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

RODRIGO OCTAVIO RECH

Um Modelo de Pontuação na Busca de Competências Acadêmicas de Pesquisadores

Dissertação apresentada como requisito parcial para a obtenção do grau de Mestre em Ciência da Computação

Prof. Dr. José Valdeni de Lima
Orientador

Porto Alegre, abril de 2007.

CIP – CATALOGAÇÃO NA PUBLICAÇÃO

Rech, Rodrigo Octavio

Um Modelo de Pontuação na Busca de Competências Acadêmicas de Pesquisadores / Rodrigo Octavio Rech – Porto Alegre: Programa de Pós-Graduação em Computação, 2007.

92 f.:il.

Dissertação (mestrado) – Universidade Federal do Rio Grande do Sul. Programa de Pós-Graduação em Computação. Porto Alegre, BR – RS, 2007. Orientador: José Valdeni de Lima.

1.Avaliação acadêmica. 2.Funções de similaridade. 3.Extração de dados na Web. 4. Recuperação de informação. I. Lima, José Valdeni de. II. Título.

AGRADECIMENTOS

Gostaria de agradecer a todas as pessoas que de uma forma ou de outra me ajudaram a concluir este trabalho. Em primeiro lugar, agradeço a minha família pelo apoio e incentivo: obrigado Carlos (pai), Marilú (mãe), Paulo e Tiago (irmãos).

Do meio acadêmico, agradeço a todos os professores pelos quais tive a oportunidade de enriquecer meus conhecimentos durante esta jornada. Em especial, agradeço ao meu orientador José Valdeni de Lima, por ser paciente e sempre acreditar nesta pesquisa. Não poderia deixar de agradecer aos colegas do grupo de pesquisa SIGHA: Tiago, Nick, Carlos, Adriana, Maximira, Fábio, Dóris, Edicarsia, Elmário e Kelly – valeu pessoal!!!

Agradeço também a CAPES pela bolsa de pesquisa.

Por fim, agradeço aos amigos do peito Júlio, Marcelo, Diego, Roger, Daniel Kerber e Marcel, pois a amizade e companheirismo de vocês são fundamentais em minha vida.

SUMÁRIO

LISTA DE ABREVIATURAS E SIGLAS

API	Application Programming Interface
APP	Artigos publicados em periódicos
CAPES	Coordenação de Aperfeiçoamento de Pessoal de Nível Superior
BD	Banco de Dados
CC	Coeficiente de competência
CC_c	Coeficiente de competência considerando indicadores do currículo
CC_b	Coeficiente de competência considerando indicadores da produção bibliográfica
C&T	Ciência e Tecnologia
CLP	Capítulos de livros publicados
CNPq	Conselho Nacional de Desenvolvimento Científico e Tecnológico
CV	Curriculum Vitae
DEA	Análise por Envoltória de Dados
DNA	Ácido Desoxirribonucleico
DOM	Document Object Model
DTD	Document Type Definition
ECT	Embedded Catalog Tree
EQARI	Evento Qualidade A e Repercussão Internacional
EQARN	Evento Qualidade A e Repercussão Nacional
EQARL	Evento Qualidade A e Repercussão Local
EQBRI	Evento Qualidade B e Repercussão Internacional
EQBRN	Evento Qualidade B e Repercussão Nacional
EQBRL	Evento Qualidade B e Repercussão Local
EQCRI	Evento Qualidade C e Repercussão Internacional
EQCRN	Evento Qualidade C e Repercussão Nacional
EQCRL	Evento Qualidade C e Repercussão Local
GUI	Grafical User Interface

HEL	HTML Extraction Language
HTML	HyperText Markup Language
IR	Information Retrieval
ISBN	International Standard Book Number
ISI	Institute for Scientific Information
ISSN	International Standard Serial Number
LPO	Livros publicados ou organizados
Mo-DROP	Modelo para Determinação da Relevância da Opinião
NSF	National Science Foundation
NSL	Nested String List
OAD	Orientações em andamento de doutorado
OAM	Orientações em andamento de mestrado
OCD	Orientações concluídas de doutorado
OCM	Orientações concluídas de mestrado
OOA	Outras orientações em andamento
OOC	Outras orientações concluídas
OPB	Outros tipos de produção bibliográfica
OPBTC	Outras participações em bancas de trabalhos de conclusão
OPTAP	Outros tipos de produção técnica (apresentação de trabalhos)
OPTOE	Outros tipos de produção técnica (organização de eventos)
OPTRP	Outros tipos de produção técnica (relatórios de pesquisa)
PBCJ	Participações em bancas de comissões julgadoras
PBTCD	Participações em bancas de trabalhos de conclusão de doutorado
PBTCM	Participações em bancas de trabalhos de conclusão de mestrado
PE	Participações em eventos
PQARI	Periódico Qualidade A e Repercussão Internacional
PQARN	Periódico Qualidade A e Repercussão Nacional
PQARL	Periódico Qualidade A e Repercussão Local
PQBRI	Periódico Qualidade B e Repercussão Internacional
PQBRN	Periódico Qualidade B e Repercussão Nacional
PQBRL	Periódico Qualidade B e Repercussão Local
PQCRI	Periódico Qualidade C e Repercussão Internacional
PQCRN	Periódico Qualidade C e Repercussão Nacional
PQCRL	Periódico Qualidade C e Repercussão Local
PTEC	Produtos tecnológicos

RCIT	Razão entre número total de citações e número total de trabalhos
RR	Recommender's Rank
SCI	Science Citation Index
SW	Software
TJR	Textos em jornais ou revistas
TOTCIT	Número total de citações
TPAE	Trabalhos publicados em anais de eventos
TT	Trabalhos técnicos
UFRGS	Universidade Federal do Rio Grande do Sul
URL	Universal Resource Locator
XHTML	eXtensible Hypertext Markup Language
XML	eXtensible Markup Language
XPath	XML Path Language
XQuery	XML Query Language
XSLT	eXtensible Stylesheet Language Transformations

LISTA DE FIGURAS

LISTA DE TABELAS

RESUMO

Esta pesquisa descreve um modelo para descobrir e pontuar competências acadêmicas de pesquisadores, baseado na combinação de indicadores quantitativos que permitem mensurar a produção acadêmica dos cientistas.

Um diferencial do modelo é a inclusão de indicadores quantitativos relacionados com a importância da produção bibliográfica dos pesquisadores. Estes indicadores possibilitam uma avaliação da produção considerando aspectos como repercussão na comunidade acadêmica e nível dos veículos de publicação.

A pesquisa também contribui com a especificação de uma arquitetura flexível e extensível fundamentada em técnicas de extração de dados na Web e casamento aproximado de dados (através de funções de similaridade). A arquitetura foi implementada em um sistema Web cuja principal característica é a integração de diversas tecnologias *open source*. O sistema desenvolvido permite que qualquer pesquisador avalie quantitativamente sua produção científica, automatizando diversos aspectos relacionados à tarefa de avaliação, como a obtenção dos indicadores e a integração das diferentes bases de informações.

Palavras-Chave: avaliação acadêmica, funções de similaridade, extração de dados na Web, recuperação de informação.

A SCORE-BASED MODEL FOR ASSESSING ACADEMIC RESEARCHERS COMPETENCES

ABSTRACT

The present research describes a model that aims finding out and scoring academic researchers skills or competences based on the combination of quantitative indicators that make it possible to measure the production of academic scientists.

A special feature concerning our model is the inclusion of quantitative indicators related to the importance of the researchers' bibliographic production. These indicators allow the evaluation of the production considering both the outcome it has had in the academic community, and the quality level of the place it was published.

The study also presents a flexible and extensible architecture specification based on techniques of web data extraction, and on approximate data matching (which is carried out through similarity functions). The architecture has been implemented in a web system whose main feature relies on the integration of several open-source technologies. The developed system allows any researcher to evaluate his/her own scientific production in quantitative terms, automating as well the so many aspects regarding the evaluation task, by making it easier to obtain the indicators and to integrate the different information databases, for instance.

Keywords: academic evaluation, similarity functions, web data extraction, information retrieval.

1 INTRODUÇÃO

A rápida evolução das redes de computadores, especialmente a partir do surgimento da *World Wide Web*, em conjunto com o avanço das tecnologias para desenvolvimento de aplicações distribuídas, possibilitaram a construção de sistemas de informação colaborativos inovadores.

