   PBD <= 0
|   |   |   MCC <= 2: 1 (12.0/4.0)
|   |   |   MCC > 2: 3 (6.0/2.0)
|   |   PBD > 0
|   |   |   OD <= 2
|   |   |   |   LIV <= 1: 5 (9.0)
|   |   |   |   LIV > 1
|   |   |   |   |   AP <= 30
|   |   |   |   |   |   SOFT <= 4: 4 (10.0/4.0)
|   |   |   |   |   |   SOFT > 4: 5 (6.0)
|   |   |   |   |   AP > 30: 4 (6.0/2.0)
|   |   |   OD > 2
|   |   |   |   OM <= 57
|   |   |   |   |   PP <= 4: 3 (17.0/10.0)
|   |   |   |   |   PP > 4
|   |   |   |   |   |   AP <= 34
|   |   |   |   |   |   |   TCC <= 115
|   |   |   |   |   |   |   |   HI <= 8: 5 (10.0/3.0)
|   |   |   |   |   |   |   |   HI > 8
|   |   |   |   |   |   |   |   |   TCC <= 71: 4 (14.0/5.0)
|   |   |   |   |   |   |   |   |   TCC > 71: 5 (9.0/2.0)
|   |   |   |   |   |   |   TCC > 115: 4 (11.0/1.0)
|   |   |   |   |   |   AP > 34
|   |   |   |   |   |   |   RC <= 62
|   |   |   |   |   |   |   |   OD <= 7: 5 (7.0/3.0)
|   |   |   |   |   |   |   |   OD > 7: 4 (7.0/4.0)
|   |   |   |   |   |   |   RC > 62: 3 (16.0/6.0)
|   |   |   |   OM > 57: 2 (7.0/4.0)
|   MCEP > 5
|   |   GI <= 12: 2 (6.0/2.0)
|   |   GI > 12: 4 (10.0/4.0)

Number of Leaves  :     18
Size of the tree :      35
```

Figura 5.9: Árvore gerada pelo algoritmo J48 (C4.5) com os elementos mais relevantes do Rep-Model dos 404 pesquisadores do CNPq da área de Ciência da Computação.

Na árvore gerada pelo algoritmo J48 (C4.5) pode-se observar que a classificação produzida manteve somente os elementos do Rep-Model que foram considerados relevantes e alinhados com a classificação nos níveis do CNPq, dentre eles: AP (Artigo em Periódico), HI (H-Index), OM (Orientação de Mestrado), OD (Orientação de Doutorado), PBD (Participação em Banca de Doutorado), MCC (Membro de Comitê de Conferência), MCEP (Membro de Corpo Editorial de Periódico), LIV (Livro), TCC (Trabalho Completo em Conferência), RC (Rede de Coautoria), PP (Projeto de

Pesquisa), GI (Grau de Instrução) e SOFT (Software). Dos 18 (dezoito) elementos do Rep-Model, 13 (treze) foram considerados relevantes, tendo-se como cenário os dados dos 404 pesquisadores da área de Ciência da Computação.

Levando-se em consideração o cenário apresentado no resultado do experimento, os elementos CCC (Coordenação de Comitê de Conferência), OP (Orientação de Pós-doutorado), PBM (Participação em Banca de Mestrado), CLIV (Capítulo de Livro) e RP (Revisão de Periódico) não foram considerados relevantes no processo de classificação dos pesquisadores nos níveis do CNPq. Isso representa 5 (cinco) em um total de 18 (dezoito) elementos do Rep-Model.

Outra observação importante em relação à Figura 5.9 é que os números que aparecem entre parênteses (com destaque em círculo), após os nós folha, indicam o número de casos atribuídos a esse nó, seguido de quantos desses casos são incorretamente classificadas como resultado. Como exemplo, nesse experimento, em relação ao elemento AP (Artigo em Periódico), o algoritmo classificou corretamente 241 pesquisadores e apenas 20 incorretamente.

A Figura 5.10 apresenta, de forma gráfica, a representação da árvore de decisão gerada pelo experimento.

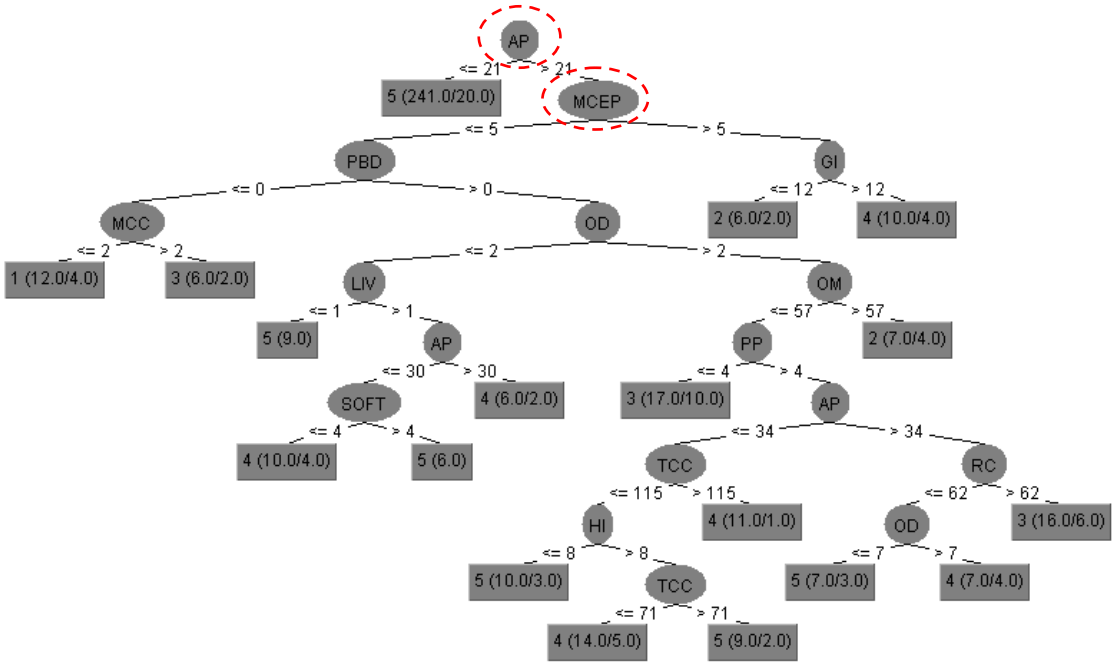


Figura 5.10: Visualização gráfica da árvore gerada para os elementos do Rep-Model dos 404 pesquisadores do CNPq da área de Ciência da Computação.

Pode-se observar na Figura 5.10 que o elemento mais relevante, levando-se em consideração todos os dados dos 830 pesquisadores, é o elemento AP (Artigo em Periódico). Esse resultado é devido ao algoritmo J48 (C4.5) construir a árvore de decisão usando uma abordagem *top-down*, onde o atributo mais significativo, quando comparado aos outros atributos, é considerado a raiz da árvore. Na sequência do processo, o próximo nó da árvore é o segundo elemento mais significativo, e assim por diante. Nesse caso, o segundo elemento mais significativo é MCEP (Membro de Corpo Editorial de Periódico).

A Figura 5.11 apresenta o resultado da validação cruzada (*Stratified cross-validation*), onde pode ser visualizado alguns dados relevantes sobre o experimento.

```
=== Stratified cross-validation ===

Correctly Classified Instances         280               69.3069 %
Incorrectly Classified Instances       124               30.6931 %
Kappa statistic                          0.3446
Mean absolute error                      0.1479
Root mean squared error                  0.301
Relative absolute error                 72.6806 %
Root relative squared error             94.6967 %
Total Number of Instances              404
```

Figura 5.11: Resultado da validação cruzada do experimento com dados do Rep-Model dos 404 pesquisadores do CNPq da área de Ciência da Computação.

A Figura 5.11 apresenta o percentual de precisão do experimento realizado, atingindo 69.3069% (destaque em vermelho). Esse total representa que o algoritmo J48 (C4.5) conseguiu classificar corretamente 280 pesquisadores (destaque em azul), dos 404 considerados no teste. Isto representa que 124 pesquisadores não foram corretamente classificados.

A Figura 5.12 ilustra a precisão (*Precision*) em relação aos resultados encontrados para a classificação dos pesquisadores de acordo com os níveis do CNPq. O resultado de cada nível é apresentado na coluna *Class* e corresponde ao nível 1A (*Class* 1), nível 1B (*Class* 2), nível 1C (*Class* 3), nível 1D (*Class* 4) e nível 2 (*Class* 5).

```
=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.438     0.028       0.389    0.438      0.412      0.716     1
                 0.267     0.018       0.364    0.267      0.308      0.771     2
                 0.278     0.054       0.333    0.278      0.303      0.65      3
                 0.254     0.082       0.364    0.254      0.299      0.644     4
                 0.887     0.446       0.807    0.887      0.845      0.806     5
Weighted Avg.    0.693     0.322       0.663    0.693      0.675      0.762
```

Figura 5.12: Resultado da precisão (Precision) encontrada para cada nível do CNPq (Class) da área de Ciência da Computação.

Os resultados apresentados na Figura 5.12 indicam que a classe (nível do CNPq) onde houve maior precisão foi a de número 5 (nível 2 do CNPq), alcançando 0.807. Já a classe onde houve menor precisão foi a de número 3 (nível 1C do CNPq), alcançando 0.333. Em relação a todas as classes do experimento, a média de precisão encontrada foi igual a 0.663.

Na Figura 5.13 está representada a matriz de confusão gerada pelos resultados do experimento. A matriz de confusão contém informações para o melhor entendimento do resultado do algoritmo, como a quantidade de instâncias classificadas corretamente e incorretamente, bem como a quantidade de instâncias que o algoritmo acreditava ser de um tipo, mas que foram classificadas como sendo de outro.

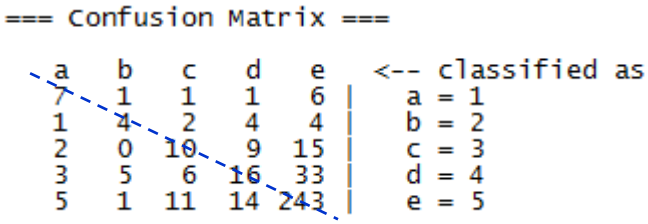


Figura 5.13: Matriz de confusão com resultados da classificação dos 404 pesquisadores do CNPq da área de Ciência da Computação pelo algoritmo J48 (C4.5).

De acordo com os dados de saída apresentados na matriz de confusão da Figura 5.13, levando-se em consideração a classificação do CNPq dos 404 bolsistas, foram identificados os seguintes resultados:

- Dos 16 bolsistas do nível 1A – o algoritmo classificou 7 como sendo do nível 1A, 1 do nível 1B, 1 do nível 1C, 1 do nível 1D e 6 do nível 2;
- Dos 19 bolsistas do nível 1B – o algoritmo classificou 4 como sendo do nível 1B, 1 do nível 1A, 2 do nível 1C, 4 do nível 1D e 4 do nível 2;
- Dos 36 bolsistas do nível 1C – o algoritmo classificou 10 como sendo do nível 1C, 2 do nível 1A, nenhum do nível 1B, 9 do nível 1D e 15 do nível 2;
- Dos 63 bolsistas do nível 1D – o algoritmo classificou 16 como sendo do nível 1D, 3 do nível 1A, 5 do nível 1B, 6 do nível 1C e 33 do nível 2;
- Dos 274 bolsistas do nível 2 – o algoritmo classificou 243 como sendo do nível 2, 5 do nível 1A, 1 do nível 1B, 11 do nível 1C e 14 do nível D.

5.2.2.2 Experimento 2.2 – Pesquisadores da área de Economia

Para realizar os experimentos com os 210 pesquisadores da área de Economia foram utilizados diversos testes para verificar o maior grau de exatidão levando-se em consideração o número de classes utilizadas, bem como as características dos elementos do Rep-Model. Os experimentos foram executados usando a seguinte quantidade de instâncias por folha: M2 (padrão da ferramenta Weka), M3, M4, M5, M6, M7, M8, M9, M10, M11, M12, M13, M14, M15, M16, M17, M18, M19 e M20. Após análise de todos os testes, a quantidade de instâncias por folha que apresentou melhor resultado foi utilizando M3, onde obteve-se 67.1429% (numa escala de 1 a 100) de grau de exatidão.

A configuração utilizada no experimento é apresentada a seguir:

- Esquema: weka.classifiers.trees.J48 -C 0.25 -M3;
- Instâncias: 210 pesquisadores;
- Atributos: 19 (18 atributos do Rep-Model + 1 do nível do CNPq);
- Modo do teste: 10-fold cross-validation;

A Figura 5.14 apresenta a árvore construída pelo algoritmo J48 (C4.5) levando-se em consideração a configuração acima descrita.

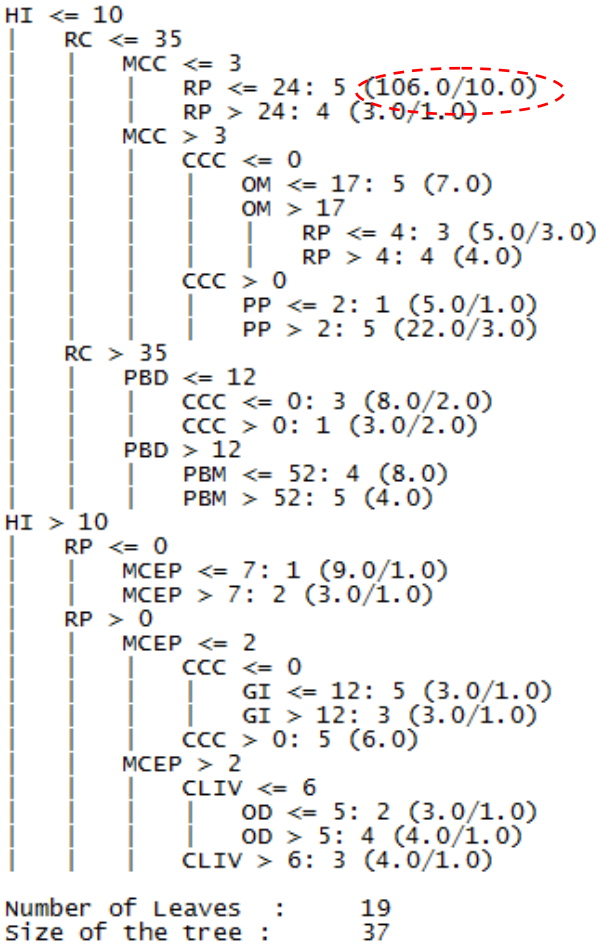


Figura 5.14: Árvore gerada pelo algoritmo J48 (C4.5) com os elementos mais relevantes do Rep-Model dos 210 pesquisadores do CNPq da área de Economia.

Na árvore gerada pelo algoritmo J48 (C4.5) pode-se observar que a classificação produzida manteve somente os elementos dos Rep-Model que foram considerados relevantes e alinhados com a classificação nos níveis do CNPq, dentre eles: HI (H-Index), RC (Rede de Coautoria), CCC (Coordenação de Comitê de Conferência), MCC (Membro de Comitê de Conferência), MCEP (Membro de Corpo Editorial de Periódico), CLIV (Capítulo de Livro), GI (Grau de Instrução), PBM (Participação em Banca de Mestrado), PBD (Participação em Banca de Doutorado), PP (Projeto de Pesquisa), OM (Orientação de Mestrado), RP (Revisão de Periódico) e OD (Orientação de Doutorado). Dos 18 (dezoito) elementos do Rep-Model 13 (treze) foram considerados relevantes, tendo-se como cenário os dados dos 210 pesquisadores da área de Economia.

Levando-se em consideração o cenário apresentado no resultado do experimento, os elementos OP (Orientação de Pós-doutorado), LIV (Livro), TCC (Trabalho Completo em Conferência), AP (Artigo em Periódico) e SOFT (Software) não foram considerados relevantes no processo de classificação dos pesquisadores nos níveis do CNPq. Isso representa 5 (cinco) em um total de 18 (dezoito) elementos do Rep-Model.

Outra observação importante em relação à Figura 5.14 é que os números que aparecem entre parênteses (com destaque em vermelho), após os nós folha, indicam o número de casos atribuídos a esse nó, seguido de quantos desses casos são

incorretamente classificadas como resultado. Como exemplo, nesse experimento, em relação ao elemento RP (Revisão de Periódico), o algoritmo classificou corretamente 106 pesquisadores e 10 incorretamente.

A Figura 5.15 apresenta, de forma gráfica, a representação da árvore do experimento.

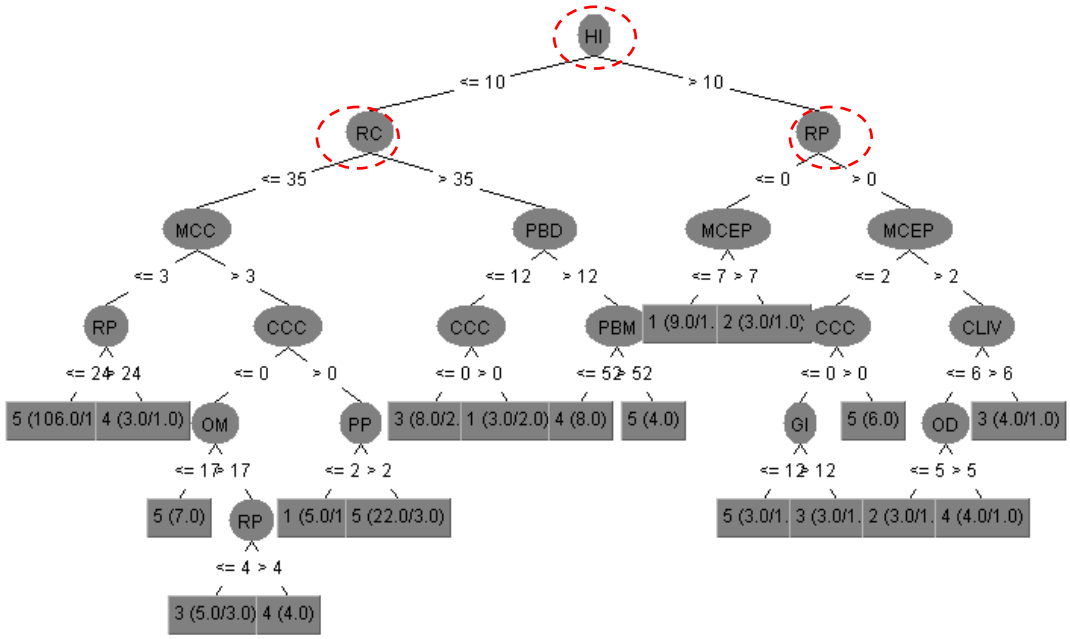


Figura 5.15: Visualização gráfica da árvore gerada para os elementos do Rep-Model dos 210 pesquisadores do CNPq da área de Economia.