Um exemplo de aplicação colaborativa que revolucionou a forma de se produzir e organizar conhecimento é a Wikipedia[1], uma enciclopédia livre e gratuita que adota um modelo no qual qualquer usuário pode editar qualquer artigo da enciclopédia. Porém, esta natureza extremamente aberta da Wikipedia não fornece indicadores de que um usuário tenha conhecimento e/ou experiência na área de um artigo de sua autoria. Em diversos domínios específicos, estes indicadores são essenciais para determinação da qualidade dos documentos produzidos.

A proposta de Oliveira et al. (2005) sugere a construção de um sistema para editoração colaborativa de artigos acadêmicos, com revisão interativa pelos pares, discussão pública e processo de avaliação aberta. Neste caso, segundo Oliveira et al., dispor das competências acadêmicas dos usuários do sistema (tais como autores, comentaristas e revisores), em relação aos diversos tópicos em que podem opinar, é de fundamental importância para o processo de avaliação aberta.

Isto porque a idéia central da proposta de Oliveira et al. é permitir que um ambiente automatizado avalie a qualidade dos documentos produzidos baseado na qualificação dos usuários que participaram do processo de editoração colaborativa. Portanto, faz-se necessário integrar no sistema um modelo que capture as qualificações ou competências acadêmicas destes usuários. A elaboração deste modelo é um desafio, e constitui na motivação e foco principal da presente dissertação.

Além disso, atualmente está cada vez mais fácil compartilhar artigos acadêmicos em larga escala. Segundo dados da NSF (2007), em 2003 cerca de 699.000 artigos foram publicados nos principais veículos de divulgação científica. Grande parte desta produção acadêmica pode ser consultada *online*. Sistemas como o CiteSeer[2] e o Google Scholar[3] são exemplos de repositórios digitais e mecanismos de busca de artigos científicos na Web. Entretanto, uma questão crítica que se coloca é a determinação da relevância dos artigos, ou seja, de que forma um sistema poderia classificar (em níveis como iniciante, intermediário ou avançado), ordenar e recomendar artigos de acordo

---

[1] http://www.wikipedia.org, acessado em mar. 2007.

[2] http://citeseer.ist.psu.edu, acessado em mar. 2007.

[3] http://scholar.google.com, acessado em mar. 2007.

com critérios que justifiquem o impacto dos mesmos na comunidade acadêmica. O presente trabalho assume que a relevância de um artigo provavelmente está relacionada com as qualificações de seus autores. Portanto, uma forma de inferir esta relevância passa pela construção de um modelo que permita capturar as competências acadêmicas dos pesquisadores.

Outro aspecto motivador desta pesquisa engloba o processo de avaliação da ciência. Conforme Davyt e Velho (2000), avaliar e monitorar a atividade científica são práticas rotineiras e indissociáveis da gestão de ciência e tecnologia, pois permitem corrigir rumos e orientar a aplicação de recursos no setor.

Assim sendo, existe uma grande demanda por parte das agências de fomento (públicas e privadas) em quantificar a produção científica de centros de pesquisa. As próprias universidades cada vez mais necessitam de ferramentas que possam auxiliar na gestão das competências de seus pesquisadores, na busca de diferenciais competitivos e excelência acadêmica.

Uma das principais formas de avaliação da ciência emprega métodos baseados em indicadores quantitativos para mensurar a produtividade acadêmica de um pesquisador. Segundo Davyt e Velho, tais indicadores medem a quantidade bem como o impacto da produção de um cientista. Logo, se um sistema informatizado for construído com base em um modelo que considera múltiplos indicadores quantitativos relevantes, ele poderá contribuir no processo de avaliação da ciência (NIEDERAUER, 2002).

1.1 Objetivos

O objetivo principal desta dissertação é descrever um modelo capaz de descobrir e pontuar competências acadêmicas de pesquisadores. Para isso, o modelo precisa oferecer mecanismos adequados à avaliação quantitativa da produção científica dos pesquisadores, considerando aspectos como produtividade e impacto da produção na comunidade acadêmica.

Como objetivo complementar, especifica-se o desenvolvimento de um sistema que implemente o referido modelo e permita a realização de experimentos (possibilitando assim a aplicação do modelo em uma situação real).

1.2 Organização do Texto

A pesquisa está organizada em 6 capítulos, incluindo os capítulos da Introdução e Conclusão.

No Capítulo 2 é apresentado um levantamento bibliográfico, onde são relatados trabalhos relacionados e pesquisas sobre tópicos fundamentais para a realização desta dissertação, como: processo de avaliação em ciência e tecnologia, integração de bases de dados através de funções de similaridade e extração de dados na Web.

Em seguida, o Capítulo 3 descreve em detalhes o modelo de pontuação na busca de competências acadêmicas de pesquisadores, relatando os indicadores quantitativos considerados e as métricas utilizadas para calcular as pontuações.

O Capítulo 4 relata a construção de um sistema Web que implementa o modelo. A arquitetura e as principais funcionalidades do sistema são apresentadas em detalhes, assim como as dificuldades e limitações encontradas durante a implementação.

Já o Capítulo 5 expõe os experimentos realizados (e seus resultados), que tiveram dois objetivos principais: (i) avaliar a performance das funções de similaridade na integração das bases de dados consideradas; e (ii) aplicar o modelo em uma situação real.

Finalmente, o Capítulo 6 tece as conclusões e aponta os trabalhos futuros da pesquisa.

2 LEVANTAMENTO BIBLIOGRÁFICO

Para atingir os objetivos desta dissertação, os seguintes temas foram investigados:

- Processo de avaliação em ciência e tecnologia – estudo introdutório sobre formas de avaliação da ciência, destacando o processo de revisão pelos pares e os indicadores quantitativos da área da cienciometria;
- Trabalhos relacionados – identificação de pesquisas com objetivos similares ao presente trabalho. Nesta categoria encontram-se pesquisas que visam ou necessitam determinar a competência acadêmica (ou então a "autoridade", o "*expertise*", a "relevância de opinião", o "impacto", a "produtividade", a "eficiência", a "performance") de pesquisadores para diversos propósitos, como recomendação de artigos científicos, alocação de recursos financeiros, formação de comunidades virtuais, entre outros;
- Funções de similaridade – um problema freqüente na utilização de bases de dados distintas (sem padronização na entrada dos dados e sem relacionamentos através de atributos chaves) surge quando se necessita integrar estas bases para obtenção de informação. Nesta dissertação foi preciso integrar três bases de dados distintas: uma contendo a produção científica dos pesquisadores; outra contendo dados adicionais da produção (número de citações); e uma contendo a classificação dos veículos de publicação. Assim sendo, procedeu-se um estudo sobre funções de similaridade entre *strings* (cadeias de caracteres), a fim de identificar abordagens adequadas a este problema;
- Extração de dados na Web – a Internet é um dos maiores repositórios de informação da humanidade. Todavia, grande parte dos dados ainda estão disponibilizados de forma não-estruturada, o que dificulta a obtenção e o processamento das informações por computadores. Como a instanciação do modelo descrito no Capítulo 3 necessita de informações disponíveis apenas na Internet, foi necessário estudar técnicas de extração de dados na Web a fim de identificar soluções apropriadas a este problema.

2.1 Processo de Avaliação em Ciência e Tecnologia (C&T)

Avaliar a produção científica de pesquisadores não é uma tarefa simples, porém de fundamental importância em um mundo onde pesquisadores competem por recursos limitados para desenvolver suas pesquisas, por reconhecimento e por prestígio (NIEDERAUER, 2002). É através da avaliação – seja de artigos para publicação, seja do currículo de um pesquisador, seja de um projeto de pesquisa submetido para

financiamento – que se definem rumos, tanto do próprio conteúdo da ciência quanto das instituições a ela vinculadas (DAVYT e VELHO, 2000).

Segundo Niederauer (2002), o processo de avaliação em ciência e tecnologia ocorre de duas formas: pelos pares, qualitativa e subjetiva por natureza; e pelo uso de métodos quantitativos baseados em procedimentos bibliométricos e cienciométricos.

A avaliação pelos pares é o processo pelo qual os próprios cientistas avaliam seus colegas. É a forma mais tradicional, difundida e aceita de avaliação da atividade e do progresso científico, sendo que seu conceito surgiu desde a criação dos primeiros periódicos científicos no século XVII (NIEDERAUER, 2002). Esta noção de que apenas os próprios cientistas podem avaliar o trabalho de seus colegas permanece até os dias atuais, sendo que a grande maioria dos periódicos e revistas científicas adotam o julgamento por pares nas decisões sobre publicação. Ainda, o processo é usualmente empregado pelas agências de fomento (públicas ou privadas) como método de alocação de recursos para financiamento de pesquisas. Neste caso, o argumento relaciona-se com o conceito de "contrato social entre ciência e sociedade", onde se acredita que o investimento público na ciência de qualidade, mais cedo ou mais tarde, retorna para a própria sociedade, bastando garantir que, de fato, seja apoiada a "ciência de qualidade", que só pode ser identificada pelos próprios cientistas (DAVYT e VELHO, 2000).

O sistema de revisão pelos pares levanta diversas questões relevantes, pois sua base está fundamentada na imparcialidade e competência dos avaliadores (que muitas vezes competem com o avaliado por financiamentos, espaço em publicações, premiações, etc.). Este fato acarreta várias conseqüências, estudadas em Davyt e Velho (2000) e aqui resumidas:

- Privilégio conferido a pesquisadores e instituições de prestígio ("efeito halo"), em detrimento a minorias éticas, mulheres, pesquisadores jovens, pesquisadores de centros acadêmicos de menor prestígio e até mesmo de países diferentes dos avaliadores;
- Resistência a idéias inovadoras, favorecendo linhas de pesquisa tradicionais;
- Controle da burocracia interna sobre os processos (incluindo os critérios de julgamento e o conteúdo dos pareceres);
- Sigilo dos avaliadores (geralmente num só sentido, ou seja, os julgados não conhecem o nome dos juízes, mas estes sabem quem são os proponentes), o que torna difícil a cobrança de resultados e isenção de julgamento, além de estimular a elaboração de pareceres maliciosos ou simplesmente de má-qualidade;
- Enorme gasto de tempo dos cientistas e de recursos financeiros;
- Promove competição em lugar da cooperação e colaboração entre cientistas.

Ao contrário da avaliação por pares, que é realizada por humanos e está sujeita a aspectos subjetivos e éticos, os métodos quantitativos das disciplinas de bibliometria e cienciometria visam quantificar a produção científica de pesquisadores, instituições acadêmicas ou países através de modelos matemáticos e estatísticos. Segundo Niederauer (2002), em geral os métodos quantitativos são aceitos como alternativas *complementares* à revisão por pares e são extremamente úteis à gestão de ciência e tecnologia.

Para Macias-Chapula (1998), a cienciometria pode ser definida como o estudo dos aspectos quantitativos da ciência enquanto uma disciplina ou atividade econômica. Por envolver dados oriundos das publicações dos cientistas, sobrepõe-se à bibliometria, que estuda aspectos quantitativos da produção, disseminação e uso da informação registrada.

De acordo com Davyt e Velho (2000), os métodos cienciométricos ou cientométricos (do inglês *scientometrics*) começaram a ser utilizados na década de 60, a partir do trabalho seminal de Price (1963), que demonstrou que a ciência podia ser medida e cunhou o termo "ciência da ciência". Nesta década também foi criado o *Institute for Scientific Information* (ISI[4]), que publica uma das principais fontes para análise quantitativa da evolução científica, o *Science Citation Index* (SCI[5]).

Conforme mencionado anteriormente, os métodos quantitativos da cienciometria geralmente são baseados em indicadores bibliométricos. Isto porque a publicação dos resultados das pesquisas é um compromisso dos cientistas, caso os mesmos desejem divulgar suas descobertas científicas, salvaguardar a propriedade intelectual e alcançar a fama (NIEDERAUER, 2002; MACIAS-CHAPULA, 1998). Os indicadores mais conhecidos são (MACIAS-CHAPULA, 1998):

- Número de trabalhos – reflete o tamanho da produção científica, medido pela contagem dos trabalhos e pelo tipo de documentos (livros, artigos, publicações, relatórios, etc.);
- Número de citações – reflete o reconhecimento dos artigos e permite calcular o impacto (relação citação/artigo);
- Co-autoria: reflete o grau de colaboração na ciência, medido pelo número de co-autores nos artigos;
- Número de patentes: reflete as tendências das mudanças técnicas ao longo do tempo e avalia os resultados dos recursos investidos em atividades de pesquisa e desenvolvimento;
- Mapas dos campos científicos e dos países – auxiliam a localizar as posições relativas de diferentes países na cooperação científica global.

Segundo Spinak (1998), cada um dos indicadores pode ser medido de forma simples, relativa ou ponderada. Também é possível medi-los através de séries temporais ou através de medidas de distribuição. Ainda, as medições podem ser feitas em níveis micro (pesquisadores, periódicos individuais), médio (instituições, grupos temáticos) ou macro (países, toda uma disciplina).

Assim como o processo de avaliação por pares, os métodos quantitativos também são alvo de diversas críticas. A primeira delas diz respeito ao fato de que muitos pesquisadores simplesmente rejeitam a idéia de ter sua atividade científica avaliada por um sistema quantitativo, aceitando apenas a avaliação pelos pares (MACIAS-CHAPULA, 1998). Outra questão está relacionada com a cobertura das bases de dados. Conforme Spinak (1998), os procedimentos de seleção de periódicos e revistas para integrar a base de dados do ISI eram parciais e não suficientemente adequados para avaliar a ciência e tecnologia dos países em desenvolvimento, e privilegiam a chamada

---

[4] http://scientific.thomson.com, acessado em mar. 2007.

[5] http://thomsonscientific.com/products/sci/, acessado em mar. 2007.

ciência *mainstream* (feita por países desenvolvidos com tradição em pesquisa). Ainda, cada área de conhecimento possui particularidades de publicação, o que inviabiliza a comparação entre áreas (ROUSSEAU, 1998). Por exemplo, na área da Ciência da Computação é muito valorizada a publicação em anais de conferências internacionais, enquanto que para outras áreas apenas a publicação em periódicos deve ser considerada.

A análise das citações como medida de avaliação de pesquisa também provoca questionamentos. Para Niederauer (2002), as citações nada dizem sobre a "qualidade" de uma pesquisa, somente dão idéia do impacto da mesma nos trabalhos que se seguem. Kostoff (1998) investiga a validade da análise das citações como medida do impacto e qualidade de pesquisa, e aponta diversos fatores negativos: (i) auto-citações sem critérios técnicos justificáveis; (ii) citações políticas, por exemplo, a inclusão de citações de trabalhos de editores de jornais ou potencias revisores apenas para ter maior chance de publicação; ou ainda a criação de "clubes de citações", onde cada membro cita os outros regularmente a fim de aumentar o volume de citações; e (iii) um artigo pode ter um alto nível de citações porque realmente contribuiu na evolução da ciência ou então porque possui erros óbvios que foram identificados por muitos pesquisadores.

Como é possível perceber nesta introdução sobre o assunto, ambas abordagens de avaliação possuem seus prós e contras. Entretanto, segundo Niederauer (2002), uma possível solução para minimizar os aspectos negativos de cada forma de avaliação está na combinação das mesmas. Ainda de acordo com Niederauer (2002), a avaliação pelos pares, apesar de suas falhas, é praticamente insubstituível. Já os métodos quantitativos, se considerarem vários indicadores distintos, podem complementar o perfil do pesquisador avaliado. A pesquisa desenvolvida nesta dissertação contribui principalmente na avaliação quantitativa da ciência.

2.2 Trabalhos Relacionados

2.2.1 Relevância de opinião

A relevância de opinião de um usuário pode ser deduzida de suas competências. Nesta direção, o trabalho de Cazella (2006) aborda uma limitação crítica dos sistemas de recomendação colaborativos: estes não indicam a relevância de opinião dos usuários que auxiliaram no processo de predição de um item recomendado. Cazella propõe um modelo para determinação da relevância da opinião do usuário (Mo-DROP), que emprega uma métrica chamada *Recommender's Rank* (RR). Esta métrica tem por finalidade representar o peso da opinião (nível de *expertise*) do usuário em áreas de interesse do mesmo. A equação 1 calcula o RR, através de uma média aritmética ponderada entre os atributos considerados e seus respectivos pesos:

$$RR = \frac{\sum_{i=1}^{n} an_i * p_i}{\sum_{i=1}^{n} p_i} \quad (1)$$

Cazella aplicou o modelo em um sistema para capturar a relevância de opinião de pesquisadores e adicioná-la no processo de recomendação de artigos científicos. Os atributos selecionados e seus respectivos pesos foram definidos através de um importante experimento. Nele, 25 doutores da área da Ciência da Computação identificaram e ponderaram os indicadores de produção acadêmica que consideravam

importantes para a definição da relevância de opinião de um pesquisador. Os resultados do experimento estão no Anexo A desta dissertação.