Pode-se observar na Figura 5.15 que o elemento mais relevante, levando-se em consideração todos os dados dos 210 pesquisadores, é o elemento HI (H-Index). Esse resultado é devido ao algoritmo J48 (C4.5) construir a árvore de decisão usando uma abordagem *top-down*, onde o atributo mais significativo, quando comparado aos outros atributos, é considerado a raiz da árvore. Na sequência do processo, o próximo nó da árvore é o segundo elemento mais significativo, e assim por diante. Nesse caso, o segundo e o terceiro elemento mais significativo são, respectivamente, RC (Rede de Coautoria) e RP (Revisão de Periódico).

A Figura 5.16 apresenta o resultado da validação cruzada (*Stratified cross-validation*), onde pode ser visualizado alguns dados relevantes sobre o experimento.

```
=== Stratified cross-validation ===

Correctly Classified Instances         141               67.1429 %
Incorrectly Classified Instances        69               32.8571 %
Kappa statistic                          0.3292
Mean absolute error                      0.1517
Root mean squared error                  0.3301
Relative absolute error                 71.1437 %
Root relative squared error            101.7161 %
Total Number of Instances              210
```

Figura 5.16: Resultado da validação cruzada do experimento com dados do Rep-Model dos 210 pesquisadores do CNPq da área de Economia.

A Figura 5.16 apresenta o percentual de precisão do experimento realizado, atingindo 67.1429% (destaque em vermelho). Esse total representa que o algoritmo J48 (C4.5) conseguiu classificar corretamente 141 pesquisadores (destaque em azul), dos 210 considerados no teste. Isto representa que 69 pesquisadores não foram corretamente classificados.

A Figura 5.17 ilustra a precisão (*Precision*) em relação aos resultados encontrados para a classificação dos pesquisadores de acordo com os níveis do CNPq. O resultado de cada nível é apresentado na coluna *Class* e corresponde ao nível 1A (*Class* 1), nível 1B (*Class* 2), nível 1C (*Class* 3), nível 1D (*Class* 4) e nível 2 (*Class* 5).

```
=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.429    0.051      0.375     0.429     0.4        0.795     1
                 0.111    0.025      0.167     0.111     0.133      0.459     2
                 0.533    0.046      0.471     0.533     0.5        0.721     3
                 0.121    0.073      0.235     0.121     0.16       0.551     4
                 0.878    0.451      0.792     0.878     0.833      0.764     5
Weighted Avg.    0.671    0.318      0.627     0.671     0.644      0.717
```

Figura 5.17: Resultado da precisão (Precision) encontrada para cada nível do CNPq (Class) da área de Economia.

Os resultados apresentados na Figura 5.17 indicam que a classe (nível do CNPq) onde houve maior precisão foi a de número 5 (nível 2 do CNPq), alcançando 0.792. Já a classe onde houve menor precisão foi a de número 2 (nível 1B do CNPq), alcançando 0.167. Em relação a todas as classes do experimento, a média de precisão encontrada foi igual a 0.627.

Na Figura 5.18 está representada a matriz de confusão gerada pelos resultados do experimento. A matriz de confusão contém informações para o melhor entendimento do resultado do algoritmo, como a quantidade de instâncias classificadas corretamente e incorretamente, bem como a quantidade de instâncias que o algoritmo acreditava ser de um tipo, mas que foram classificadas como sendo de outro.

```
=== Confusion Matrix ===

  a   b   c   d   e   <-- classified as
  6   1   0   2   5 |   a = 1
  3   1   1   2   2 |   b = 2
  2   0   8   2   3 |   c = 3
  2   1   4   4  22 |   d = 4
  3   3   4   7 122 |   e = 5
```

Figura 5.18: Matriz de confusão com resultados da classificação dos 210 pesquisadores do CNPq da área de Economia pelo algoritmo J48 (C4.5).

De acordo com os dados de saída apresentados na matriz de confusão da Figura 5.18, levando-se em consideração a classificação do CNPq dos 210 bolsistas, identificamos os seguintes resultados:

- Dos 14 bolsistas do nível 1A – o algoritmo classificou 6 como sendo do nível 1A, 1 do nível 1B, nenhum do nível 1C, 2 do nível 1D e 5 do nível 2;
- Dos 9 bolsistas do nível 1B – o algoritmo classificou 1 como sendo do nível 1B, 3 do nível 1A, 1 do nível 1C, 2 do nível 1D e 2 do nível 2;
- Dos 15 bolsistas do nível 1C – o algoritmo classificou 8 como sendo do nível 1C, 2 do nível 1A, nenhum do nível 1B, 2 do nível 1D e 3 do nível 2;

- Dos 33 bolsistas do nível 1D – o algoritmo classificou 4 como sendo do nível 1D, 2 do nível 1A, 1 do nível 1B, 4 do nível 1C e 22 do nível 2;
- Dos 139 bolsistas do nível 2 – o algoritmo classificou 122 como sendo do nível 2, 3 do nível 1A, 3 do nível 1B, 4 do nível 1C e 7 do nível D.

5.2.2.3 Experimento 2.3 – Pesquisadores da área de Odontologia

Para realizar os experimentos com os 216 pesquisadores da área de Odontologia foram utilizados diversos testes para verificar o maior grau de exatidão levando-se em consideração o número de classes utilizadas, bem como as características dos elementos do Rep-Model. Os experimentos foram executados usando a seguinte quantidade de instâncias por folha: M2 (padrão da ferramenta Weka), M3, M4, M5, M6, M7, M8, M9, M10, M11, M12, M13, M14, M15, M16, M17, M18, M19 e M20. Após análise de todos os testes, a quantidade de instâncias por folha que apresentou melhor resultado foi utilizando M9, onde obteve-se 61.5741% (numa escala de 1 a 100) de grau de exatidão.

A configuração utilizada no experimento é apresentada a seguir:

- Esquema: weka.classifiers.trees.J48 -C 0.25 -M9;
- Instâncias: 216 pesquisadores;
- Atributos: 19 (18 atributos do Rep-Model + 1 do nível do CNPq);
- Modo do teste: 10-fold cross-validation.

A Figura 5.19 apresenta a árvore construída pelo algoritmo J48 (C4.5) levando-se em consideração a configuração anteriormente descrita.

```
HI <= 11: 5 (118.0/20.0)
HI > 11
|   CCC <= 0
|   |   TCC <= 0: 5 (20.0/8.0)
|   |   TCC > 0: 4 (24.0/10.0)
|   CCC > 0
|   |   AP <= 111: 5 (11.0/3.0)
|   |   AP > 111
|   |   |   PP <= 25
|   |   |   |   AP <= 188
|   |   |   |   |   RC <= 117: 3 (9.0/3.0)
|   |   |   |   |   RC > 117: 2 (9.0/4.0)
|   |   |   |   AP > 188: 1 (15.0/5.0)
|   |   |   PP > 25: 2 (10.0/4.0)

Number of Leaves  :     8
Size of the tree :      15
```

Figura 5.19: Árvore gerada pelo algoritmo J48 (C4.5) com os elementos mais relevantes do Rep-Model dos 216 pesquisadores do CNPq da área de Odontologia.

Na árvore gerada pelo algoritmo J48 (C4.5) pode-se observar que a classificação produzida manteve somente os elementos dos Rep-Model que foram considerados relevantes e alinhados com a classificação nos níveis do CNPq, dentre eles: HI (H-Index), CCC (Coordenação de Comitê de Conferência), TCC (Trabalho Completo em Conferência), AP (Artigo em Periódico), PP (Projeto de Pesquisa) e RC (Rede de Coautoria). Dos 18 (dezoito) elementos do Rep-Model 6 (seis) foram considerados relevantes, tendo-se como cenário os dados dos 216 pesquisadores da área de Odontologia.

Levando-se em consideração o cenário apresentado no resultado do experimento, os elementos GI (Grau de Instrução), OP (Orientação de Pós-doutorado), PBM (Participação em Banca de Mestrado), PBD (Participação em Banca de Doutorado), MCC (Membro de Comitê de Conferência), MCEP (Membro de Corpo Editorial de Periódico), CLIV (Capítulo de Livro), LIV (Livro), OM (Orientação de Mestrado), RP (Revisão de Periódico), OD (Orientação de Doutorado) e SOFT (Software) não foram considerados relevantes no processo de classificação dos pesquisadores nos níveis do CNPq. Isso representa 12 (doze) em um total de 18 (dezoito) elementos do Rep-Model.

Outra observação importante em relação à Figura 5.30 é que os números que aparecem entre parênteses (com destaque em vermelho), após os nós folha, indicam o número de casos atribuídos a esse nó, seguido de quantos desses casos são incorretamente classificadas como resultado. Como exemplo, nesse experimento, em relação ao elemento HI (H-Index), o algoritmo classificou corretamente 118 pesquisadores e 20 incorretamente.

A Figura 5.20 ilustra a representação da árvore do experimento.

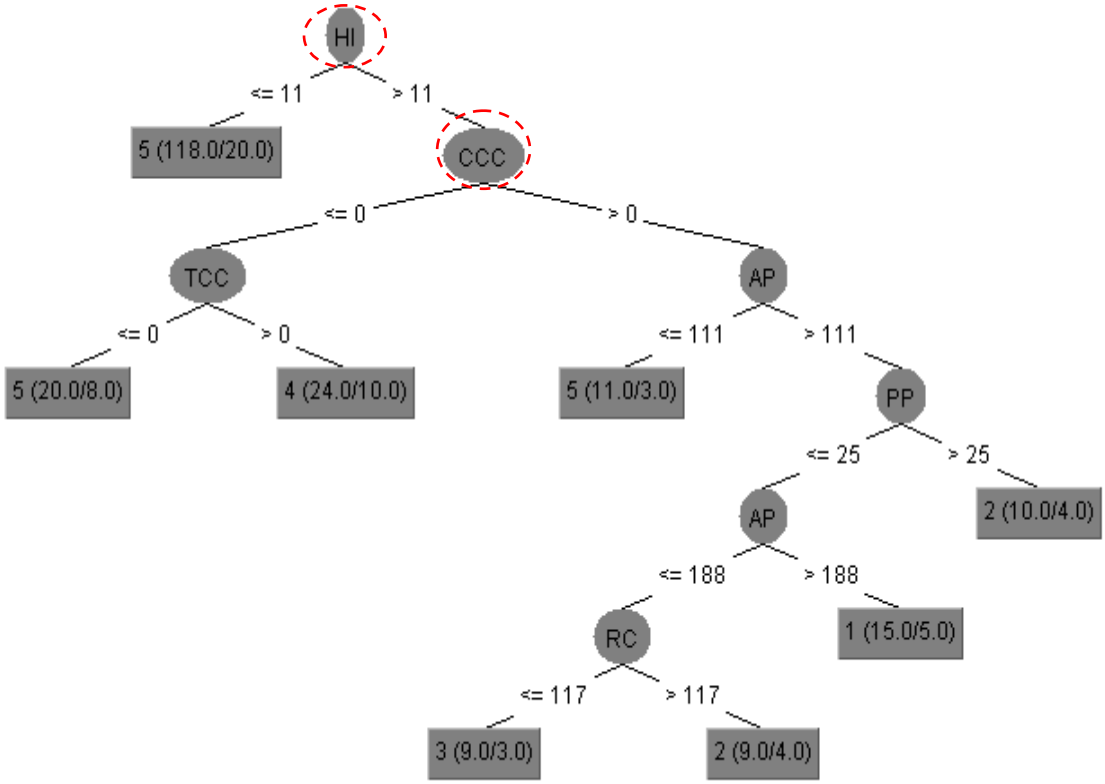


Figura 5.20: Visualização gráfica da árvore gerada para os elementos do Rep-Model dos 216 pesquisadores do CNPq da área de Odontologia.

Pode-se observar na Figura 5.20 que o elemento mais relevante, levando-se em consideração todos os dados dos 216 pesquisadores, é o elemento HI (H-Index). Esse resultado é devido ao algoritmo J48 (C4.5) construir a árvore de decisão usando uma abordagem *top-down*, onde o atributo mais significativo, quando comparado aos outros atributos, é considerado a raiz da árvore. Na sequência do processo, o próximo nó da árvore é o segundo elemento mais significativo, e assim por diante. Nesse caso, o segundo elemento mais significativo é o CCC (Coordenação de Comitê de Conferência).

A Figura 5.21 apresenta o resultado da validação cruzada (*Stratified cross-validation*), onde podem ser visualizados alguns dados relevantes sobre o experimento.

```
=== Stratified cross-validation ===

Correctly Classified Instances         133               61.5741 %
Incorrectly Classified Instances        83               38.4259 %
Kappa statistic                          0.3207
Mean absolute error                      0.1925
Root mean squared error                  0.3322
Relative absolute error                 76.7022 %
Root relative squared error             94.0945 %
Total Number of Instances              216
```

Figura 5.21: Resultado da validação cruzada do experimento com dados do Rep-Model dos 216 pesquisadores do CNPq da área de Odontologia.

A Figura 5.21 apresenta o percentual de precisão do experimento realizado, atingindo 61.5741% (destaque em vermelho). Esse total representa que o algoritmo J48 (C4.5) conseguiu classificar corretamente 133pesquisadores (destaque em azul), dos 216 considerados no teste. Isto representa que 83 pesquisadores não foram corretamente classificados.

A Figura 5.22 ilustra a precisão (*Precision*) em relação aos resultados encontrados para a classificação dos pesquisadores de acordo com os níveis do CNPq. O resultado de cada nível é apresentado na coluna *Class* e corresponde ao nível 1A (*Class* 1), nível 1B (*Class* 2), nível 1C (*Class* 3), nível 1D (*Class* 4) e nível 2 (*Class* 5).

```
=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   ROC Area  Class
                 0.333    0.071      0.3       0.333     0.316      0.796     1
                 0.174    0.083      0.2       0.174     0.186      0.711     2
                 0.05     0.02       0.2       0.05      0.08       0.539     3
                 0.323    0.054      0.5       0.323     0.392      0.682     4
                 0.903    0.424      0.742     0.903     0.815      0.804     5
Weighted Avg.    0.616    0.268      0.562     0.616     0.577      0.751
```

Figura 5.22: Resultado da precisão (Precision) encontrada para cada nível do CNPq (Class) dos 216 pesquisadores da área de Odontologia.

Os resultados apresentados na Figura 5.22 indicam que a classe (nível do CNPq) onde houve maior precisão foi a de número 5 (nível 2 do CNPq), alcançando 0.742. Já em relação ao resultado onde houve menor precisão, ocorreu um empate. As classes número 2 (nível 1B do CNPq) e número 3 (nível 1C do CNPq) alcançaram 0.2. Em relação a todas as classes do experimento, a média de precisão encontrada foi igual a 0.562.

Na Figura 5.23 está representada a matriz de confusão gerada pelos resultados do experimento. A matriz de confusão contém informações para o melhor entendimento do resultado do algoritmo, como a quantidade de instâncias classificadas corretamente e incorretamente, bem como a quantidade de instâncias que o algoritmo acreditava ser de um tipo, mas que foram classificadas como sendo de outro.

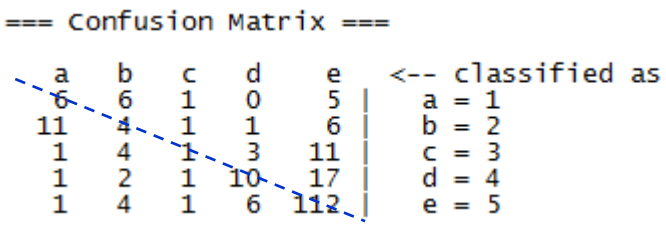


Figura 5.23: Matriz de confusão com resultados da classificação dos 216 pesquisadores do CNPq da área de Odontologia pelo algoritmo J48 (C4.5).

De acordo com os dados de saída apresentados na matriz de confusão da Figura 5.23, levando-se em consideração a classificação do CNPq dos 216 bolsistas, foram identificados os seguintes resultados:

- Dos 18 bolsistas do nível 1A – o algoritmo classificou 6 como sendo do nível 1A, 6 do nível 1B, 1 do nível 1C, nenhum do nível 1D e 5 do nível 2;
- Dos 23 bolsistas do nível 1B – o algoritmo classificou 4 como sendo do nível 1B, 11 do nível 1A, 1 do nível 1C, 1 do nível 1D e 6 do nível 2;
- Dos 20 bolsistas do nível 1C – o algoritmo classificou 1 como sendo do nível 1C, 1 do nível 1A, 4 do nível 1B, 3 do nível 1D e 11 do nível 2;
- Dos 31 bolsistas do nível 1D – o algoritmo classificou 10 como sendo do nível 1D, 1 do nível 1A, 2 do nível 1B, 1 do nível 1C e 17 do nível 2;
- Dos 124 bolsistas do nível 2 – o algoritmo classificou 112 como sendo do nível 2, 1 do nível 1A, 4 do nível 1B, 1 do nível 1C e 6 do nível D.