O protótipo desenvolvido por Cazella é baseado em sistemas multiagentes e mineração de dados. Ele solicita informações explícitas do usuário (por exemplo, as áreas de interesse e o nível de conhecimento em cada área) e utiliza atributos quantitativos relacionados com seu currículo (por exemplo, o número de artigos publicados em periódicos, a quantidade de orientações concluídas – ver Anexo A com a lista completa de atributos). Na proposta original de Cazella e Alvares (2003), o modelo não diferenciava a importância entre publicações com repercussão nacional ou internacional, porém em Cazella (2006) este critério foi incluído. O sistema calcula um RR (valor entre 0 – nenhum *expertise*, e 10 – *expertise* máximo) para cada área de interesse do usuário, conforme a equação 1.

A pesquisa realizada na presente dissertação parte dos resultados obtidos por Cazella (2006) e procura acrescentar novos aspectos no modelo. Outros indicadores são investigados, com o propósito de avaliar a repercussão dos trabalhos do pesquisador na comunidade acadêmica (através da análise do número de citações). Além disso, o modelo descrito no Capítulo 3 difere-se do modelo Mo-DROP ao separar a avaliação da produção acadêmica do pesquisador em dois critérios: (i) avaliação quantitativa relacionada ao currículo e (ii) avaliação quantitativa relacionada à produção bibliográfica. Esta separação permite obter duas percepções distintas das competências dos pesquisadores: a primeira relaciona-se com a produtividade e a segunda com a importância da produção. Finalmente, nesta dissertação adota-se uma abordagem baseada em funções de similaridade para minimizar os problemas descritos em Cazella (2006) sobre integração das bases de dados contendo a produção dos pesquisadores e a classificação dos veículos de publicação.

2.2.2 Produtividade relativa de pesquisadores utilizando DEA

A obtenção de consenso muitas vezes é uma tarefa extremamente difícil, principalmente quando se está em questão a avaliação dos méritos e competências de pessoas. Sob esta ótica, o trabalho de Niederauer (2002) propõe um modelo (chamado Ethos) baseado na Análise por Envoltória de Dados (DEA) cruzado, com o objetivo de medir o desempenho relativo dos candidatos à Bolsa de Produtividade em Pesquisa concedida pelo CNPq[6].

De acordo com Niederauer, a Análise por Envoltória de Dados pode ser contextualizada como:

> Um método de *benchmarking* utilizado para medir o desempenho relativo de unidades organizacionais semelhantes (países, empresas, departamentos, indivíduos) que utilizam vários insumos para gerar múltiplos produtos, sem necessitar de informação *a priori*, tanto na escolha da métrica (a forma funcional) quanto na determinação dos pesos de cada variável (2002, p. 34.).

A adoção desta abordagem está fundamentada no fato de que os critérios pelos quais os pesquisadores são avaliados para obtenção da bolsa variam de comitê para comitê, pois refletem as particularidades das áreas e a visão que os integrantes dos comitês têm a cada momento. Assim sendo, segundo Niederauer, as métricas baseadas em somas

[6] http://www.cnpq.br, acessado em mar. 2007.

ponderadas não representam a melhor solução, tendo em vista que exigem a intervenção humana na ponderação dos atributos.

Em seu trabalho, Niederauer considera que o pesquisador doutor candidato à bolsa pode ser entendido como um sistema que transforma seu tempo em uma série de produtos, medidos através dos indicadores de produção. Portanto, a variável de entrada (insumos) no modelo Ethos é o tempo de doutorado, e as variáveis de saída (produtos) são os indicadores de produção.

Niederauer demonstrou através de simulações e aplicações do modelo em casos reais que sua proposta gerou resultados satisfatórios. As simulações levaram a conclusões coerentes e os experimentos com dados reais capturaram corretamente as recomendações dos comitês de assessoramento em relação à distribuição das bolsas de produtividade em pesquisa. Como principais contribuições do modelo Ethos, Niederauer destaca:

- Permite que cada pesquisador escolha os indicadores de produção com os quais quer ser comparado, sendo livre de restrições de pesos, aumentando a transparência do processo de concessão da bolsa;
- Permite que o pesquisador também seja avaliado segundo a perspectiva dos demais, por meio da avaliação pelos pares, possível através da aplicação da técnica DEA cruzado.

O modelo descrito no Capítulo 3 desta dissertação adota médias ponderadas para medir a produtividade dos pesquisadores, sendo que os pesos são determinados manualmente. Destaca-se que a implementação do modelo descrita no Capítulo 4 permite que sejam cadastradas diversas combinações de pesos para cada área de pesquisa. Assim sendo, cada área pode "escolher" quais indicadores serão considerados no cálculo, e atribuir a importância para cada um deles. Todavia, a abordagem está sujeita a críticas relacionadas à dificuldade da obtenção de consenso na determinação dos pesos, e considera que a utilização do DEA como feito por Niederauer (2002) deve ser avaliada em trabalhos futuros.

Uma limitação do modelo Ethos é que o mesmo considera apenas indicadores de produção (por exemplo, número de artigos publicados em periódicos, número de orientações concluídas, entre outros). Ao contrário, o modelo da presente dissertação, além de empregar indicadores de produção, inclui outros indicadores quantitativos que permitem mensurar a importância da produção científica do pesquisador. Estes indicadores avaliam a produção em termos da repercussão ou impacto dos trabalhos na comunidade acadêmica e da classificação dos veículos de publicação, oferecendo uma visão mais completa dos perfis dos cientistas.

2.2.3 Mineração de competências, comunidades virtuais e identificação de *expertise*

Rodrigues et al. (2004) propuseram uma técnica para criar comunidades científicas na Web baseada nas competências extraídas a partir das publicações dos pesquisadores.

O objetivo foi facilitar a aquisição, compartilhamento e disseminação de conhecimento científico em um domínio de pesquisa, conectando cientistas com interesses comuns através da sugestão de comunidades virtuais, assumindo que pessoas que trabalham em uma área específica e possuem competências similares desejam compartilhar informação e conhecimento.

A figura 2.1 apresenta a arquitetura para extração de palavras relevantes (que podem indicar as competências) a partir das publicações dos pesquisadores:

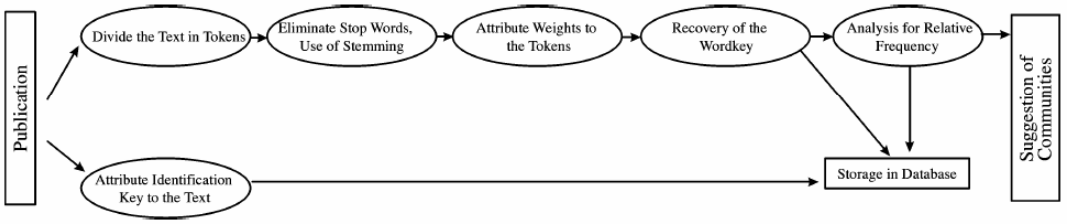

Figura 2.1: Arquitetura para extração de competências (RODRIGUES et al., 2004).

A arquitetura descrita na figura 2.1 foi implementada em um sistema que emprega métodos e técnicas tradicionais das áreas de descoberta de conhecimento em textos e *text mining*. Cada pesquisador possui um diretório onde todas suas publicações são armazenadas. Para cada publicação é atribuído uma chave de identificação (*identification key*). O texto de uma publicação é submetido a um algoritmo que identifica *tokens* (palavras) e elimina palavras irrelevantes (*stop words*). Nesta fase também é empregada uma técnica chamada *stemming*, que remove sufixos das palavras e compara radicais, permitindo que palavras como *connect*, *connection*, *connected* e *connecting* sejam computadas como *connect*. O sistema permite atribuir pesos para cada palavra extraída e computa a freqüência relativa das palavras relevantes. Ao final do processo, as competências (palavras relevantes) e o grau de *expertise* (medida da relevância de uma palavra) são armazenados no banco de dados.