5.2.3 Experimento 3 – Utilização dos Elementos Mais Relevantes para a Geração do Rep-Index Usando Cada Área Individualmente

O objetivo desse experimento é verificar a equivalência entre o resultado do Rep-Index e a classificação do CNPq nas áreas de Ciência da Computação, Economia e Odontologia, analisando o Rep-Index dos pesquisadores pela média de cada área individualmente. Para a geração do Rep-Index utilizamos uma avaliação seletiva, ou seja, utilizamos no Rep-Model somente os elementos que foram identificados como mais relevantes para a identificação da reputação, mediante os experimentos com o algoritmo J48 (C4.5) apresentados nas seções 5.2.2.1, 5.2.2.2 e 5.2.2.3. Os resultados dos experimentos são apresentados nas Figuras 5.24, 5.25 e 5.26.

A Figura 5.24 apresenta os resultados da área de Ciência da Computação levando-se em consideração a média da própria área. Observa-se que o Rep-Index dos pesquisadores seguiu a classificação do CNPq, tendo apenas uma inversão entre os pesquisadores dos níveis 1A e 1B. Os pesquisadores do nível 1B apresentaram resultado um pouco superior aos pesquisadores do nível 1A. Rep-Index 3 (3,26) para os do nível 1B e Rep-Index 3 (3,14) para os do nível 1A. Isso representa uma compatibilidade de 80% entre os resultados do Rep-Index e a classificação do CNPq para a área de Ciência da Computação.

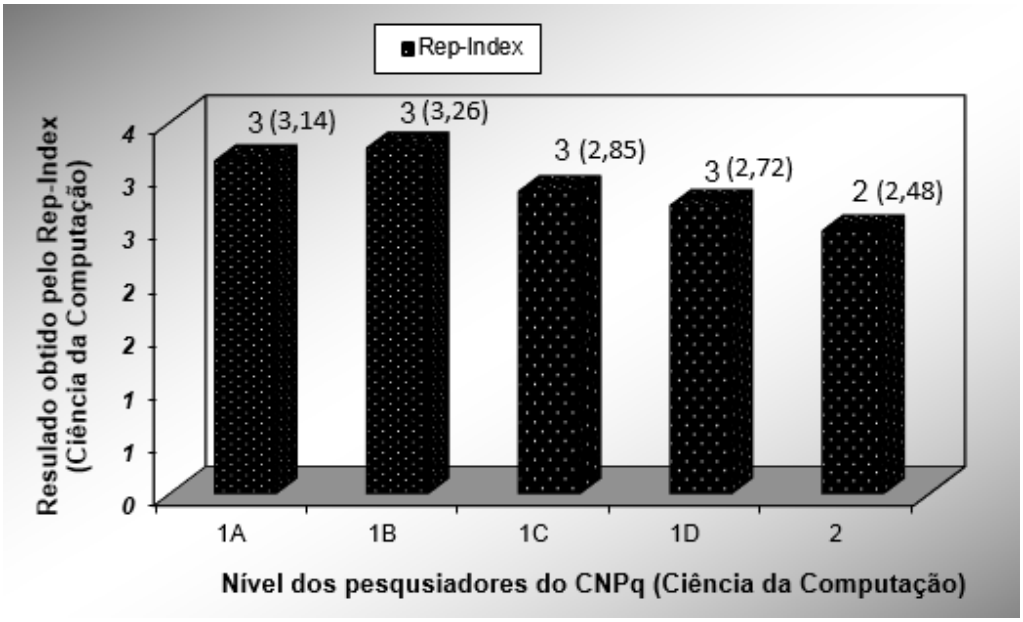


Figura 5.24: Rep-Index dos pesquisadores da área de Ciência da Computação vs classificação do CNPq da área de Ciência da Computação.

Em relação ao valor do Rep-Index usando somente os elementos do Rep-Model considerados mais relevantes, houve uma elevação no índice em relação ao experimento anterior. A Tabela 5.1 apresenta esse resultado de forma comparativa, onde percebe-se que todos os pesquisadores foram classificados em um nível superior (de 2 para 3). A única exceção foram os pesquisadores do nível 2, que permaneceram no mesmo nível.

Tabela 5.1: Comparação do Rep-Index dos pesquisadores do CNPq da área de Ciência da Computação usando todos os elementos do Rep-Model e usando somente os elementos identificados pelo algoritmo J48 (C4.5).

Nível do CNPq	Usando todos[33] os elementos do Rep-Model	Usando somente os elementos relevantes do Rep-Model
1A	2 (2,19)	3 (3,14)
1B	2 (2,27)	3 (3,26)
1C	2 (1,97)	3 (2,85)
1D	2 (1,94)	3 (2,72)
2	2 (1,59)	2 (2,48)

A Figura 5.25 apresenta os resultados da área de Economia levando-se em consideração a média da própria área. Observa-se que o Rep-Index dos pesquisadores seguiu exatamente a classificação do CNPq, não havendo nenhuma inversão, diferentemente quando ocorreu no experimento utilizando todos os elementos do Rep-Model. Isso representa uma compatibilidade de 100% entre os resultados do Rep-Index e a classificação do CNPq para a área de Economia.

[33] Dados do experimento realizado na seção 5.3.2 (Figura 5.32).

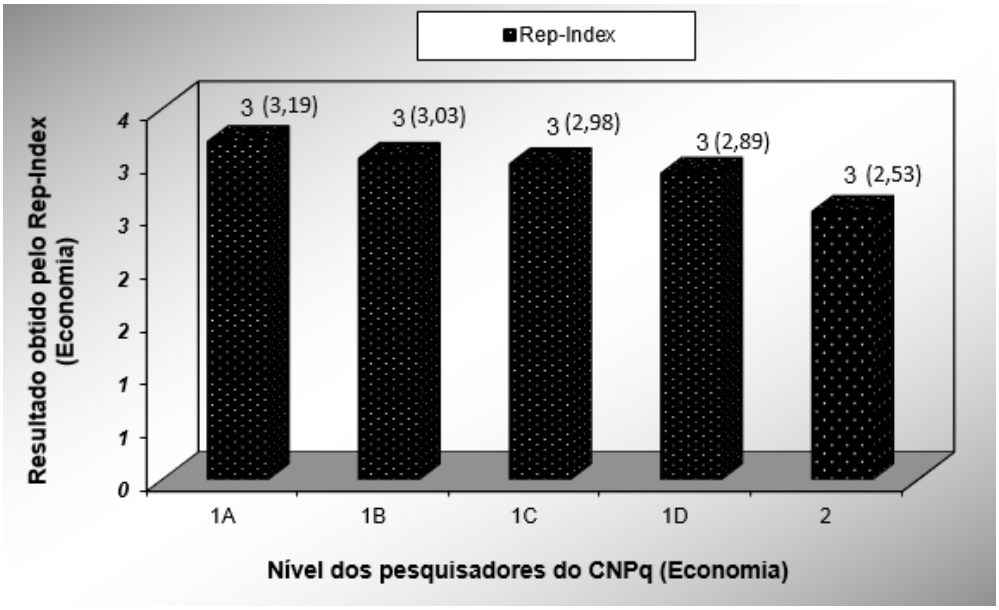


Figura 5.25: Rep-Index dos pesquisadores da área de Economia vs classificação do CNPq da área de Economia.

Em relação ao valor do Rep-Index usando somente os elementos do Rep-Model considerados mais relevantes, houve uma elevação no índice em relação ao experimento anterior. A Tabela 5.2 apresenta esse resultado de forma comparativa, onde percebe-se que todos os pesquisadores foram classificados em um nível superior (de 2 para 3).

Tabela 5.2: Comparação do Rep-Index dos pesquisadores do CNPq da área de Economia usando todos os elementos do Rep-Model e usando somente os elementos identificados pelo algoritmo J48 (C4.5).

Nível do CNPq	Usando todos[34] os elementos do Rep-Model	Usando somente os elementos relevantes do Rep-Model
1A	(2) 2,14	(3) 3,19
1B	(2) 2,00	(3) 3,03
1C	(2) 2,13	(3) 2,98
1D	(2) 2,00	(3) 2,89
2	(2) 1,65	(3) 2,53

Uma observação importante a ser feita é em relação a nova classificação. No experimento anterior, usando todos os elementos do Rep-Model, havia inversão entre os pesquisadores dos níveis 1C e 1B, onde os do nível 1C apresentavam valor maior para o Rep-Index. Já no experimento usando somente os elementos relevantes do Rep-Model, o resultado foi diferente, ou seja, a classificação do CNPq se manteve a mesma.

A Figura 5.26 mostra o resultado do Rep-Index dos pesquisadores da área de Odontologia segundo a classificação do CNPq, considerada a média da própria área de Odontologia. Observa-se que o Rep-Index dos pesquisadores seguiu exatamente a classificação do CNPq, não havendo nenhuma inversão, diferentemente quando ocorreu no experimento utilizando todos os elementos do Rep-Model. Isso representa uma

[34] Dados do experimento realizado na seção 5.3.2 (Figura 5.33).

compatibilidade de 100% entre os resultados do Rep-Index e a classificação do CNPq para a área de Economia.

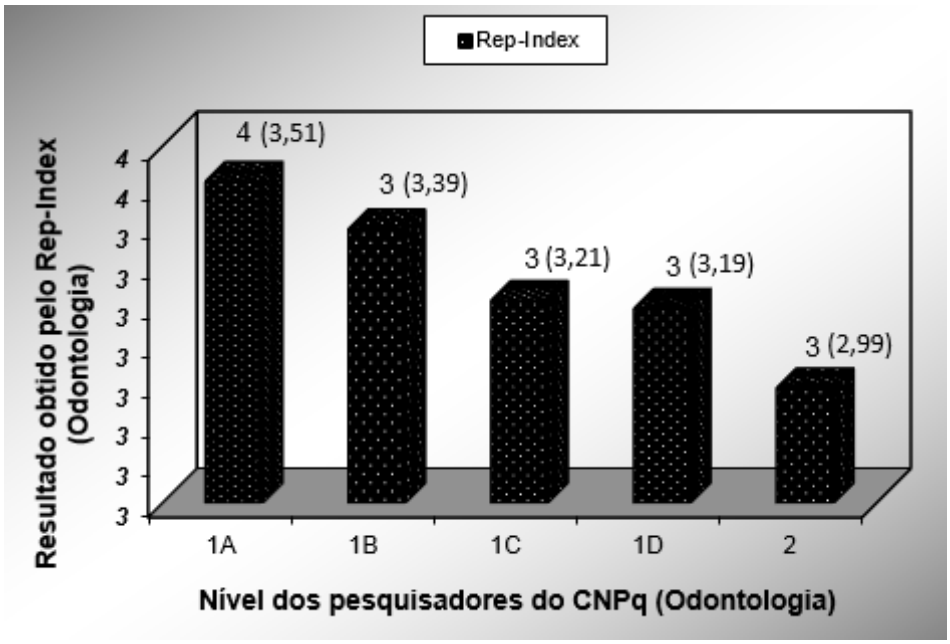


Figura 5.26: Rep-Index dos pesquisadores da área de Odontologia vs classificação do CNPq da área de Odontologia.

Em relação ao valor do Rep-Index usando somente os elementos do Rep-Model considerados mais relevantes, houve uma elevação no índice em relação ao experimento anterior. A Tabela 5.3 apresenta esse resultado de forma comparativa, onde percebe-se que os pesquisadores dos níveis 1B, 1C, 1D e 2 foram classificados em um nível superior (de 2 para 3) e os pesquisadores do nível 1A foram classificados com Rep-Index 4, aumentando 2 posições (de 2 para 4).

Tabela 5.3: Comparação do Rep-Index dos pesquisadores do CNPq da área de Odontologia usando todos os elementos do Rep-Model e usando somente os elementos identificados pelo algoritmo J48 (C4.5).

Nível do CNPq	Usando todos[35] os elementos do Rep-Model	Usando somente os elementos relevantes do Rep-Model
1A	(2) 2,44	(4) 3,51
1B	(2) 2,35	(3) 3,39
1C	(2) 2,05	(3) 3,21
1D	(2) 2,10	(3) 3,19
2	(2) 1,93	(3) 2,99

Uma observação importante a ser feita é em relação a nova classificação. No experimento anterior, usando todos os elementos do Rep-Model, havia inversão entre os pesquisadores dos níveis 1D e 1C, onde os do nível 1D apresentavam valor maior para o Rep-Index. Já no experimento usando somente os elementos relevantes do Rep-Model, o resultado foi diferente, ou seja, a classificação do CNPq se manteve a mesma.

[35] Dados do experimento realizado na seção 5.3.2 (Figura 5.34).

5.2.4 Análise dos Resultados dos Experimentos do Rep-Model

Os resultados dos experimentos evidenciaram algumas questões interessantes quando se aborda o tema reputação de pesquisadores. Um dos pontos a serem analisados é em relação ao Rep-Index dos pesquisadores das três áreas envolvidas. Os testes para identificar a relevância de todos os elementos utilizados no Rep-Model para gerar o Rep-Index dos pesquisadores apresentou resultados diferentes em cada uma das três áreas de pesquisa. Na sequência, são apresentados os resultados.

Após os testes, a área de Ciência da Computação foi a que apresentou melhor resultado para a relevância dos elementos do Rep-Model, atingindo um percentual de 69.3069% de grau de exatidão. Dessa forma, apresentou relevância para 13 dos 18 elementos que compõem o Rep-Model. Dos 13 elementos, os que mais apresentaram relevância, e que foram os definidos pelo algoritmo para iniciar a construção da árvore, foram o AP (Artigo em Periódico), seguido pelo MCEP (Membro de Corpo Editorial de Periódico). Os resultados da área de Ciência da Computação também apontaram que o nível do CNPq que mais se aproximou do resultado do Rep-Index foi o nível 2, alcançando 80,07% de compatibilidade. Já o nível que apresentou menor compatibilidade foi o 1C, alcançando apenas 33,33%.

Os resultados da área de Economia foram semelhantes aos da área de Ciência da Computação. Após os testes, a área de Economia apresentou um grau de exatidão de 67,1429 acerca da relevância entre a classificação do CNPq e o resultado do Rep-Index dos 210 pesquisadores. Dos 18 elementos do Rep-Model, 13 foram considerados relevantes. Dos 13 elementos que mais apresentaram relevância, o algoritmo iniciou a criação da árvore pelo HI (H-Index), seguido pelo RC (Rede de Coautoria) e RP (Revisor de Periódico). Os resultados da área de Economia indicaram que o nível em que foi encontrado maior precisão foi o nível 2, com 79,2% de compatibilidade. Já o nível que apresentou menor precisão foi o 1B, alcançando 16,7%.

Em relação a área de Odontologia, foi a que menos apresentou relevância entre a classificação do CNPq e o resultado do Rep-Index dos pesquisadores vinculados à área, obtendo um percentual de 61.5741% de grau de exatidão. Isso representa que dos 18 elementos do Rep-Model utilizados nos testes, apenas 6 apresentaram relevância. Destes, o que mais apresentou relevância foi o elemento HI (H-Index), pois foi o escolhido pelo algoritmo para iniciar a construção da árvore. Na sequência, o segundo mais relevante foi o elemento CCC (Coordenação de Comitê de Conferência). Os resultados da área de Odontologia também apontaram que o nível do CNPq que mais se aproximou do resultado do Rep-Index foi o nível 2, atingindo 75% de compatibilidade. Já os níveis que apresentaram menor precisão foram os de níveis 1B e 1C, alcançando 20%.

Também foram analisadas as relações entre os resultados dos elementos mais relevantes e dos menos relevantes entre as três áreas. A Figura 5.27 apresenta os resultados.

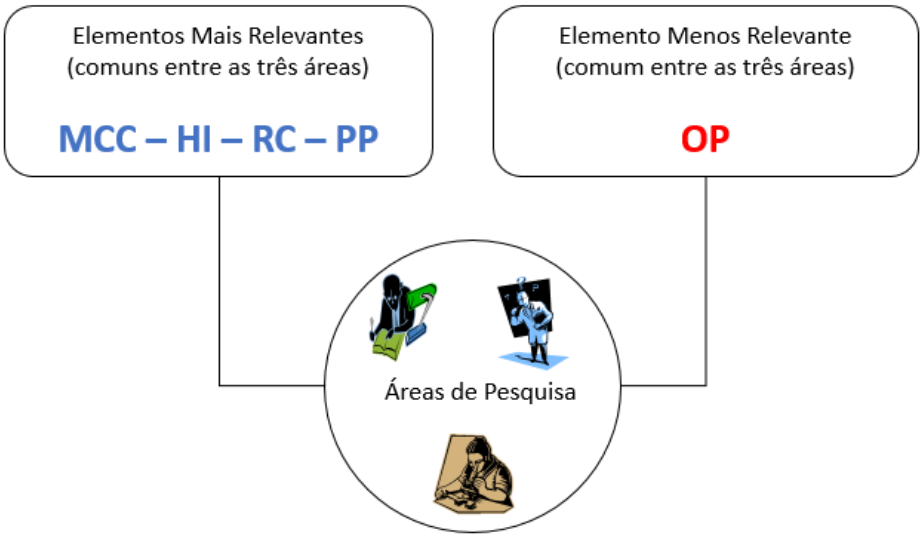


Figura 5.27: Resultado dos elementos que apresentaram mais relevância e menos relevância entre as três áreas envolvidas: Ciência da Computação, Economia e Odontologia.