Seguindo uma abordagem semelhante, Borges et al. (2004) desenvolveram uma ferramenta que permite identificar automaticamente *expertises* de usuários e áreas de competência dos mesmos. A ferramenta aplica uma técnica de *text mining* para comparar palavras extraídas do currículo do usuário com palavras relacionadas a conceitos presentes em uma ontologia de domínio da área da Ciência da Computação. A técnica representa currículos e conceitos através de vetores de termos. Assim, a ferramenta utiliza uma função de similaridade que calcula a distância entre dois vetores, avaliando a similaridade entre um currículo e os conceitos presentes na ontologia. Esta proposta foi refinada no trabalho de Ribeiro Junior et al. (2005), que permite atribuir pesos diferenciados para termos relacionados a elementos especiais do currículo (por exemplo, palavras-chaves, áreas de atuação, entre outros).

Tanto a proposta de Rodrigues et al. (2004) como a de Borges et al. 2004 são adequadas para identificação de áreas de experiência ou de interesse, porém não foram desenvolvidas com a intenção de avaliar a produtividade ou competência de pesquisadores. Elas não consideram o uso de indicadores de produção, ao contrário do modelo descrito no Capítulo 3 desta dissertação.

2.3 Funções de Similaridade

2.3.1 Definições

Em diversos sistemas de gerenciamento de informação, freqüentemente faz-se necessário aplicar um processo chamado *casamento aproximado de dados*, cujo objetivo é definir se duas instâncias de dados (cadeias de caracteres – *strings*, tuplas, árvores, etc.) representam o mesmo objeto do mundo real (SILVA et al., 2007). Um

exemplo de aplicação do processo surge em tarefas de integração de bases de dados oriundas de fontes distintas, sem relacionamentos e/ou restrições de integridade.

O processo de casamento aproximado de dados pode ser implementado através de funções de similaridade, definidas na equação 2:

$$f(s,t) \rightarrow r \tag{2}$$

A função *f* da equação 2 recebe como entrada duas instâncias de dados *s* e *t* (geralmente *strings*) e retorna um número real *r*, usualmente normalizado no intervalo $[0,1]$, onde valores próximos a 1 indicam maior similaridade.

A utilização de funções de similaridade implica na definição de um valor de limiar (ou fator de corte – *threshold*), pois o objetivo não é identificar apenas instâncias de dados exatamente iguais, mas também semelhantes. O *threshold* é um valor decimal entre 0 e 1 que deve ser ajustado de acordo com o domínio da aplicação. Quando a similaridade entre *s* e *t* é maior ou igual que o *threshold* informado considera-se que *s* e *t* representam o mesmo objeto.

Um conceito análogo ao conceito de função de similaridade é a definição de função de distância. Neste caso, o valor de retorno, ao invés de medir a similaridade, quantifica a distância (ou diferença) entre os objetos comparados. Ou seja, quanto menor o valor de retorno, menor a diferença (portanto maior a similaridade) entre os objetos. Quando normalizada, a distância pode ser convertida em similaridade e vice-versa.

2.3.2 Exemplos de funções de similaridade

2.3.2.1 Levenshtein (edit distance)

A distância de edição entre duas *strings* é definida como o menor número de operações de edição necessárias para transformar a primeira *string* na segunda. As operações de edição são: inserir um caractere na primeira *string*, deletar um caractere da primeira *string* ou substituir um caractere na primeira *string* por um caractere da segunda *string*. A cada uma dessas operações deve ser atribuído um custo, que no caso da função *Levenshtein* (LEVENSHTEIN, 1966) é igual a 1.

Por exemplo, a distância de edição para transformar a *string s*="*test*" na *string t*="*tent*" é igual a 1, porque apenas uma operação é necessária (substituir o caractere "*s*" em *s* pelo caractere "*n*" de *t*).

Para calcular a distância de edição entre duas *strings*, utiliza-se uma técnica conhecida como programação dinâmica (GUSFIELD, 1997). A idéia básica da técnica consiste em obter a solução de um problema a partir de soluções menores de sub-problemas análogos. Dado duas *strings* S_1 e S_2 de tamanhos *n* e *m* respectivamente, uma matriz *D* de tamanho $(n + 1) \times (m + 1)$ é computada de acordo com o algoritmo da figura 2.2:

```
LevenshteinDistance(s1, s2)
Início
  n := Tamanho(s1)
  m := Tamanho(s2)
  d[0..n, 0..m] // Matriz de inteiros
  i, j, custo // Inteiros

  // Inicializa a primeira linha da matriz
  Para i=0 até n Faça
    d[i,0] := i
  // Inicializa a primeira coluna da matriz
  Para j=0 até m Faça
    d[0,j] := j

  Para i=1 até n Faça
  Início
    Para j=1 até m Faça
    Início
      Se s1[i]=s2[j] Então
        custo := 0
      Senão
        custo := 1
      d[i,j] := Mínimo(d[i-1, j  ] + 1,     // Apagar
                       d[i  , j-1] + 1,     // Inserir
                       d[i-1, j-1] + custo  // Substituir
                      )
    Fim Para
  Fim Para

  Retorna d[n,m]
Fim
```

Figura 2.2: Algoritmo para computar a distância de edição entre duas *strings*.

A figura 2.3 apresenta a matriz calculada para as *strings* S_1="*winter*" e S_2="*writer*":

		w	r	i	t	e	r
	0	**1**	**2**	**3**	**4**	**5**	**6**
w	**1**	0	1	2	3	4	5
i	**2**	1	1	1	2	3	4
n	**3**	2	2	2	2	3	4
t	**4**	3	3	3	2	3	4
e	**5**	4	4	4	3	2	3
r	**6**	5	4	5	4	3	**2**

Figura 2.3: Matriz calculada.

A distância de edição entre as *strings* sempre estará no canto inferior direito da matriz. No caso da figura 2.3, a distância de edição é igual a 2.

É importante ressaltar que este valor representa a distância entre duas *strings*, e não a similaridade. Para calcular a similaridade é preciso aplicar uma função de conversão de escalas de distância $[M_d..L_d]$ para similaridade $[M_s..L_s]$, onde M é o limite mínimo inferior e L é o limite máximo superior. Para a função de distância *Levenshtein*, o valor

de M_d é zero (*strings* idênticas) e o valor de L_d é o tamanho da maior *string*. A equação 3 permite converter as escalas (GROSSI, 2005):

$$\frac{d - M_d}{L_d - M_d} = \frac{s^{-1} - M_s}{L_s - M_s} \tag{3}$$

No caso das *strings* da figura 2.3:

$$\frac{2-0}{6-0} = \frac{s^{-1}-0}{1-0} \Rightarrow s^{-1} = \frac{1}{3}$$

Sabendo que na função de distância quanto maior a distância mais diferentes são as entradas e na função de similaridade quanto maior o valor, mais similares elas serão, pode-se deduzir que $s = 1 - s^{-1} = 0{,}66$, ou seja, as *strings* da figura 2.3 são 66% similares de acordo com a função *Levenshtein*.

2.3.2.2 Smith-Waterman

A função *Smith-Waterman* (SMITH e WATERMAN, 1981) foi desenvolvida com o objetivo de encontrar alinhamentos ótimos entre seqüências de DNA ou proteínas.

Segundo Gusfield (1997), um alinhamento (global) de duas *strings* S_1 e S_2 é obtido inserindo-se espaços (ou o caractere "-") dentro ou no final das *strings* (para que as duas fiquem com o mesmo tamanho), e então colocando-as uma sob a outra de modo que cada caractere ou espaço esteja emparelhado a um único caractere ou espaço da outra *string*, sendo que não podem existir espaços na mesma posição em S_1 e S_2. Para uma definição formal de alinhamento, consulte Ticona (2003). Um exemplo de alinhamento para as *strings "qacdbd"* e *"qawxb"* é (GUSFIELD, 1997):

```
q a c - d b d
q a w x - b -
```

É possível alinhar duas *strings* com tamanhos diferentes de diversas formas, dependendo das posições onde os espaços são inseridos. Para descobrir o melhor alinhamento, utiliza-se um esquema de pontuação. Este esquema é dependente do alfabeto das *strings*, e geralmente penaliza o alinhamento entre um símbolo e um espaço (*gap*), penaliza o alinhamento de dois símbolos diferentes (*mismatch*) e enfatiza o alinhamento de dois símbolos idênticos (*match*). Assim sendo, o melhor alinhamento é aquele que atinge a maior pontuação.

O alinhamento basicamente pode ser de dois tipos: global (alinha as duas *strings* como um todo) e local. O alinhamento local tem como objetivo encontrar *substrings* com alta similaridade, sendo definido da seguinte forma (GUSFIELD, 1997): dado duas *strings* S_1 e S_2, encontre as *substrings* α e β de S_1 e S_2 respectivamente cuja similaridade (melhor alinhamento global) é máxima entre todos os pares de *substrings* de S_1 e S_2.

A função *Smith-Waterman* é utilizada para encontrar o melhor alinhamento local, sendo uma adaptação do algoritmo *Needleman-Wunsch* (NEEDLEMAN e WUNSCH, 1970), que calcula o melhor alinhamento global. A função é computada através da programação dinâmica. Considerando duas *strings* s e t de tamanhos n e m respectivamente, a matriz D de tamanho $(n + 1) \times (m + 1)$ é calculada de acordo com a equação 4:

$$D[i,j] = \max \begin{cases} 0, \\ D[i-1,j-1] + p(s_i, t_j) \\ D[i-1,j] - g \\ D[i,j-1] - g \end{cases} \quad (4)$$

onde $p(s_i, t_j)$ representa uma função de custo e g a penalidade de alinhamento com um *gap*. Os valores de $D[i,0]$ e $D[0,j]$ são inicializados com zero.

A figura 2.4 mostra um exemplo de matriz com as *strings* *s*="*aaaa mnop zzzz*" e *t*="*bbbb mnop yyyy*", *g*=0,5 e função de custo definida da seguinte forma:

$p(s_i, t_j)$ = 1, se o caractere *i* da string *s* for igual ao caractere *j* da string *t*;

$p(s_i, t_j)$ = -2, se o caractere *i* da string *s* for diferente do caractere *j* da string *t*.

		a	a	a	a		m	n	o	p		z	z	z	z
	b	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	b	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	b	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	b	0	0	0	0	0	0	0	0	0	0	0	0	0	0
		0	0	0	0	1	0.5	0	0	0	1	0.5	0	0	0
	m	0	0	0	0	0.5	2	1.5	1	0.5	0.5	0	0	0	0
	n	0	0	0	0	0	1.5	3	2.5	2	1.5	1	0.5	0	0
	o	0	0	0	0	0	1	2.5	4	3.5	3	2.5	2	1.5	1
	p	0	0	0	0	0	0.5	2	3.5	5	4.5	4	3.5	3	2.5
		0	0	0	0	1	0.5	1.5	3	4.5	6	5.5	5	4.5	4
	y	0	0	0	0	0.5	0	1	2.5	4	5.5	5	4.5	4	3.5
	y	0	0	0	0	0	0	0.5	2	3.5	5	4.5	4	3.5	3
	y	0	0	0	0	0	0	0	1.5	3	4.5	4	3.5	3	2.5
	y	0	0	0	0	0	0	0	1	2.5	4	3.5	3	2.5	2

Figura 2.4: Matriz calculada[7].

O maior valor em qualquer célula da matriz corresponde ao melhor alinhamento local. No caso da figura 2.4, o valor é 6. Para calcular a similaridade das *strings*, divide-se este valor pelo comprimento da menor *string* (que corresponde ao máximo valor possível de alinhamento local). Portanto, a similaridade é igual a 6/14 = 0,43. Ou seja, as *strings* são 43% similares.

2.3.2.3 Jaro

O comparador de *strings* proposto por Jaro (1989) é baseado no número e ordem de caracteres comuns entre duas *strings*. Considerando as strings $s = a_1 ... a_K$ e $t = b_1 ... b_L$, um caractere a_i em *s* é "comum com *t*" caso exista um caractere $b_j = a_i$ em *t* tal que

[7] http://www.dcs.shef.ac.uk/~sam/stringmetrics.html#smith, acessado em mar. 2007.

$i - H \leq j \leq i + H$, onde $H = \frac{\min(|s|,|t|)}{2}$. Defina $s' = a'_1 ... a'_{K'}$ como sendo os caracteres em *s* que são comuns com *t* (na mesma ordem em que aparecem em *s*) e defina $t' = b'_1 ... b'_{L'}$ de forma análoga. Defina uma transposição para s', t' como sendo uma posição *i* tal que $a'_i \neq b'_i$. Defina $T_{s',t'}$ como sendo a metade do número de transposições para s', t'. A métrica de similaridade *Jaro* para *s* e *t* é:

$$Jaro(s,t) = \frac{1}{3} \cdot \left(\frac{|s'|}{|s|} + \frac{|t'|}{|t|} + \frac{|s'| - T_{s',t'}}{|s'|} \right) \quad (5)$$

Segundo Bilenko et al. (2003), a função *Jaro* foi desenvolvida principalmente para *strings* pequenas (por exemplo, nomes ou sobrenomes de pessoas).

2.3.2.4 Cosine Similarity

A função *Cosine Similarity* (SALTON e MCGILL, 1983) representa os objetos (no caso, *strings*) como vetores de termos em um espaço euclidiano *n*-dimensional, onde *n* corresponde ao número de termos existentes. As funções de similaridade baseadas em termos são utilizadas quando a ordem dos termos nas *strings* não é importante para determinação da similaridade. A similaridade é baseada apenas na proximidade dos vetores, conforme mostra a figura 2.5:

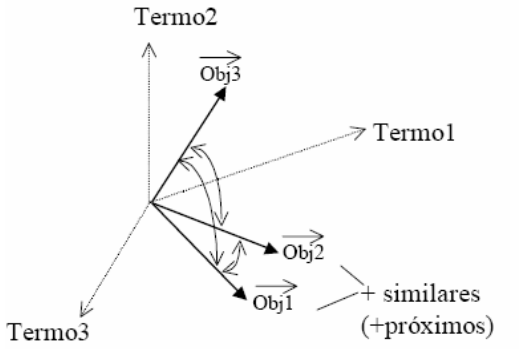

Figura 2.5: Distância entre vetores de termos (WIVES, 1999).

A similaridade pode ser quantificada pelo co-seno do ângulo entre dois vetores, conforme mostra a equação 6:

$$sim(Q,D) = \frac{\sum_{k=1}^{n} w_{qk} \times w_{dk}}{\sqrt{\sum_{k=1}^{n} (w_{qk})^2 \times \sum_{k=1}^{n} (w_{dk})^2}} \quad (6)$$

onde:

- Q é o vetor de termos de uma *string s*;
- D é o vetor de termos de uma *string t*;
- w_{qk} são os pesos dos termos do vetor Q;
- w_{dk} são os pesos dos termos do vetor D.

Existem diversos mecanismos para atribuição de pesos aos termos dos vetores. A abordagem mais simples seria adotar pesos binários (o termo está presente no vetor ou não). Outras propostas utilizam a freqüência relativa dos termos como valor dos pesos, considerando todas as *strings* envolvidas em um processo de análise de similaridade (COHEN, RAVIKUMAR e FIENBERG, 2003). O trabalho de Salton e Buckley (1988) apresenta as principais abordagens utilizadas para atribuição de pesos aos elementos dos vetores de termos.

2.3.3 Avaliação da performance das funções de similaridade

2.3.3.1 Conceitos

O objetivo de um sistema IR (*information retrieval*) é recuperar informação que pode ser útil ou relevante para o usuário. Neste tipo de sistema, os documentos (ou itens) recuperados não são necessariamente respostas exatas à consulta do usuário. Ou seja, eles precisam ser ordenados de acordo com um critério de relevância, por exemplo, através da análise do grau da similaridade entre o documento e a consulta. Logo, sistemas de recuperação de informação requerem uma avaliação da precisão do conjunto de resultados retornado, conhecida como *avaliação da performance de recuperação* (BAEZA-YATES e RIBEIRO-NETO, 1999).

A avaliação da performance usualmente envolve a definição de uma coleção de teste de referência, que consiste de: (i) uma coleção de documentos; (ii) um conjunto de exemplos de requisição de informação (consultas); e (iii) um conjunto de documentos relevantes (fornecido por especialistas) para cada exemplo de requisição de informação (BAEZA-YATES e RIBEIRO-NETO, 1999).