Na Figura 5.27 estão representados os elementos do Rep-Model que apresentaram mais e menos relevância com a classificação do CNPq levando-se em consideração os 830 pesquisadores das áreas de Ciência da Computação, Odontologia e Economia. Os elementos relevantes que foram comuns entre as três áreas foram o HI (H-Index), MCC (Membro de Comitê de Conferência), RC (Rede de Coautoria) e PP (Projeto de Pesquisa). Já em relação aos elementos não relevantes, o único que foi comum entre as três áreas envolvidas foi o elemento OP (Orientação de Pós-doutorado).

Em relação aos resultados do experimento para a avaliação seletiva, ou seja, utilizando somente os elementos considerados relevantes de cada área de pesquisa, chegou-se aos seguintes resultados:

- Todos os pesquisadores da área de Ciência da Computação tiveram seu Rep-Index alterado do nível 2 para o nível 3, com exceção dos bolsistas do nível 2 do CNPq, que permaneceram no mesmo nível;
- Todos os pesquisadores da área de Economia tiveram seu Rep-Index alterado do nível 2 para o nível 3;
- Todos os pesquisadores da área de Odontologia tiveram seu Rep-Index alterado do nível 2 para o nível 3, com exceção dos bolsistas do nível 1A que passaram do nível 2 para o nível 4.

5.3 Experimentos do Rep-Index

Esta seção descreve o grupo de experimentos que conduziram a avaliação do Rep-Index para identificar a reputação de pesquisadores. Para analisar a existência ou não de correlação do Rep-Index com o h-index e com o g-index, foram processados os dados dos 830 pesquisadores do CNPq, usando todos os elementos do Rep-Model. Assim, foram identificados o Rep-Index de cada um deles. Após, os resultados foram comparados. Primeiro, o nível de correlação entre o Rep-Index e o h-index foi analisado. Depois, o Rep-Index foi comparado com o g-index. Conforme descrito

anteriormente, o coeficiente de correlação de Spearman foi utilizado como método estatístico.

Para os experimentos foram utilizadas todas as categorias e todos os elementos do Rep-Model. A calibração dos pesos foi definida de acordo com os critérios do CNPq, onde os requisitos para concessão de bolsas de produtividade foram transformamos em valores inteiros. Assim, os pesos dos elementos foram ajustados até que o valor do Rep-Index fosse compatível com a classificação dos níveis das bolsas. Enfatiza-se que a calibração dos pesos dos elementos do Rep-Model são adaptáveis e podem ser ajustados pelo usuário de acordo com o contexto e com critérios específicos de utilização.

Em relação aos valores dos elementos, estes foram definidos pela análise dos dados dos 830 pesquisadores. Para cada pesquisador foi analisada a totalidade do conjunto de elementos do modelo, identificando o valor de ocorrência de cada elemento, entre todos os pesquisadores. Tendo os valores de todos os elementos dos pesquisadores, foi identificado o valor mais elevado de cada elemento. Para identificar o valor mais elevado, foi calculada a média aritmética entre os elementos de cada área. O valor mais alto do elemento é necessário para calcular o Rep-Index, conforme detalhado na seção 4.3.

A Tabela 5.4 apresenta as categorias e seus elementos, seus respectivos pesos e o valor mais elevado de cada elemento.

Tabela 5.4: Categorias, elementos, pesos e maior valor do elemento.

Categoria	Elemento (Sigla)	Peso	Maior Valor do Elemento
Identificação (ID)	NM	-	-
	INST	-	-
	GI	15	15
Orientação (ORI)	OM	4	116
	OD	5	59
	OP	6	13
Banca (BAN)	PBM	4	159
	PBD	6	95
Comitê (COM)	CCC	1	23
	MCC	1	57
	MCEP	5	5
	RP	3	3
Publicação (PUB)	AP	15	237
	CLIV	5	84
	LIV	7	39
	TCC	8	299
	HI	8	31
	RC	3	262
	PP	1	130
	SOFT	1	19
Total	-	100	-

Fonte: adaptado de (CERVI; GALANTE; OLIVEIRA, 2013-a).

Dado o valor máximo para cada elemento identificado, pode-se definir o número de intervalos para a classificação dos pesquisadores nos níveis de reputação. Nesse grupo de experimentos foram definidos cinco níveis de reputação, para que ficassem alinhados com os níveis de classificação do CNPq.

A Figura 5.28 apresenta o grupo de experimentos que foram executados, conforme especificação a seguir:

- Experimento 4 – o Rep-Index é comparado com o ranking do CNPq usando a média dos valores das três áreas;
- Experimento 5 – o Rep-index é comparado com o ranking do CNPq usando os valores de cada área individualmente;
- Experimento 6 – calculamos a correlação do Rep-Index usando os critérios do CNPq e comparamos com o h-index e com o g-index dos 830 pesquisadores;
- Experimento 7 – comparamos a correlação entre os pesos dos elementos do Rep-Model com a classificação do CNPq.

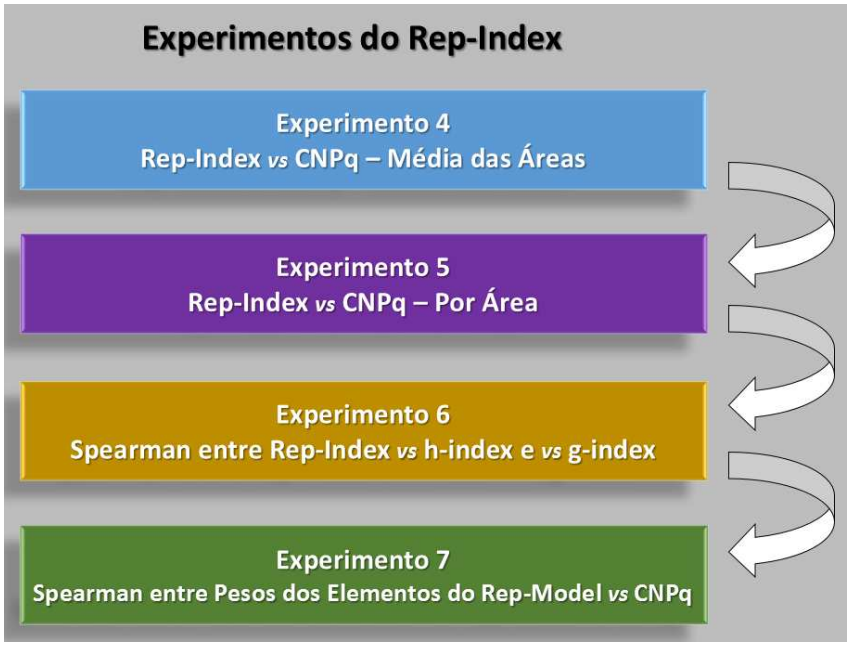


Figura 5.28: Resultado dos elementos que apresentaram mais relevância e menos relevância entre as três áreas envolvidas: Ciência da Computação, Economia e Odontologia.

5.3.1 Experimento 4 – Rep-Index Comparado ao Ranking do CNPq (média das áreas)

O objetivo deste experimento é verificar a equivalência entre o resultado do Rep-Index dos pesquisadores usando dados da média entre as três áreas (Ciência da Computação, Economia e Odontologia) com a classificação de cada área do CNPq. A hipótese é que a classificação de cada área do CNPq é equivalente ao resultado do Rep-Index usando a média dos valores dos elementos entre as três áreas. Para realizar o experimento foram comparados os pesquisadores de cada nível do CNPq de cada área com o resultado do Rep-Index dos pesquisadores avaliados.

Para uma melhor compreensão dos resultados, as Figuras 5.29, 5.30 e 5.31 apresentam, entre parênteses, o valor do Rep-Index em formato decimal. Destaca-se que o valor do Rep-Index é um valor inteiro e positivo compreendido do intervalo 1 a 5.

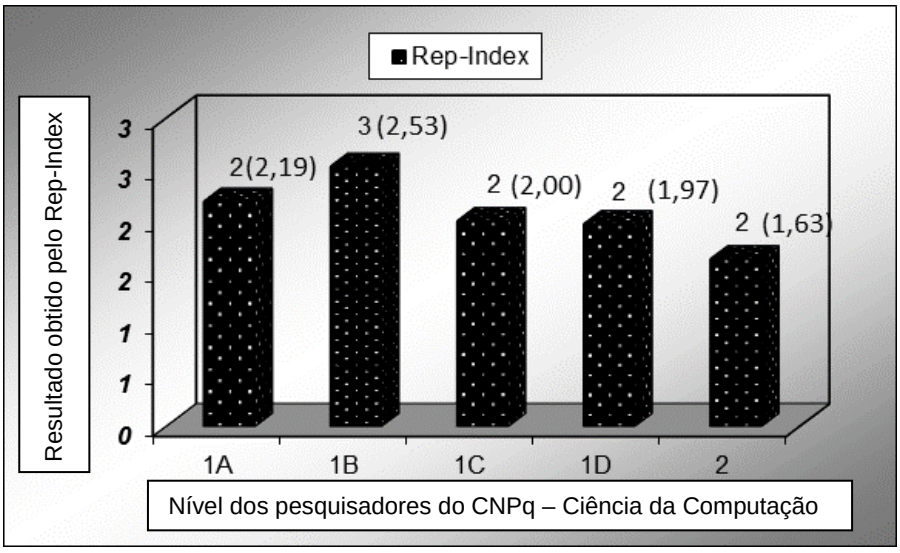


Figura 5.29: Rep-Index vs classificação do CNPq da área de Ciência da Computação comparado com a média das três áreas (adaptado de Cervi, Galante e Oliveira (2013-a)).

A Figura 5.29 mostra o resultado do Rep-Index dos pesquisadores da área de Ciência da Computação comparado com a classificação do CNPq. Pode-se observar que o Rep-Index dos pesquisadores sofreu alteração entre os níveis 1A e 1B. Enquanto os pesquisadores do nível 1A apresentaram Rep-Index igual a 2 (2,19), os pesquisadores do nível 1B atingiram Rep-Index 3 (2,53). Este resultado apresenta compatibilidade de 80% entre os pesquisadores da área de Ciência da Computação se comparados com o resultado da média entre as três áreas envolvidas no processo.

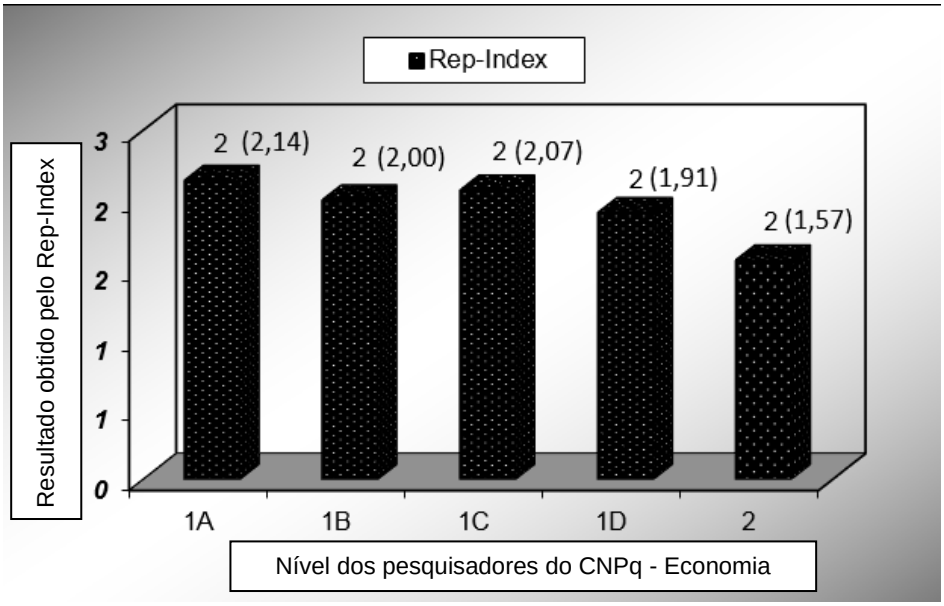


Figura 5.30: Rep-Index vs classificação do CNPq da área de Economia comparado com a média das três áreas (adaptado de (Cervi, Galante e Oliveira (2013-a)).

A Figura 5.30 mostra o resultado do Rep-Index dos pesquisadores da área de Economia comparado com a classificação do CNPq. A única exceção foi uma inversão

entre os grupos 1B e 1C, onde o grupo de pesquisadores 1C obteve um resultado pouco superior (2,07, com Rep-Index 2) ao grupo de pesquisadores 1B (2,00 com Rep-Index 2). Tal resultado mostra que houve uma compatibilidade de 80% entre os pesquisadores da área de Economia se comparados com o resultado da média entre as três áreas envolvidas no processo.

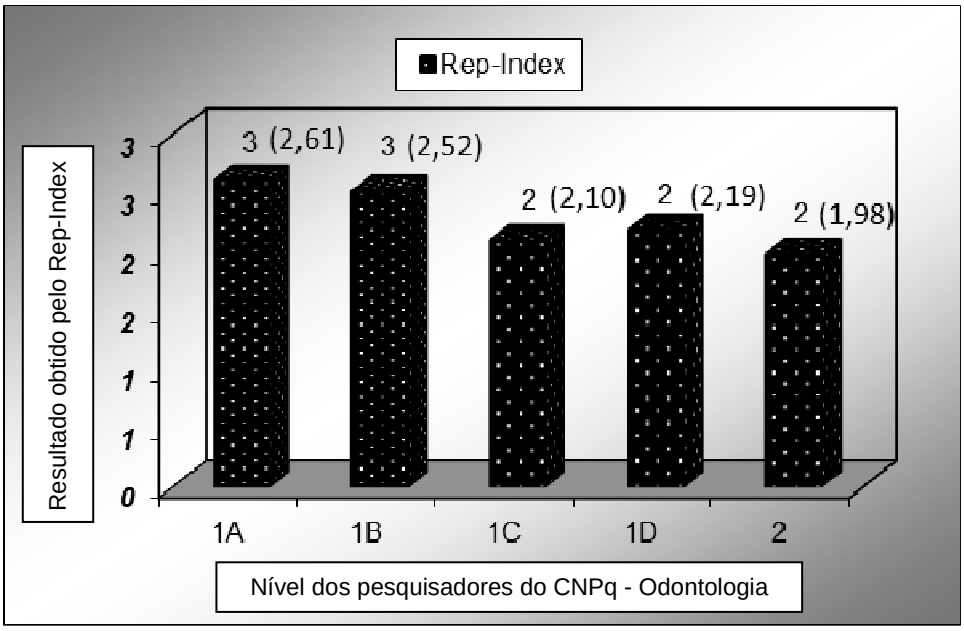


Figura 5.31: Rep-Index vs classificação do CNPq da área de Odontologia comparado com a média das três áreas (adaptado de (Cervi, Galante e Oliveira (2013-a)).

O resultado apresentado na Figura 5.31 mostra que o Rep-Index dos pesquisadores da área de Odontologia também sofreu uma inversão. Os bolsistas do nível 1D (Rep-Index 2 ou 2,19) apresentaram maior reputação que os bolsistas do nível 1C (Rep-Index 2 ou 2,10). Este resultado, mesmo que pouco significante, apresentou compatibilidade de 80% entre os pesquisadores da área de Odontologia se comparados com o resultado da média entre as três áreas envolvidas no processo.

5.3.2 Experimento 5 – Rep-Index Comparado ao Ranking do CNPq (individualizado por área)

O objetivo deste experimento é verificar a equivalência entre o resultado do Rep-Index e a classificação do CNPq nas áreas de Ciência da Computação, Economia e Odontologia, analisando o Rep-Index dos pesquisadores pela média de cada área individualmente. A hipótese é que o Rep-Index é equivalente a classificação do CNPq em cada área de pesquisa.

Para executar o experimento, foram coletados dados dos 830 bolsistas do CNPq usando todos os elementos do Rep-Model de cada área individualmente e foi gerado o Rep-Index dos pesquisadores. Os resultados do experimentos são apresentados nas Figuras 5.32, 5.33 e 5.34.

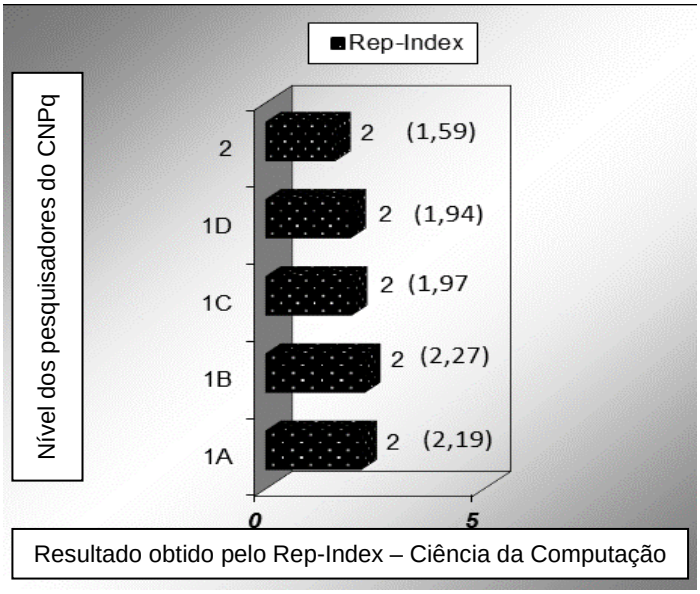


Figura 5.32: Rep-Index dos pesquisadores da área de Ciência da Computação vs classificação do CNPq da área de Ciência da Computação (adaptado de Cervi, Galante e Oliveira (2013-a)).

A Figura 5.32 apresenta os resultados da área de Ciência da Computação levando-se em consideração a média da própria área. Observa-se que o Rep-Index dos pesquisadores seguiu a classificação do CNPq, tendo apenas uma inversão entre os pesquisadores dos níveis 1A e 1B. Os pesquisadores do nível 1B apresentaram resultado um pouco superior aos pesquisadores do nível 1A. Rep-Index 2 (2,27) para os do nível 1B e Rep-Index 2 (2,19) para os do nível 1A. Isso representa uma compatibilidade de 80% entre os resultados do Rep-Index e a classificação do CNPq para a área de Ciência da Computação. Tal percentual de compatibilidade foi o mesmo encontrado no experimento onde os pesquisadores da área de Ciência da Computação foram avaliados utilizando-se a média entre as três áreas (Figura 5.29).