Além da coleção de referência, também são utilizadas métricas que permitem quantificar a performance do sistema de recuperação de informação. As métricas clássicas são conhecidas como revocação (*recall*) e precisão (*precision*). Diversos trabalhos utilizam estas medidas para avaliar a performance das funções de similaridade (DORNELES, 2006; BERRETA, 2006; BILENKO et al. 2003; COHEN, RAVIKUMAR e FIENBERG, 2003). Antes de explicitar as definições de revocação e precisão, faz-se necessário esclarecer alguns conceitos:

- Documentos relevantes (*Rel*): são aqueles que deveriam ser retornados pelo sistema de recuperação em resposta à consulta do usuário;
- Documentos recuperados (*Rec*): são todos documentos retornados pelo sistema de recuperação, independente se são relevantes ou não;
- Documentos relevantes recuperados (*RR*): representados na figura 2.6 pela intersecção, são aqueles recuperados que são relevantes.

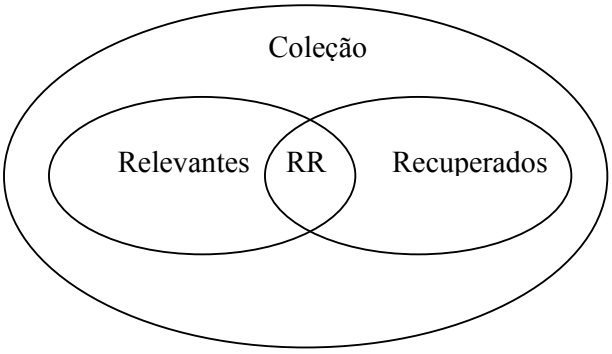

Figura 2.6: Revocação e precisão (BAEZA-YATES e RIBEIRO-NETO, 1999).

Com base na figura 2.6, é possível definir as métricas revocação e precisão:

1. Revocação é a fração de documentos relevantes que foi recuperada.

$$Revocação = \frac{|RR|}{|Rel|} \tag{7}$$

2. Precisão é a fração dos documentos recuperados que são relevantes. Mede a capacidade do sistema de recuperação em deixar fora do resultado da consulta os documentos que são irrelevantes.

$$Precisão = \frac{|RR|}{|Rec|} \tag{8}$$

No caso das funções de similaridade, a melhor função para determinada aplicação será aquela que maximiza a quantidade de itens relevantes retornados (alta revocação), e ao mesmo tempo minimize a quantidade de itens irrelevantes (alta precisão).

Um dos principais artefatos utilizados para avaliar a performance dos sistemas de recuperação de informação é a curva Revocação vs. Precisão. A construção da curva baseia-se na técnica que calcula a *precisão nos onze níveis padrões de revocação* (0%, 10%, 20%, ..., 100%) (BAEZA-YATES e RIBEIRO-NETO, 1999).

2.3.3.2 Análise de trabalhos que avaliam a performance das funções de similaridade

Partindo dos mecanismos de avaliação de sistemas de recuperação de informação definidos na seção anterior, Berreta (2006) estudou funções de similaridade entre *strings* com o objetivo de integrar duas bases de dados distintas: uma contendo a produção científica de pesquisadores e a outra contendo uma classificação de eventos e veículos de publicação. A referida integração consiste do mesmo problema enfrentado nesta dissertação, conhecido aqui como problema da Similaridade Lattes-Qualis, detalhado na Seção 4.3.3 do Capítulo 4. Portanto, as melhores funções identificadas na pesquisa de Berreta foram adotadas no presente trabalho.

Berreta avaliou a performance de 18 funções de similaridade. As melhores funções, considerando o domínio da aplicação, foram: *Smith-Waterman* e *Cosine Similarity*. O comportamento destas funções também foi estudado detalhadamente no trabalho de Berreta, através da análise das curvas Revocação vs. Precisão. Os dados indicaram que a função *Smith-Waterman* apresentou o melhor resultado final na integração. Este resultado motivou a utilização da função na presente dissertação.

Cohen, Ravikumar e Fienberg (2003) fizeram uma comparação entre diversas funções de distância entre *strings* em tarefas de casamento e clusterização de nomes de entidades, que resultou na implementação de um *toolkit* bastante conhecido chamado SecondString[8]. As funções foram investigadas em relação à performance através da análise das curvas Revocação vs. Precisão, considerando diferentes bases de dados.

Um aspecto fundamental sobre a avaliação das funções de similaridade é que a performance de uma função pode diferir consideravelmente dependendo do problema investigado. Uma função de similaridade que apresenta boa performance em uma determinada base de dados pode fracassar completamente em outra. Portanto, pesquisas como a de Cohen, Ravikumar e Fienberg, que estudam a performance *média* das funções de similaridade considerando diversas bases de dados, fornecem indicativos sobre a eficiência das funções. Porém, a escolha da melhor função depende do domínio da aplicação onde ela será utilizada.

Neste sentido, Bilenko et al. (2003), trabalhando na questão de detectar registros duplicados em banco de dados, empregaram técnicas de inteligência artificial que se adaptam a domínios específicos. A abordagem consiste em combinar múltiplas funções de similaridade entre *strings* que capturam diferentes "noções" de similaridade. Para isto, as referidas técnicas necessitam de conhecimento prévio sobre o problema a ser investigado. No caso, Bilenko et al. consideraram cada registro como um conjunto de campos distintos, ao invés de transformar o registro inteiro em uma única *string*. Assim, a similaridade entre registros foi definida como a similaridade média entre os campos dos registros (sendo que para cada campo foi possível aplicar uma função de similaridade diferente), o que aumentou a performance do sistema em determinadas bases de dados.

Ainda, Bilenko et al. passaram a representar a similaridade entre registros com um vetor de características, ao invés de uma única medida de similaridade. Construindo um conjunto de treinamento onde alguns pares de registros são previamente identificados como similares ou não-similares, Bilenko et al. adotaram métodos de aprendizagem que adaptativamente encontram a melhor combinação de funções de similaridade para determinado problema, bem como a importância (peso) de cada função. Também, Bilenko et al. utilizaram uma abordagem baseada em cadeias de Markov para desenvolver uma função de distância de edição adaptativa, onde o custo de cada operação foi adaptado em relação ao domínio da aplicação considerado. Novamente, os resultados demonstraram aumento de performance.

2.3.4 Definição do *threshold* e avaliação da qualidade das funções de similaridade

Silva et al. (2007) propuseram um método para medir a qualidade das funções de similaridade, isto é, expressar a eficiência das funções em diferenciar resultados relevantes e irrelevantes para determinado conjunto de dados.

Partindo do problema de que a distribuição dos resultados (*scores*) de uma função de similaridade depende do conjunto de dados, e pode ser completamente diferente de uma distribuição obtida por outra função, o método também define uma técnica para estimação semi-automática dos melhores valores de limiares (*thresholds*). De acordo com Silva et al., o objetivo da técnica é fornecer um intervalo de valores de *thresholds*

[8] http://secondstring.sourceforge.net, acessado em mar. 2007.

$\left[t_{best}^{\min}, t_{best}^{\max}\right]$ que maximiza o número de casos em $s_{irrel} \leq t_{best} \leq s_{rel}$, onde s_{rel} é o menor *score* de um item relevante e s_{irrel} é o maior *score* de um item irrelevante.

Segundo Silva et al., a técnica é baseada em um processo de amostragem, onde uma amostra de dados Q é extraída a partir de um conjunto de dados V. Para cada elemento $q \in Q$ é calculada a similaridade com cada elemento $v \in V$, através de uma função de similaridade $L : (Q \subseteq V) \times V \rightarrow \Re^{+}$. Assim, para cada $q \in Q$ define-se o conjunto $R_q = \{s \in \Re^{+} / s = L(q,v), \forall v \in V\}$. O conjunto R_q é ordenado de forma decrescente, criando um *ranking*, e um *expert* humano marca cada elemento i do *ranking* como relevante (*rel*), se os elementos q e v_i representam o mesmo objeto do mundo real, ou irrelevante (*irrel*) caso contrário. A tabela 2.1 mostra um exemplo de *ranking* gerado por uma função de similaridade, onde *q="Journal of Informetrics"* e V contém nove elementos, sendo que cinco são relevantes (correspondem ao mesmo objeto *q*):

Tabela 2.1: Exemplo de *ranking*.

Score	*Item de Dado*	*Relevância*
1,0000	Journal of Informetrics	Relevante
0,8636	Jrnl of Infometrics	Relevante
0,7391	J. of Informetrics	Relevante
0,1304	Informetrics Journal	Relevante
0,1304	JOI	Relevante
0,1250	Decision Support Systems	Irrelevante
0,0869	TODS	Irrelevante
0,0869	SIGMOD	Irrelevante
0,0434	TKDE	Irrelevante

Fonte: SILVA et AL., 2007.