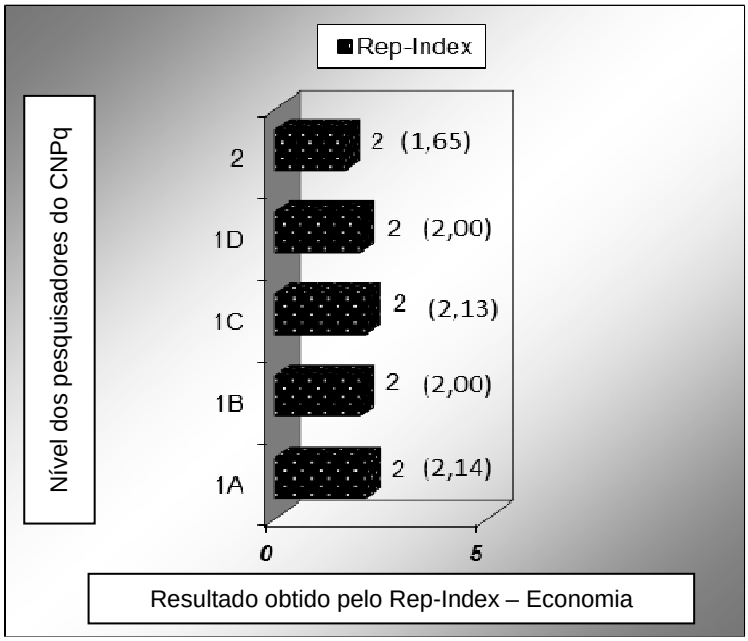


Figura 5.33: Rep-Index dos pesquisadores da área de Economia vs classificação do CNPq da área de Economia (adaptado de Cervi, Galante e Oliveira (2013-a)).

A Figura 5.33 mostra o resultado do Rep-Index dos pesquisadores da área de Economia segundo a classificação do CNPq, considerada a média da própria área de Economia. Nesse experimento, o resultado do Rep-Index foi compatível em 80%. A única alteração foi encontrada entre os pesquisadores dos níveis 1B e 1C, onde os bolsistas do nível 1C obtiveram um valor mais elevado de Rep-Index, 2 (2,13) para os do nível 1C e 2 (2,00) para os do nível 1B. Esse resultado também representa uma compatibilidade de 80% entre o Rep-Index e a classificação do CNPq dos pesquisadores da área de Economia. Tal percentual de compatibilidade foi o mesmo encontrado no experimento onde os pesquisadores da área de Economia foram avaliados utilizando-se a média entre as três áreas (Figura 5.30).

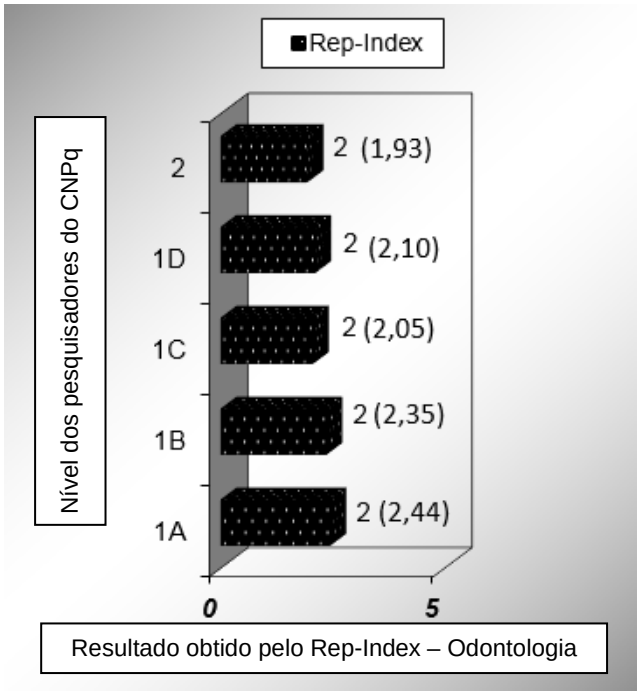


Figura 5.34: Rep-Index dos pesquisadores da área de Odontologia vs classificação do CNPq da área de Odontologia (adaptado de Cervi, Galante e Oliveira (2013-a)).

Nos resultados apresentados na Figura 5.34 o Rep-Index dos pesquisadores da área de Odontologia, usando a média da própria área de Odontologia, seguiu a classificação do CNPq, com exceção dos pesquisadores dos níveis 1C e 1D. Os bolsistas do nível 1D apresentaram um valor de Rep-Index um pouco superior aos bolsistas do nível 1C. Rep-Index 2 (2,10) para os do nível 1D e Rep-Index 2 (2,05) para os do nível 1C. Esse resultado também apresenta uma compatibilidade de 80% entre o valor do Rep-Index dos pesquisadores e a classificação do CNPq. Esse percentual de compatibilidade foi o mesmo encontrado no experimento onde os pesquisadores da área de Odontologia foram avaliados utilizando-se a média entre as três áreas (Figura 5.31).

5.3.3 Experimento 6 – Correlação de Spearman entre Rep-Index, h-index e g-index

O objetivo desse experimento é calcular a correlação do Rep-Index com o h-index e com o g-index dos 830 pesquisadores das áreas de Ciência da Computação, Economia e Odontologia. A hipótese esperada é que os três índices sejam compatíveis.

Para executar os experimentos foi usado como método de comparação o coeficiente de correlação de Spearman, conforme justificado anteriormente. Os pesquisadores de

cada área foram analisados e comparamos com a classificação do CNPq. Para a definição do Rep-Index dos pesquisadores foram usadas todas as categorias e todos os elementos do Rep-Model para cada área. Também foi identificado o h-index e o g-index de todos os pesquisadores, sendo finalizado com as comparações. Após, foi gerado o coeficiente de correlação de Spearman entre os três índices para todos os 830 pesquisadores. Os resultados são apresentados nas Figuras 5.35, 5.36 e 5.37.

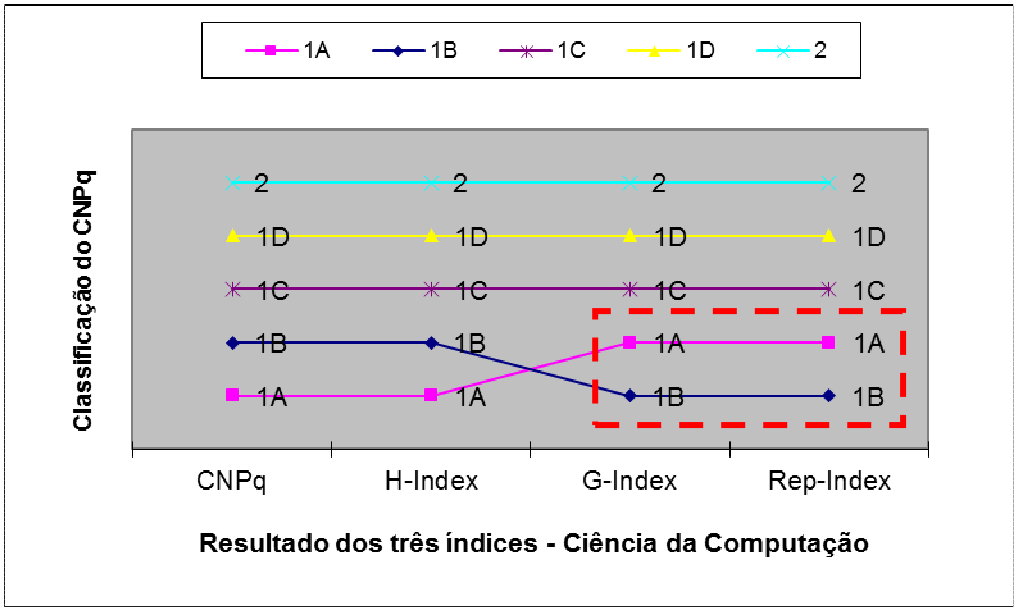


Figura 5.35: Resultado usando correlação de Spearman dos bolsistas do CNPq da área de Ciência da Computação com os índices Rep-Index, g-index e h-index (adaptado de Cervi, Galante e Oliveira (2013-a)).

A Figura 5.35 mostra os resultados dos índices da área de Ciência da Computação. Em relação ao h-index não houve alteração em relação a classificação do CNPq. Isto mostra que o h-index possui 100% de compatibilidade com os critérios do CNPq para a área de Ciência da Computação, pois o coeficiente de correção de Spearman encontrado foi 1.0. Já em relação ao g-index, o resultado obtido foi diferente. Mesmo apresentando um coeficiente de correlação de Spearman de 0.9, o que indica uma forte correlação, houve inversão de posição entre os pesquisadores do nível 1A com os pesquisadores do nível 1B.

Os resultados demonstram que o g-index pode ser mais adequado que o h-index quando utilizado para os pesquisadores da área de Ciência da Computação. Isto fica evidenciado pelo desempenho dos bolsistas do CNPq do nível 1B, pois sua produção foi mais elevada que os pesquisadores do nível 1A, quando analisados pelo g-index. Nos demais níveis não houve alteração na classificação. Tal resultado também mostra que o g-index é compatível em 100% com o resultado do Rep-Index para o mesmo grupo de bolsistas, conforme destacado na Figura 5.36.

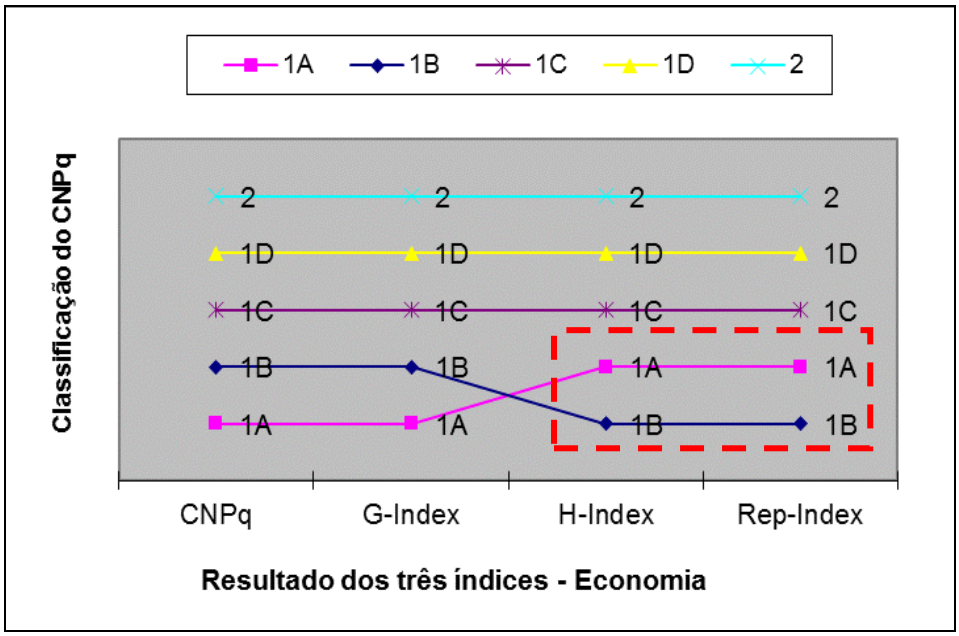


Figura 5.36: Resultado usando correlação de Spearman dos bolsistas do CNPq da área de Economia com os índices Rep-Index, g-index e h-index (adaptado de Cervi, Galante e Oliveira (2013-a)).

A Figura 5.36 mostra os resultados dos índices da área de Economia. Em relação ao g-index não houve alteração em relação a classificação do CNPq. Isto mostra que o g-index possui 100% de compatibilidade com os critérios do CNPq para a área de Economia, pois o coeficiente de correção de Spearman encontrado foi 1.0. Já em relação ao h-index, o resultado obtido foi diferente. Mesmo apresentando um coeficiente de correlação de Spearman de 0.9, o que indica uma forte correlação, houve inversão de posição entre os pesquisadores do nível 1A com os pesquisadores do nível 1B.

Os resultados demonstram que o h-index pode ser mais adequado que o g-index quando utilizado para os pesquisadores da área de Economia. Isto fica evidenciado pelo desempenho dos bolsistas do CNPq do nível 1B, pois sua produção foi mais elevada que os pesquisadores do nível 1A, quando analisados pelo h-index. Nos demais níveis não houve alteração na classificação. Tal resultado também mostra que o h-index é compatível em 100% com o resultado do Rep-Index para o mesmo grupo de bolsistas, conforme destacado na Figura 5.37.

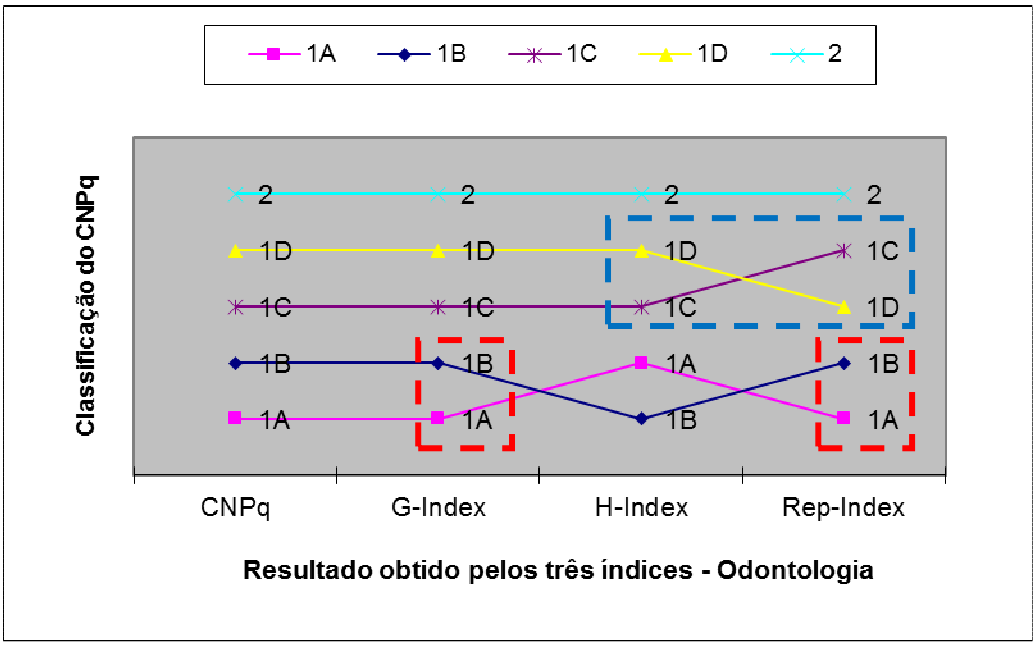


Figura 5.37: Resultado usando correlação de Spearman dos bolsistas do CNPq da área de Odontologia com os índices Rep-Index, g-index e h-index (adaptado de Cervi, Galante e Oliveira (2013-a)).

A Figura 5.37 mostra os resultados dos índices da área de Odontologia. Em relação ao g-index não houve alteração em relação a classificação do CNPq. Isto mostra que o g-index possui 100% de compatibilidade com os critérios do CNPq para a área de Odontologia, pois o coeficiente de correção de Spearman encontrado foi 1.0. Já em relação ao h-index, o resultado obtido foi diferente. Mesmo apresentando um coeficiente de correlação de Spearman de 0.9, o que indica uma forte correlação, houve inversão de posição entre os pesquisadores do nível 1A com os pesquisadores do nível 1B.

Os resultados demonstram que o h-index pode ser mais adequado que o g-index quando utilizado para os pesquisadores da área de Odontologia. Isto fica evidenciado pelo desempenho dos bolsistas do CNPq do nível 1B, pois sua produção foi mais elevada que os pesquisadores do nível 1A, quando analisados pelo h-Index. Nos demais níveis não houve alteração na classificação.

Em relação ao resultado do Rep-Index, este apresentou compatibilidade de 100% com o g-Index. Quando comparado com o h-index, o resultado do Rep-Index foi diferente da classificação do CNPq, pois houve inversão entre os pesquisadores dos níveis 1A e 1B, mesmo que o resultado de Spearman tenha indicado um índice de correlação igual a 0.9, o que representa uma forte correlação. Ainda com relação ao Rep-Index, também houve alteração na classificação do CNPq entre os níveis 1C e 1D, onde os bolsistas do nível 1D apresentaram produção superior aos bolsistas do nível 1C. Tal resultado não foi identificado para os demais índices (g-index e h-index), onde os níveis 1C, 1D e 2, não apresentaram alteração na classificação.

5.3.4 Experimento 7 – Correlação de Spearman entre os Elementos do Rep-Model

Este experimento tem como objetivo apresentar a correlação de cada elemento do Rep-Model na classificação do CNPq nas áreas de Ciência da Computação, Economia e

Odontologia. A hipótese é que os elementos do Rep-Model são fortemente correlacionados com a classificação do CNPq nas três áreas.

Para a realização desse experimento, foi calculado o coeficiente de correlação de Spearman dos 18 (dezoito) elementos do Rep-Model para os 830 pesquisadores do CNPq. Os resultados são apresentados nas Figuras 5.38, 5.39 e 5.40.

A Figura 5.38 mostra o resultado da correlação de Spearman para os bolsistas da área de Ciência da Computação.

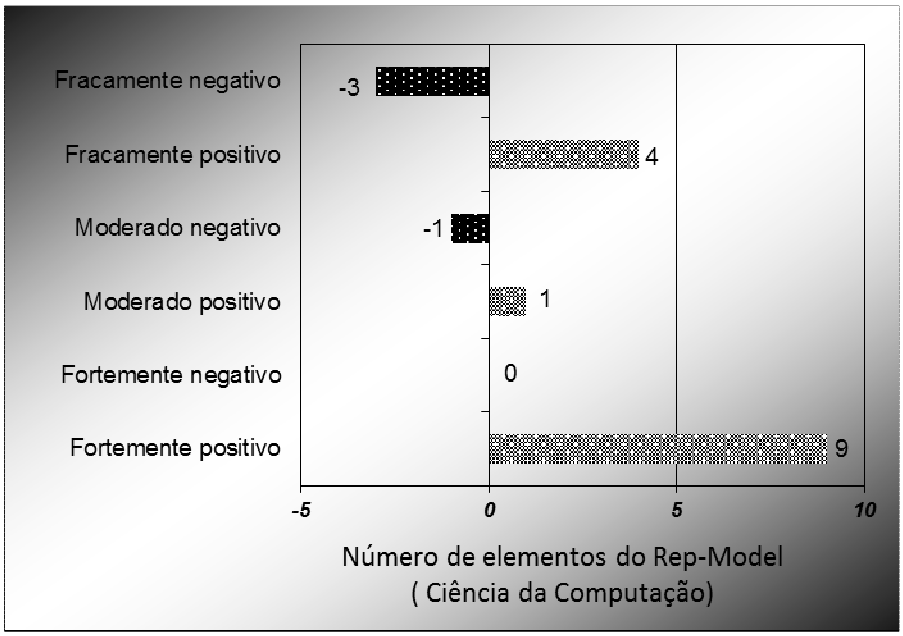


Figura 5.38: Resultado usando correlação de Spearman para todos os elementos do Rep-Model dos bolsistas do CNPq da área de Ciência da Computação (adaptado de Cervi, Galante e Oliveira (2013-a)).

Pode-se observar na Figura 5.38 que 9 (nove) elementos do Rep-Model apresentaram correlação forte, 4 (quatro) fraca positiva, 3 (três) fraca negativa, 1 (um) fortemente positiva, 1 (um) fortemente negativa, enquanto que nenhum elemento apresentou correlação fortemente negativa.

Os elementos e as suas correlações são apresentamos a seguir:

- Fortemente positiva: Orientação de Pós-doutorado, Orientação de Doutorado, Orientação de Mestrado, Membro de Corpo Editorial de Periódico, Artigo em Periódico, Capítulo de Livro, Livro, H-Index e Rede de Coautoria;
- Fortemente negativa: Nenhum elemento;
- Moderada positiva: Grau de Instrução;
- Moderada negativa: Projeto de Pesquisa;
- Fracamente positiva: Participação em Banca de Doutorado, Coordenação de Comitê de Conferência, Revisão de Periódico e Trabalho Completo em Conferência.
- Fracamente negativa: Participação em Banca de Mestrado, Membro de Comitê de Conferência e Software.

A Figura 5.39 mostra o resultado da correlação de Spearman para os bolsistas da área de Economia.

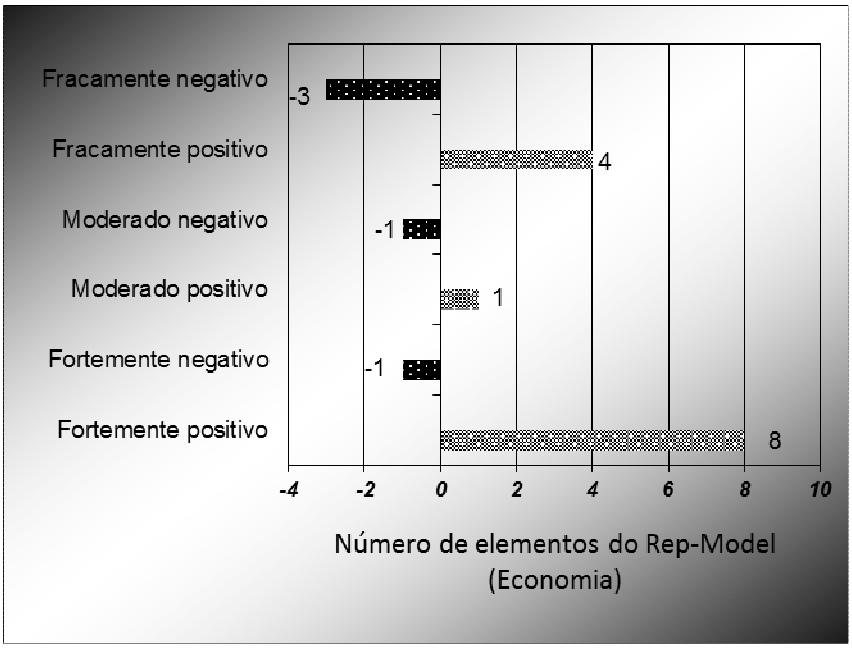


Figura 5.39: Resultado usando correlação de Spearman para todos os elementos do Rep-Model dos bolsistas do CNPq da área de Economia (adaptado de Cervi, Galante e Oliveira (2013-a)).

Pode observar na Figura 5.39 que 8 (oito) elementos do Rep-Model apresentaram correlação fortemente positiva, 4 (quatro) fracamente positiva, 3 (três) fracamente negativa, 1 (um) fortemente negativa, 1 (um) moderada positiva e 1 (um) moderada negativa.

Os elementos e as suas correlações são apresentados a seguir:

- Fortemente positiva: Orientação de Mestrado, Coordenação de Comitê de Conferência, Membro de Comitê de Conferência, Artigo em Periódico, Capítulo de Livro, Livro, H-Index, Rede de Coautoria;
- Fortemente negativa: Revisão de Periódico;
- Moderada positiva: Membro de Corpo Editorial de Periódico;
- Moderada negativa: Projeto de Pesquisa;
- Fracamente positiva: Grau de Instrução, Orientação de Pós-doutorado, Orientação de Doutorado e Software;
- Fracamente negativa: Participação em Banca de Mestrado, Participação em Banca de Doutorado e Trabalho Completo em Conferência.

A Figura 5.40 mostra o resultado da correlação de Spearman para os bolsistas da área de Odontologia.

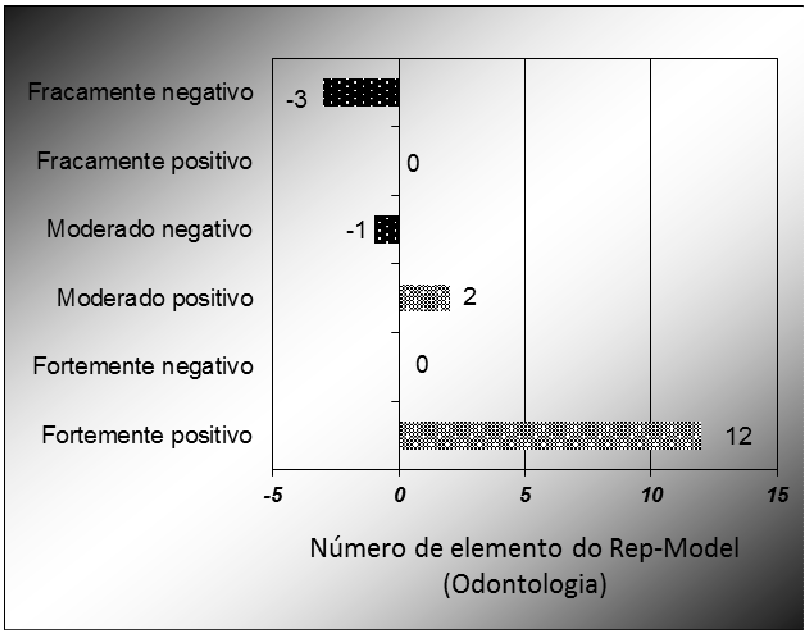


Figura 5.40: Resultado usando correlação de Spearman para todos os elementos do Rep-Model dos bolsistas do CNPq da área de Odontologia (adaptado de Cervi, Galante e Oliveira (2013-a)).

Pode-se observar na Figura 5.40 que 12 (doze) elementos do Rep-Model apresentaram correlação fortemente positiva, 3 (três) fracamente negativa, 2 (dois) moderada positiva, 1 (um) moderada negativa e nenhum apresentou correlação fortemente negativa e fracamente positiva.

Os elementos e as suas correlações são apresentados a seguir:

- Fortemente positiva: Orientação de Pós-doutorado, Orientação de Doutorado, Orientação de Mestrado, Participação em Banca de Mestrado, Participação em Banca de Doutorado, Coordenação de Comitê de Conferência, Membro de Comitê de Conferência, Artigo em Periódico, Capítulo de Livro, Livro, H-Index e Rede de Coautoria;
- Fortemente negativa: Nenhum elemento;
- Moderada positiva: Membro de Corpo Editorial de Periódico e Trabalho Completo em Conferência;
- Moderada negativa: Software;
- Fracamente positiva: Nenhum elemento;
- Fracamente negativa: Grau de Instrução, Revisão de Periódico e Projeto de Pesquisa.

5.3.5 Análise dos Resultados dos Experimentos do Rep-Index

Após realizar os experimentos, analisamos os resultados a fim de identificar padrões, tendências e comportamento nos dados analisados. Quanto ao comportamento dos pesquisadores nas áreas de Ciência da Computação, Economia e Odontologia, tendo em vista a classificação do CNPq, os resultados demonstram que os critérios utilizados pelo CNPq para a concessão de bolsas de produtividade são relevantes e consistentes. Mesmo existindo algumas diferenças entre as áreas, o que é normal dentro de qualquer

população, os resultados apresentaram forte correlação, chegando a um percentual de mais de 80%. Essa correlação forte foi identificada quando os pesquisadores foram avaliados individualmente em cada área de pesquisa, bem como pela média entre as três áreas.

Em relação a comparação do Rep-Index com o h-index e o g-index, identificou-se que o Rep-Index dos 830 pesquisadores das áreas de Ciência da Computação, Economia e Odontologia apresentaram forte correlação. Os resultados demonstram que o g-index é mais adequado que o h-index quando utilizado para os pesquisadores da área de Ciência da Computação. Esse resultado aponta que o Rep-Index é compatível em 100% com o resultado do g-index para o mesmo grupo de bolsistas. Com relação à área de Economia, os resultados demonstram que o h-index é mais adequado que o g-index. Esse resultado também mostra que o Rep-Index é compatível em 100% com o resultado do h-index para o mesmo grupo de bolsistas.

Para a área de Odontologia, levando-se em comparação o h-index e o g-index, os resultados demonstram que o h-index é mais adequado que o g-index. Em relação ao Rep-Index, este apresentou compatibilidade de 100% com o g-Index para os bolsistas da área de Odontologia. No caso da comparação com o h-index, o resultado do Rep-Index foi diferente da classificação do CNPq, pois houve inversão entre os pesquisadores dos níveis 1A e 1B, bem como dos níveis 1C e 1D, mesmo que o resultado da correlação de Spearman tenha indicado um índice de correlação igual a 0.9, o que representa uma forte correlação.

Foi analisada também a correlação entre os elementos do Rep-Model nas áreas de Ciência da Computação, Economia e Odontologia. Dos 18 (dezoito) elementos do modelo, foi identificada uma forte correlação em 6 (seis) destes elemento nas três áreas de pesquisa: OM (Orientação de Mestrado), AP (Artigo em Periódico), CLIV (Capítulo de Livro), LIV (Livro), HI (H-Index) e RC (Rede de Coautoria). Outros elementos do Rep-Model também se destacaram. Quatro (4) deles apresentaram forte correlação entre duas das três áreas utilizadas: OP (Orientação de Pós-doutorado), OD (Orientação de Doutorado), CCC (Coordenação de Comitê de Conferência) e MCC (Membro de Comitê de Conferência).

Alguns elementos do Rep-Model demonstraram fraca correlação ou correlação negativa. Este é o caso dos elementos GI (Grau de Instrução), PBM (Participação em Banca de Mestrado), PBD (Participação em Banca de Doutorado), RP (Revisão de Periódico), TCC (Trabalho Completo em Conferência) e SOFT (Software). Já os elementos PP (Projeto de Pesquisa) e MCEP (Membro de Corpo Editorial de Periódico) apresentaram correlação moderada entre as três áreas.

A Figura 5.41 apresenta a correlação entre os percentuais dos elementos do Rep-Model entre as três áreas de pesquisa utilizadas nos experimentos: Ciência da Computação, Economia e Odontologia.

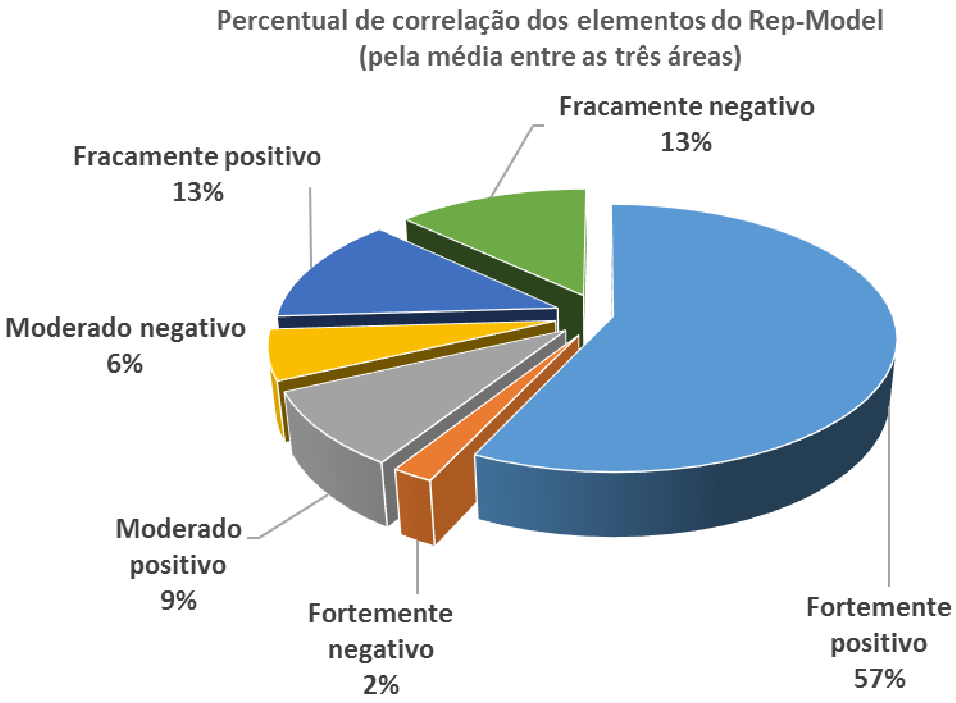


Figura 5.41: Percentual de correlação dos elementos do Rep-Model entre as áreas de Ciência da Computação, Economia e Odontologia.

Pode-se observar na Figura 5.41 que a correlação fortemente positiva entre os elementos do Rep-Model das três áreas foi de 57%. Além disso, a correlação positiva moderada foi de 9%. Somando o percentual das duas correlações, chega-se a um valor de 66%. Este resultado mostra que os elementos do Rep-Model podem ser usados em diferentes áreas de pesquisa e em diferentes contextos. Destaca-se que, pela quantidade de dados envolvidos nos experimentos, pela diversidade das áreas escolhidas, pelas especificidades de cada área, bem como pela heterogeneidade dos dados, o resultado de 66% de correlação fortemente positiva e positiva moderada pode ser considerado satisfatório.

6 CONCLUSÕES

Neste capítulo, são apresentadas as considerações finais, destacando-se as contribuições e resultados alcançados pela tese. Dentre as contribuições e resultados, são apresentados os principais trabalhos que foram desenvolvidos durante o curso de doutorado, como publicações, orientações e produção técnica. Por fim, são discutidos alguns trabalhos futuros identificados ao longo do desenvolvimento da tese.

6.1 Contribuições

Nesta tese, foi apresentada uma abordagem para identificar reputação acadêmica de pesquisadores levando-se em consideração duas premissas: abrangência e adaptabilidade. A abordagem pressupõe um modelo de perfil, denominado Rep-Model e uma métrica, denominada Rep-Index.

Para a definição do Rep-Model foram especificados dezoito indicadores que fazem parte da carreira de um pesquisador. Esses indicadores são representados no Rep-Model como categorias e elementos e são estruturados para suportarem ponderação. A ponderação se dá por meio de pesos associados aos elementos e pode ser definida de acordo com os critérios determinados pelo utilizador.

O Rep-Index é a métrica que identifica a reputação de pesquisadores usando os indicadores constantes no Rep-Model. A métrica classifica pesquisadores em níveis de reputação e pode ser utilizada em diferentes áreas e ser adaptada para suportar diferentes contextos.

A proposta tem como premissa ser abrangente e adaptável, pois engloba a vida científica do pesquisador construída ao longo de sua carreira científica. Tais premissas permitem a utilização da abordagem em diferentes áreas e em diferentes contextos, uma vez que áreas diferentes usam critérios diferentes. Dessa forma, dependendo do propósito de utilização, ou seja, o que se quer avaliar, o que se quer medir e para que se quer medir, os critérios podem ser adaptados para que contemplem os requisitos do usuário e suas ponderações.

Para avaliar o modelo de perfil de pesquisadores e a métrica para identificar reputação acadêmica, experimentos exaustivos foram realizados envolvendo análise de trajetória científica. Nos experimentos, bolsistas de produtividade em pesquisa e tecnologia do CNPq das áreas de Ciência da Computação, Economia e Odontologia foram utilizados. Para esses pesquisadores, foram realizados experimentos para comparar o resultado do Rep-Index de cada um deles com duas métricas conhecidas e amplamente utilizadas na comunidade científica de diversas áreas, o g-index e o h-index. O método estatístico escolhido para a validação foi o Coeficiente de Correlação de Postos de Spearman (ρ).

Dentre os experimentos realizados, usando como baseline os bolsistas de produtividade do CNPq das áreas de Ciência da Computação, Odontologia e Economia, totalizando 830 pesquisadores, algumas considerações interessantes foram observadas. Em relação a abrangência, ou seja, da quantidade de elementos que fazem parte da trajetória científica de um pesquisador, os resultados indicaram que existe relevância entre a classificação do CNPq e o resultado do Rep-Index dos 830 pesquisadores. A área que mais apresentou resultados satisfatórios foi a área de Ciência da Computação, seguida pela área de Economia. Em relação a área de Odontologia, os resultados dos experimentos também foram satisfatórios, no entanto, a relevância do número total de elementos do Rep-Model foi muito menor que nas outras áreas. Isto indica que para avaliar a reputação de pesquisadores da área de Odontologia não são necessários todos os elementos do Rep-Model.

No aspecto relacionado à adaptabilidade, os resultados mostraram que a utilização do Rep-Index em diferentes áreas de pesquisa é viável e interessante, principalmente porque o Rep-Index pode proporcionar uma avaliação seletiva dos elementos, bem como incorpora ponderação. Esses fatores proporcionam ao usuário a possibilidade de adaptar o modelo de acordo com seus critérios de utilização, bem como possa representar as especificidades de sua área de pesquisa.

Outra análise observada é que o Rep-Index dos pesquisadores do CNPq, comparados a seu h-index e a seu g-index, apresentou os seguintes resultados:

- na área de Ciência da Computação, o g-index apresentou melhor resultado;
- na área de Economia, o h-index apresentou melhor resultado;
- na área de Odontologia, o g-index apresentou melhor resultado.

Em relação aos resultados do Rep-Index, do h-index e do g-index com os pesquisadores do CNPq, o h-index e o g-index favorecem pesquisadores com mais produção, pois consideram no cálculo da reputação, exclusivamente, os trabalhos publicados e a quantidade de citações a esses trabalhos. Nessa comparação direta, os pesquisadores em início de carreira e não bolsistas de produtividade tendem a obter um resultado menos satisfatório, uma vez que não possuem alta produção. Já o Rep-Index, por envolver diversos elementos da carreira do pesquisador no cálculo da reputação, pode proporcionar uma avaliação mais justa.

Outra análise que pode ser feita sobre a abordagem proposta na tese é que a utilização do Rep-Model de forma integral (com todos os elementos) proporciona uma avaliação mais equilibrada e abrangente. Já a utilização de forma seletiva (com menos elementos) proporciona uma avaliação mais direcionada para aspectos específicos de acordo com os critérios do utilizador.

Os experimentos mostram que avaliar pesquisadores é um tema que merece reflexão e critérios bem definidos. Usar apenas uma métrica, como o h-index ou o g-index, pode ser arriscado para o processo de tomada de decisão. A utilização do Rep-Index pode proporcionar uma avaliação mais justa, pois outros elementos são considerados no cálculo da reputação.

6.2 Trabalhos Futuros

Vislumbra-se como trabalhos futuros, a realização de experimentos para avaliar reputação de grupos de pesquisadores, como por exemplo, pesquisadores vinculados a

programas de pós-graduação vinculados à Capes. Outro trabalho que pode ser desenvolvido é avaliar a abrangência e a adaptabilidade do Rep-Index utilizando pesquisadores iniciantes, que ainda estão em processo de construção de sua trajetória científica.

Em relação a melhorias na abordagem proposta na tese, a incorporação de aspectos temporais no Rep-Model pode ser interessante, pois o tempo poderia ser um elemento importante para representar a trajetória científica de um pesquisador. Alguns elementos que podem ser incorporados ao Rep-Model é o tempo de doutorado, o tempo das orientações, o tempo em que foi/é revisor/editor de periódico, o tempo de participação em bancas e o tempo das publicações e citações.

6.3 Publicações e Orientações Relacionadas

Esta seção apresenta as publicações e orientações desenvolvidas durante o curso de doutorado e quem possuem relação com os temas estudados ao longo desse período.

6.3.1 Publicações

1. CERVI, C. R.; GALANTE, R.; OLIVEIRA, J. P. M. Application of Scientific Metrics to Evaluate Academic Reputation in Different Research Areas. Proceedings of XXXIV International Conference on Computational Science (ICCS 2013), 2013, Bali, Indonesia. Proceedings of XXXIV International Conference on Computational Science (ICCS 2013), 2013. (*Qualis A2 – Ciência da Computação*)
2. CERVI, C. R.; GALANTE, R.; OLIVEIRA, J. P. M. Comparing the Reputation of Researchers Using a Profile Model and Scientific Metrics. Proceedings of XIII IEEE International Conference on Computer and Information Technology (CIT 2013), 2013, Sydney, Australia. IEEE Conference Proceedings, 2013. (*Qualis B1 – Ciência da Computação*)
3. CERVI, C. R.; GALANTE, R.; OLIVEIRA, J. P. M. An Adaptive Approach for Identifying Reputation of Researchers. Proceedings of International Conference WWW/Internet, 2012, Madri. Proceedings of the Iadis International Conference WWW/Internet, 2012. (*Qualis B2 – Ciência da Computação*)
4. CERVI, C. R.; GALANTE, R.; OLIVEIRA, J. P. M. Identificando a Reputação de Pesquisadores Usando um Modelo de Perfil Adaptativo. Anais do XXXVIII Seminário Integrado de Software e Hardware, 2011, Natal, RN. Anais do XXXVIII Seminário Integrado de Software e Hardware, 2011. (*Qualis B4 – Ciência da Computação*)
5. GUGEL, J.; CERVI, C. R.; GALANTE, R.; OLIVEIRA, J. P. M. Uma Ferramenta Para Análise Quantitativa da Produção Científica de Pesquisadores. Anais da VII Escola Regional de Banco de Dados, 2011, Novo Hamburgo, RS. Anais da VII Escola Regional de Banco de Dados, 2011.
6. MANICA, Edimar; CERVI, C. R.; DORNELES, C. F.; GALANTE, R. Ferramenta para Suporte a Consultas Temporais em SGBDs Convencionais. Anais da V Escola Regional de Banco de Dados, 2009, Ijuí, RS. Anais da V Escola Regional de Banco de Dados, 2009.

7. MANICA, Edimar; CERVI, C. R.; DORNELES, C. F.; GALANTE, R. EMap - Uma Interface de Consultas Temporais em SGBDs Relacionais. Anais da Sessão de Demos - XXIV Simpósio Brasileiro de Banco de Dados, 2009, Fortaleza, CE. Anais do XXIV Simpósio Brasileiro de Banco de Dados, 2009.
8. CERVI, C. R.; MANICA, Edimar; DORNELES, C. F.; GALANTE, R. BDTC - Uma Biblioteca Digital de Trabalhos Científicos com Serviços Integrados. Revista Brasileira de Computação Aplicada. [Passo Fundo], v.1, p. 65-76, 2009.
9. CERVI, C. R.; GALANTE, R.; OLIVEIRA, J. P. M. Mecanismo para Gestão do Perfil Evolutivo de Pesquisadores e Análise Preditiva Baseada em Comportamento Científico. Anais do SBBD – Simpósio Brasileiro de Banco de Dados (Resumo), 2008, Campinas. (*Qualis B3 Computação*)
10. OLIVEIRA, A. P.; CERVI, C. R. Implementação da BDMODONTO: Biblioteca Digital Multimídia Odontológica. Anais da IV Escola Regional de Banco de Dados, 2008, Florianópolis, SC. Anais da IV Escola Regional de Banco de Dados, 2008.
11. MANICA, Edimar; CERVI, C. R. Um Processo Automático para Extração de Metadados de Documentos PDF Usando um Template XML. Anais da IV Escola Regional de Banco de Dados, 2008, Florianópolis, SC. Anais da IV Escola Regional de Banco de Dados, 2008.
12. CERVI, C. R.; MANICA, Edimar; DORNELES, C. F.; PAVAN, W. Mapeamento de uma Linguagem de Consulta Temporal para o Banco de Dados PostgreSQL. Anais do VIII Simpósio de Informática do Planalto Médio, 2008, Passo Fundo, RS. Anais do VIII Simpósio de Informática do Planalto Médio, 2008.
13. CERVI, C. R.; SILVEIRA, L. S.; DORNELES, C. F.; PAVAN, W. Modelando o Perfil de Pesquisadores Através de Fontes de Dados Heterogêneas. Anais do VIII Simpósio de Informática do Planalto Médio, 2008, Passo Fundo, RS. Anais do VIII Simpósio de Informática do Planalto Médio, 2008.

6.3.2 Orientações

1. Laís Andressa Brock. Uma Ferramenta para Modelagem do Perfil de Alunos e Identificação da Evolução do Aprendizado. 2013. Trabalho de Conclusão de Curso. (Ciência da Computação) - Universidade de Passo Fundo.
2. Daniel Krauze Freire. SomeLikeYou - Uma Rede Social Baseada em Recomendações de Relacionamentos. 2011. Monografia. (Aperfeiçoamento/Especialização em Desenvolvimento de Software) - Universidade de Passo Fundo. Orientador: Cristiano Roberto Cervi.
3. Alisson Sebben da Cunha. IR - Uma Ferramenta Web para Identificar a Reputação de Pesquisadores. 2010. Trabalho de Conclusão de Curso. (Graduação em Ciência da Computação) - Universidade de Passo Fundo. Orientador: Cristiano Roberto Cervi.
4. Jardel Gugel. Uma Ferramenta Para Análise Quantitativa da Produção Científica de Pesquisadores. 2010. Trabalho de Conclusão de Curso. (Graduação em Ciência da Computação) - Universidade de Passo Fundo. Orientador: Cristiano Roberto Cervi.

5. Roger Pereira de Oliveira. Identificação de Comportamento Semelhante de Pesquisadores Através de Dados de Produção Científica. 2010. Trabalho de Conclusão de Curso. (Graduação em Ciência da Computação) - Universidade de Passo Fundo. Orientador: Cristiano Roberto Cervi.
6. Rodrigo Matt. Um Mecanismo Para Orientação de Carreira Científica Baseado em Perfil de Pesquisadores e Recomendação. 2010. Trabalho de Conclusão de Curso. (Graduação em Ciência da Computação) - Universidade de Passo Fundo. Orientador: Cristiano Roberto Cervi.
7. Paulo Roberto Silveira Paz Junior. Uma Ferramenta Web Para Extração de Redes Sociais de Pesquisadores. 2009. Trabalho de Conclusão de Curso. (Graduação em Ciência da Computação) - Universidade de Passo Fundo. Orientador: Cristiano Roberto Cervi.
8. Jean Carlo de Borba Espíndola (co-orientador). Um Estudo Analítico Sobre o Comportamento de Pesquisadores Baseado em Dados de Produção Científica. 2009. Trabalho de Conclusão de Curso. (Graduação em Ciência da Computação) - Universidade Federal do Rio Grande do Sul. Orientador: Cristiano Roberto Cervi.
9. Rodrigo Robusto Piovesan. Uma Aplicação para Dispositivos Móveis de Consulta à BDTC. 2008. Monografia. (Aperfeiçoamento/Especialização em Desenvolvimento de Software) - Universidade de Passo Fundo. Orientador: Cristiano Roberto Cervi.
10. Edimar Manica. Suporte a Consultas Temporais Através de um Mapeamento da Linguagem TSQL2 Para o PostgreSQL. 2008. Trabalho de Conclusão de Curso. (Graduação em Ciência da Computação) - Universidade de Passo Fundo. Orientador: Cristiano Roberto Cervi.
11. Fahad Kalil. Uma Ferramenta de Suporte à Análise do Comportamento Científico de Pesquisadores. 2008. Trabalho de Conclusão de Curso. (Graduação em Ciência da Computação) - Universidade de Passo Fundo. Orientador: Cristiano Roberto Cervi.
12. Leandro Serena da Silveira. Modelagem do Perfil de Pesquisadores Baseada em Dados de Produção Científica. 2008. Trabalho de Conclusão de Curso. (Graduação em Ciência da Computação) - Universidade de Passo Fundo. Orientador: Cristiano Roberto Cervi.

6.3.3 Produção Técnica

1. CUNHA, A. S.; CERVI, C. R.; GALANTE, R.; OLIVEIRA, J. P. M. Uma Ferramenta Web para Identificar a Reputação de Pesquisadores. (Software sem registro). 2010.
2. GUGEL, J.; CERVI, C. R.; GALANTE, R.; OLIVEIRA, J. P. M. Uma Ferramenta Para Análise Quantitativa da Produção Científica de Pesquisadores. (Software sem registro). 2010.
3. MATT, R.; CERVI, C. R.; GALANTE, R.; OLIVEIRA, J. P. M. Um Mecanismo Para Orientação de Carreira Científica Baseado em Perfil de Pesquisadores e Recomendação. (Software sem registro). 2010.

4. OLIVEIRA, R. P.; CERVI, C. R.; GALANTE, R.; OLIVEIRA, J. P. M. Identificação de Comportamento Semelhante de Pesquisadores Através de Dados de Produção Científica. (Software sem registro). 2010.
5. MANICA, Edinei; MANICA, Edimar; CERVI, C. R.; DORNELES, C. F.; GALANTE, R. BDTC - Uma Biblioteca Digital de Trabalhos Científicos com Serviços Integrados. (Software sem registro). 2009.
6. PAZ JUNIOR, P. R. S.; CERVI, C. R. Uma Ferramenta Web Para Extração de Redes Sociais de Pesquisadores. (Software sem registro). 2009.
7. MANICA, Edimar; CERVI, C. R. E-Map - Uma Ferramenta de Consultas Temporais Através de Um Mapeamento da Linguagem TSQL2 Para o PostgreSQL. (Software sem registro). 2008.
8. KALIL, F.; CERVI, C. R. Uma Ferramenta de Suporte a Análise do Comportamento Científico de Pesquisadores. (Software sem registro). 2008.
9. SILVEIRA, L. S.; CERVI, C. R. Uma Ferramenta Para Extração de Dados de Pesquisadores. (Software sem registro). 2008.
10. MANICA, Edimar; CERVI, C. R.; GALANTE, R. Um Processo Automático para Extração de Metadados de Documentos PDF Usando um Template XML. (Software sem registro). 2008.
11. PIOVESAN, R. R.; CERVI, C. R.; DORNELES, C. F.; PAVAN, Willingthon; GALANTE, R. BibMobile Uma Aplicação para Dispositivos Móveis de Consulta a uma Biblioteca Digital. (Software sem registro). 2008.

REFERÊNCIAS

ALONSO, S.; CABRERIZO, F. J.; HERRERA-VIEDMA, E.; HERRERA, F. Hg-index: A new index to characterize the scientific output of researchers based on the h- and g-indices. **Scientometrics**. [S.l.], v.82, n.2, p. 391-400, 2010.

ALONSO, S.; CABRERIZO, F.; HERRERA-VIEDMA, E.; HERRERA, F. H-index: A review focused in its variants, computation and standardization for different scientific fields. **Journal of Informetrics**. [S.l.], v.3, n.4, p. 273-289, 2009.

BLUM, A. L.; LANGLEY, P. Selection of relevant features and examples in machine learning. **Journal of Artificial Intelligence**. [S.l.], v.97, n.1-2, p. 245-271, 1997.

BAEZA-YATES, R.; RIBEIRO-NETO, B. Modern Information Retrieval. Prentice Hall, 2011.

BATISTA, P. D.; CAMPITELI, M. G.; KINOUCHI, O. Is it possible to compare researchers with different scientific interests? **Scientometrics**. [S.l.], v.68, n.1, p. 179-189, Jul. 2006.

BOLLEN J.; VAN DE SOMPEL, H.; HAGBERG, A.; CHUTE, R. A Principal Component Analysis of 39 Scientific Impact Measures. **PLoS ONE**. [S.l.], v.4, n.6, e6022, 2009.

BORNMANN L.; MUTZ, R.; DANIEL, H. D. The h-index research output measurement: Two approaches to enhance its accuracy. **Journal of Informetrics**. [S.l.], v.4, p. 407-414, 2010.

BORNMANN, L.; MUTZ, R.; DANIEL, H.-D. Are there better indices for evaluation purposes than the h-index? A comparison of nine different variants of the h index using data from biomedicine. **Journal of the American Society for Information Science & Technology.** [S.l.],, v.59, n.5, p. 830-837, 2008.

CARMEL, D., JOSIFOVSKI, V.; MAAREK, Y. User Modeling For Web Applications. **Proceedings of the fourth ACM International Conference on Web Search and Data Mining** (WSDM 2011). Hong Kong, China, 2011.

CERVI, C. R.; GALANTE, R.; OLIVEIRA, J. P. M. An Adaptive Approach for Identifying Reputation of Researchers. Proceedings of International Conference on WWW/Internet, Madrid, Spain, 2012.

CERVI, C. R.; GALANTE, R.; OLIVEIRA, J. P. M. Application of Scientific Metrics to Evaluate Academic Reputation in Different Research Areas. Proceedings of XXXIV

International Conference on Computational Science (ICCS 2013), Bali, Indonesia, 2013-a.

CERVI, C. R.; GALANTE, R.; OLIVEIRA, J. P. M. Comparing the Reputation of Researchers Using a Profile Model and Scientific Metrics. Proceedings of XIII IEEE International Conference on Computer and Information Technology (CIT 2013), Sydney, Australia, 2013-b.

CHEN, L.; SYCARA, K. WebMate: A personal agent for browsing and searching. In K. P. Sycara and M. Wooldridge, editors, Proceedings of Agents, p. 132-139, New York, 1998.

CHEN, P.; XIE, H.; MASLOV, S.; REDNER, S. Finding Scientific Gems with Google's PageRank Algorithm. **Informetrics**. v.41, n.1, p. 8-15, 2007.

CHEN, C.; SONG, I.; YUAN, X.; ZHANG, J. The Thematic and Citation Landscape of Data and Knowledge Engineering (1985-2007). **Data & Knowledge Engineering**. v.67, n.2, p. 234-259, 2008.

COSTAS, R.; BORDONS, M. Is g-index better than h-index? An exploratory study at the individual level. **Scientometrics**. v.77, n.2, p. 267-288, Nov. 2008.

DING, Y.; YAN, E.; FRAZHO, A.; CAVERLEE, J. PageRank for Ranking Authors in Co-citation Networks. **Journal of the American Society for Information Science and Technology**. v.60, n.11, p. 2229-2243, 2009.

EGGHE, L. An improvement of the h-index: The g-index. ISSI Newsletter. Local, v.2, n.1, p. 8-9, 2006a.

EGGHE, L. Theory and Practise of the g-index. Scientometrics. [S.l.], v.69, n.1, p.131-152, 2006b.

FRANCESCHET, M.; COSTANTINI, A. The First Italian Research Assessment Exercise: A bibliometric perspective. **Journal of Informetrics**. v.5, n.2, p.275-291, 2011.

FU, Y.; XIANG, R.; LIU, Y.; ZHANG, M.; MA, S. Finding Experts Using Social Network Analysis. In Proceedings of the International Conference on Web Intelligence, Silicon Valley, USA, 2007.

GARFIELD, E. Citation Indexes For Science: A New Dimension in Documentation through Association of Ideas. Science, v.122, n.3159, p. 108-111, 1955.

GASPARINI, I.; PERNAS, A. M.; PIMENTA, M. S.; OLIVEIRA, J. P. M.; KEMCZINSKI, A; CAVALHEIRO, G. G. H. m-AdaptWeb: An Adaptive E-Learning Environment Facing Mobility - Adaptation and Recomendation Processes Based on Context. Proceedings of 4th International Conference on Computer Supported Education, p. 395-400, Porto, Portugal, 2012.

GAUCH, S.; SPERETTA, M.; CHANDRAMOULI, A.; MICARELLI, A. User profiles for personalized information access. In The adaptive web. LN in Computer Science 4321. Springer-Verlag, Berlin, Heidelberg p. 54-89, 2007.

GRUBER, T. R. Toward Principles For The Design Of Ontologies Used For Knowledge Sharing. International Journal of Human-Computer Studies, v.43, n.4-5, p. 907-928, 1995.

GUYON, I.; ELISSEEFF, A. An introduction to variable and feature selection. Journal of Machine Learning, v.3, p. 1157-1182, 2003.

HALL, M.; FRANK, E.; HOLMES, G.; PFAHRINGER, B.; REUTEMANN, P.; WITTEN, I. H. The WEKA Data Mining Software: An Update; SIGKDD Explorations, v.11, n.1, p.10-18, 2011.

HANNEL, K.; LIMA, J. V. Qualificação de Pesquisadores por área da Ciência da Computação com Base em uma Ontologia de Perfil. 13th Brazilian Symposium on Multimedia and the Web, Gramado, RS, Brasil, 2007.

HARZING, A. W. (2007) Publish or Perish, disponível em http://www.harzing.com/pop.htm

HIRSCH, J. E. An Index to Quantify an Individual's Scientific Research Output. Proceedings of the National Academy of Science, v.102 n.46, p.16569-16572, 2005.

JIN, B. The AR-Index: Complementing the H-Index. International Society for Scientometrics and Informetrics (ISSI Newsletter), v.3 n.1, p.6, 2007.

JIN, B.; LIMING, L.; ROUSSEAU, R.; EGGHE, L. The R- and AR-indices: Complementing the H-Index. Chinese Science Bulletin. v.52 n.6, p.855-863, 2007.

JOACHIMS, T.; FREITAG, D.; MITCHELL, T. M. Web watcher: A tour guide for the world wide web. In IJCAI v.1, p. 770-777, 1997.

JØSANG, A.; ISMAIL, R.; BOYD, C. A Survey of Trust and Reputation Systems for Online Service Provision. **Decision Support Systems Journal.** Amsterdam, v.43, n.2, p. 618-644, mar. 2007.

KIM, H.-N.; HA, I.; LEE, S.-H; JO, G.-S. A Collaborative Approach to User Modeling for Personalized Content Recommendations. In Proceedings of the 11th International Conference on Asian Digital Libraries (ICADL 2008). Beijing, China, 2008.

KOBSA, A.; WAHLSTER, W. User Models in Dialog Systems. Springer Verlag, Heidelberg, Berlin, 1989.

KORN, A.; SCHUBERT, A.; TELCS, A. Lobby index in networks. Physica A, Local, v.388, n.11, p. 2221-2226, 2009.

KRAPIVIN, M.; MARCHESE, M; CASATI, F. Exploring and Understanding Scientific Metrics in Citation Networks. **Complex Sciences**. v.5, p. 1550-1563, 2009.

KUFLIK, T.; SHOVAL, P. Generation of user profiles for information filtering - Research agenda. In Proceedings of the 23th Annual International ACM SIGIR Conference on Research and Development Information Retrieval. p. 313-315. Athens, Greece, 2000.

KURKI, T.; JOKELA, S.; SULONEN, R.; M. TURPEINEN. Agents in delivering personalized content based on semantic metadata. In Proceedings of AAAI Spring Symposium Workshop on Intelligent Agents in Cyberspace, 1999.

LANG, K. NewsWeeder - Learning to Filter Netnews. In Proceedings of the 12th International Conference on Machine Learning, pg. 331–339. Morgan Kaufmann publishers Inc.: San Mateo, CA, USA, 1995.

LI, J.; TANG, J.; ZHANG, J.; LUO, Q.; LIU, Y.; HONG, M. EOS - Expertise Oriented Search Using Social Networks. In Proceedings of the 16th International Conference on WWW. Banff, Canada, 2007.

LI, J.; SANDERSON, M.; WILLETT, P.; NORRIS, M.; OPPENHEIM, C. Ranking of library and information science researchers: Comparison of data sources for correlating citation data, and expert judgments. **Journal of Informetrics**. Local, v.4, n.4, p. 554-563, 2010.

LI, Y.; ZHONG, N. Ontology-Based Web Mining Model: Representations of User Profiles. In Proceedings of the IEEE/WIC International Conference on Web Intelligence, Washington, DC, USA, 2003.

LIMA, H.; SILVA, T. H. P.; MORO, M. M.; SANTOS, R. L.T.; MEIRA JR, W.; LAENDER, A. H. F. Aggregating Productivity Indices for Ranking Researchers Across Multiple Areas. In Proceedings of the 13th ACM/IEEE-CS Joint Conference on Digital Libraries (JCDL '13). ACM, New York, NY, USA, 97-106, 2013.

LOH, S.; GRANADA, R. L.; LITCHNOW, D.; WIVES, L. K.; OLIVEIRA, J. P. M.; LORENZI, F. Using Scientific Publications to Identify People. Lecture Notes in Business Information Processing, v.45, p. 229-241, 2010.

LOPES, G. R.; DA SILVA, R.; MORO, M. M.; OLIVEIRA, J. P. M. Scientific Collaboration in Research Networks: A Quantification Method by Using Gini Coefficient. International Journal of Computer Science & Applications, v.9, p.15-31, 2012.

LOPES, G. R.; MORO, M. M.; DA SILVA, R.; BARBOSA, E. M.; OLIVEIRA, J. P. M. Ranking Strategy for Graduate Programs Evaluation. Proceedings of The 7th International Conference on Information Technology and Application (ICITA 2011), Sydney, Australia, 2011.

MAKINO, J. Productivity of Research Groups-Relation Between Citation Analysis and Reputation Within Research Communities. **Scientometrics**. [S.l.], v.43, n.1, p. 87-93, set. 1998.

MCTEAR, M. Arti¢cial Intelligence Review. Special Issue on User Modeling, v.7, n.3., 1993.

MENEZES, G. V.; ZIVIANI, N.; LAENDER, A. H. F.; ALMEIDA, V. A Geographical Analysis of Knowledge Production in Computer Science. In Proceedings of the 18th International Conference on WWW. Madrid, Spain, 2009.

MIDDLETON, S. E.; SHADBOLT, N. R.; DE ROURE, D. C. Ontological user profiling in recommender systems. ACM Transactions Information Systems, v.22, n.1, p. 54-88, 2004.

MLADENIC, D. Personal webwatcher: Design and implementation. In Technical Report IJS-DP-7472, School of Computer Science, Carnegie Mellon University, Pittsburgh, USA, 1996.

MOED, H. F. New developments in the use of citation analysis in research evaluation. **Archivum Immunologiae et Therapiae Experimentalis**. Local, v.57, n.1, p. 13-18, 2009.

MONTANER, M. Collaborative Recommender Agents Based On Case-Based Reasoning and Trust. PhD thesis. Universitat de Girona, September, 2003.

MOUKAS, A. Amalthaea: Information discovery and filtering using a multiagent evolving ecosystem. In Journal of Applied Artificial Intelligence, v.11, n.5, p. 437-457,1997.

MOUKAS, A.; MAES, P. Amalthaea: An evolving multi-agent information filtering and discovering system for the WWW. In Autonomous Agents and Multi-Agent Systems, v.1 n.1, p. 59-88, 1998.

NORUZI, A. Google Scholar: The New Generation of Citation Indexes. LIBRI, v.55, n.4, p. 170-180, 2005.

OLIVEIRA, E. A. et al. Comparison of Brazilian researchers in clinical medicine: are criteria for ranking well-adjusted? Scientometrics v.90 n.2, p.429-443, 2012.

PAGE, L.; BRIN, S.; MOTWANI, R.; WINOGRAD, T. The PageRank citation ranking: Bringing order to the Web. Tech. report, Stanford University, 1998.

PATTERSON, M.; HARRIS, S. The relationship between reviewers' quality-scores and number of citations for papers published in the journal physics in medicine and biology from 2003-2005. **Scientometrics**. [S.l.], v.80, n.2, p. 343-349, 2009.

PAZZANI, M. J.; MURAMATSU, J.; BILLSUS, D. Syskill & Webert: Identifying interesting websites. In Proceedings of 13th National Conference on Artificial Intelligence, 1996.

PRETSCHNER, A.; GAUCH, S. Ontology Based Personalized Search. In Proceedings of the 11th IEEE International Conference on Tools with Artificial Intelligence. Washington, DC, USA, p. 391-398, 1999.

PUNNARUT, R; SRIHAREE, G. A Researcher Expertise Search System Using Ontology-Based Data Mining. In Proceedings of the 17th Asia-Pacific Conference on Conceptual Modelling. Darlinghurst, Australia, p. 71-78, 2010.

QUINLAN, J. R. C4.5 Programs for Machine Learning, San Diego, CA: Morgan Kaufmann Publishers. 1993.

RAZMERITA, L.; ANGEHRN, A.; MAEDCHE, A. Ontology-Based User Modeling For Knowledge Management Systems. In Proceedings of the 9th International Conference on User modeling. Springer-Verlag, Berlin, Heidelberg, p. 213-217, 2003.

RESNICK, P.; LACOVOU, N.; SUCHAK, M.; BERGSTORM, P.; RIEDL, J. GroupLens: An Open Architecture for Collaborative Filtering of Netnews. In Proceedings of ACM Conference on Computer Supported Cooperative Work, p. 175-186, Chapel Hill, North Carolina, 1994.

RICH, E. Building and Exploiting UserModels. Ph.D. Thesis Department of Computer Science, Carnegie-Mellon University, Pittsburgh, PA, 1979.

RINIA, E.; VAN LEEUWEN, T.; VAN VUREN, H.; VAN RAAN, A. Comparative analysis of a set of bibliometric indicators and central peer review criteria: Evaluation of condensed matter physics in the Netherlands. **Research Policy**, v.27, n.1, p. 95-107, 1998.

ROUSSEAU, R.; JIN, B. H. The age-dependent h-type AR(2)-index: Basic properties and a case study. Journal of the American Society for Information Science and Technology, v.59, n.14, p. 2305-2311, 2008.

SCHOEFEGGER, K., 2011. A User Modeling Approach to Support Knowledge Work in Socio-Computational Systems. Proceedings of the 19th Conference on User Modeling, Adaptation and Personalization (UMAP 2011). Girona, Spain, 2011.

SHETH, B. D. A Learning Approach to Personalized Information Filtering. Master of Science in Computer Science and Engineering, MIT, 1994.

SMALL, H. Cocitation in science literature: New measures of relationship between two documents. Journal of the American Society for Information Science, v.24, n.4, p. 265-269, 1973.

SPILKI, F. R. Perfil dos Bolsistas de Produtividade do Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) na Área de Medicina Veterinária. Pesquisa Veterinária Brasileira, v.33 n.2, p.205-213, 2013.

STEFANI, A.; STRAPPARAVA, C. Personalizing access to web sites: The siteIF project. In Proceedings of 2nd Workshop on Adaptive Hypertext and Hypermedia, Pittsburg, USA, 1998.

SUGIYAMA, K.; KAN, M. Scholarly Paper Recommendation Via User's Recent Research Interests. In Proceedings of the 10th JCDL. Surfer's Paradise, Australia, 2010.

SUGIYAMA, K.; KAN, M. Serendipitous Recommendation for Scholarly Papers Considering Relations Among Researchers. In Proceeding of the 11th JCDL. Ottawa, Canada, 2011.

TADELIS, S. Firm Reputation with Hidden Information. **Economic Theory**. v.21, n.2-3, p. 635-651, 2003.

TANG, J.; ZHANG, D.; YAO, L. Social Network Extraction of Academic Researchers. In Proceedings of 7th ICDM. Omaha, USA, 2007.

TANG, J.; ZHANG, J.; YAO, L; LI, J. Extraction and Mining of an Academic Social Network. In Proceedings of the 17th International Conference on WWW. Beijing, China, 2008-a.

TANG, J.; ZHANG, J.; YAO, L.; LI, J.; ZHANG, L.; SU, Z. ArnetMiner - Extraction and Mining of Academic Social Networks. Proceeding of the 14th KDD. Las Vegas, USA, 2008-b.

TRAJKOVA, J.; GAUCH, S. Improving Ontology-Based User Profiles. In Proceedings of Recherche d'Information Assistée par Ordinateur - RIAO 2004, University of Avignon (Vaucluse), France, p. 380-389, 2004.

TRAJKOVA, J.; GAUCH, S. Improving Ontology-Based User Profiles. M.S. Thesis. EECS, University of Kansas. Lawrence, 2003.

VAN RAAN, A. Comparison of the Hirsch-index with standard bibliometric indicators and with peer judgment for 147 chemistry research groups. **Scientometrics**. [S.l.], v.67, n.3, p. 491-502, 2006.

VINKLER, P. Eminence of scientists in the light of h-index and other scientometric indicators. **Journal of Information Science**. [S.l], v.33, n.4, p. 481-491, 2007.

WAINER, J.; VIEIRA, P. Correlations between Bibliometrics and Peer Evaluation for All Disciplines: The Evaluation of Brazilian Scientists. **Scientometrics.** [S.l.], v.96, n.2, p.395-410, out. 2013.

WALTMAN, L.; VAN ECK, N.; VAN LEEUWEN, T.; VISSER, M.; VAN RAAN, A. On the correlation between bibliometric indicators and peer review: Reply to Opthof and Leydesdor. **Scientometrics**. v.88, n.3, p.1017-1022, 2011.

WHITE, H.D.; Mccain, K.W. Visualizing a discipline: An author cocitation analysis of information science 1972-1995. Journal of the American Society for Information Science, v.49, p. 327-355, 1998.

WIDYANTORO, D.; YIN, J.; NASR, M.; YANG, L.; ZACCHI, A.; YEN, J. Alipes: A swift messenger in cyberspace, 1999. In Proceedings of Workshop on Intelligent Agents in Cyberspace, p. 62-67, Stanford, USA, 1999.

YAN, T.; GARCIA-MOLINA, H. SIFT – A Tool for Wide-Area Information Dissemination. In Proceedings of the USENIX 1995 Technical Conference Proceedings. Berkeley, USA, 1995.

YAN, X. B.; ZHAI, L.; FAN, W. G. C-index: A weighted network node centrality measure for collaboration competence. **Journal of Informetrics**. Local, v.7, n.1, p. 223-239, 2013.

YE F. Y.; ROUSSEAU, R. Probing the h-core: an investigation of the tail-core ratio for rank distributions. Scientometrics. Local, v.84, p. 431-439, 2010.

Ye, F. Y. A unification of three models for the h-index. Journal of the American Society for Information Science and Technology, v.62, n.1, p. 205-207, 2011.

ZHAI, L.; YAN, X.; ZHU, B. The H l - index: improvement of H-index based on quality of citing papers. Scientometrics, p. 1-11, 2013.

ZHANG, C. The e-index, complementing the h-index for excess citations. PLoS ONE, v.5 n.5, 2009.

ZHANG, C. T. A novel triangle mapping technique to study the h-index based citation distribution. Sci Rep v.3, p. 1023, 2013-a.

ZHANG, C. T. The h'-Index, Effectively Improving the h-Index Based on the Citation Distribution. PLoS ONE v.8, n.4, e59912, 2013-b.

ZHANG, H.; SONG, Y.; SONG, H. Construction of Ontology-Based User Model for Web Personalization. In Proceedings of the 11th International Conference on User modeling. Lecture Notes in Computer Science Volume v.4511, p. 67-76, 2007.

ZHANG, J.; TANG, J.; LI, J. Expert Finding in a Social Network. Proceedings of 12th Database Systems for Advanced Applications. Bangkok, Thailand, 2007.