Definindo $v_q(rel) = \min\{v / v \text{ é relevante}\}$ e $v_q(irrel) = \max\{v / v \text{ é irrelevante}\}$, se $n = |Q|$ e k é o índice de q tal que $k \in [1, n]$, é possível identificar dois pontos importantes no *ranking* (SILVA et al., 2007):

- $s_{rel}^{L}(k) = L(q, v_q(rel))$, que é o menor *score* correspondente a um item relevante;
- $s_{irrel}^{L}(k) = L(q, v_q(irrel))$, que é o maior *score* correspondente a um item irrelevante.

Na tabela 2.1, s_{rel} = 0,1304 (menor *score* de um item relevante) e s_{irrel} = 0,1250 (maior *score* de um item irrelevante).

Com base nestes dois pontos, Silva et al. definiram uma função (do tipo *reward*) para medir quão bom um valor de *threshold* é em separar itens relevantes de irrelevantes:

$$f^L(n,t) = \sum_{k=1}^{n} d(s^L_{rel}(k), s^L_{irrel}(k)) \tag{9}$$

onde:

- L é a função de similaridade usada;
- n é o número de consultas (tamanho da amostra);
- t é o *threshold* sendo analisado;
- $d(.,.)$ mede quanto $s^L_{rel}(k)$ e $s^L_{irrel}(k)$ estão adequados com o *threshold* t, tal que:

$$d(s^L_{rel}(k), s^L_{irrel}(k)) = R^t_{rel}(k) + R^t_{irrel}(k) \tag{10}$$

com:

$$R^t_{rel}(k) = \begin{cases} 1 & \text{se } s^L_{rel}(k) > t \\ -1 & \text{senão } s^L_{rel}(k) \le t \end{cases} \text{ e } R^t_{irrel}(k) = \begin{cases} -1 & \text{se } s^L_{irrel}(k) \ge t \\ 1 & \text{senão } s^L_{irrel}(k) < t \end{cases} \tag{11}$$

Ainda segundo Silva et al., o intervalo de *thresholds* ótimo, que atinge o máximo valor na função $f^L(n,t)$, pode ser definido como:

$$f^L_{\max} = \max_{t \in [t_{\min}, t_{\max}]} \left\{ f^L(n,t) \right\} \tag{12}$$

onde $t_{\min}$ e $t_{\max}$ representam os limites do intervalo de *thresholds* a ser testado. A necessidade de empregar um intervalo de valores está no fato de que vários valores de *threshold* podem atingir $f_{\max}$.

Já sobre a qualidade das funções de similaridade, Silva et al. consideram que a mesma está relacionada com a habilidade da função em separar itens relevantes e irrelevantes. Diante disso, a abordagem de Silva et al. define dois aspectos essenciais para análise da qualidade de uma função de similaridade:

1. Medir quão bem uma função de similaridade separa itens relevantes de irrelevantes;
2. Medir quão distantes no *ranking* a função de similaridade coloca itens relevantes e irrelevantes.

Conforme Silva et al., o primeiro aspecto está relacionado ao valor de $f_{\max}$. Uma função de similaridade que possui $f_{\max}$ maior comparado a outra função pode ser considerada melhor. Já o segundo aspecto está relacionado ao tamanho do intervalo de *thresholds* ótimo. Considerando que uma boa função de similaridade deve colocar itens relevantes e irrelevantes bem distantes no *ranking*, quanto maior o intervalo, melhor. Assim sendo, Silva et al. definiram uma equação chamada *discernability* que permite aferir a qualidade de uma função de similaridade. Os coeficientes $c1$ e $c2$ permitem que o usuário expresse a importância dada a cada um dos aspectos considerados. O valor de retorno da equação estará no intervalo $[-1,1]$:

$$discernability^L(t^{\min}_{best}, t^{\max}_{best}, f_{\max}) = \frac{c1}{c1+c2}(t^{\max}_{best} - t^{\min}_{best}) + \frac{c2}{c1+c2} \cdot \frac{f_{\max}}{2n} \tag{13}$$

As técnicas de Silva et al. aqui descritas foram implementada nesta dissertação. Detalhes dos experimentos realizados são encontrados no Capítulo 5.

2.4 Extração de Dados na Web

A Internet atualmente é um dos maiores repositórios de informações da humanidade, e continua crescendo de forma vertiginosa. Todavia, grande parte das informações ainda estão disponibilizadas de forma não-estruturada, no formato HTML[9]. Este formato foi desenvolvido com foco na apresentação das informações para humanos, e não para extração automática dos dados por máquinas. Mesmo com o surgimento de novos padrões para representar os dados de forma estruturada, como o XML[10], e diversos estudos sobre anotação automática de páginas no escopo da Web Semântica (BERNERS-LEE, HENDLER e LASSILA, 2001; HANDSCHUH, STAAB e CIRAVEGNA, 2002; ARLOTTA et al., 2003), poucas páginas na Web são desenvolvidas com o intuito de facilitar a extração das informações para posterior processamento por aplicações computacionais.

Uma alternativa bastante explorada para enfrentar este problema promove a construção de programas especializados, chamados *wrappers*, que identificam dados de interesse em páginas HTML e os mapeiam para um formato estruturado (usualmente XML ou tabelas relacionais de bancos de dados). Segundo Laender et al. (2002), o maior desafio destes programas é reconhecer dados de interesse entre muitos trechos de informações irrelevantes (*tags* de marcações de *layout*, código JavaScript[11] embutido, dicas de navegação, entre outros), e também identificar estruturas complexas e muitas vezes implícitas da hierarquia dos dados.

A Seção 2.4.1 descreve as principais abordagens e ferramentas encontradas na literatura para construção de *wrappers*. Não foi objetivo desta dissertação fazer uma análise detalhada de cada proposta, mas sim identificar uma solução adequada ao problema enfrentado nesta pesquisa (descrito na Seção 4.3.2 do Capítulo 4).

2.4.1 Descrição das principais abordagens e ferramentas

2.4.1.1 TSIMMIS

A abordagem pioneira proposta no projeto TSIMMIS (HAMMER et al., 1997) permite extrair dados semi-estruturados de um conjunto de páginas HTML e, em seguida, convertê-los para objetos em banco de dados. A entrada do *wrapper* aqui desenvolvido é um arquivo contendo uma especificação declarativa que indica onde os dados de interesse estão localizados nas páginas HTML e como os dados devem ser empacotados em objetos. Esta especificação consiste de uma seqüência de comandos de extração no formato *[variables, source, pattern]*, onde *variables* representa variáveis que irão armazenar o resultado da execução de cada comando, *source* especifica o texto de entrada e *pattern* indica como encontrar o dado de interesse dentro do texto de entrada. Os dados armazenados nas variáveis podem ser usados como texto de entrada para comandos subseqüentes. Ao final da execução dos comandos, o conjunto de

[9] http://www.w3.org/MarkUp/, acessado em mar. 2007.

[10] http://www.w3.org/XML/, acessado em mar. 2007.

[11] http://java.sun.com/javascript/, acessado em mar. 2007.

variáveis armazena os dados extraídos, que são mapeados para objetos em banco de dados.

2.4.1.2 W4F

A ferramenta W4F (SAHUGUET e AZAVANT, 1999) permite gerar *wrappers* contendo três camadas independentes: (i) a camada de recuperação é responsável por buscar o conteúdo HTML a partir de uma fonte Web; (ii) a camada de extração extrai a informação desejada; e (iii) a camada de mapeamento especifica como os dados são exportados. A figura 2.7 apresenta o fluxo de processamento dos *wrappers* gerados pela ferramenta W4F:

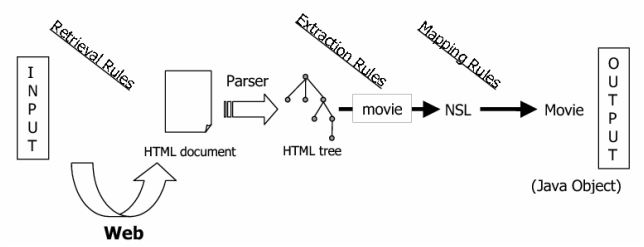

Figura 2.7: Fluxo de execução do *wrapper* (SAHUGUET e AZAVANT, 1999).

A figura 2.7 mostra que em primeiro lugar o documento HTML é obtido através de regras de recuperação. Após, ele passa por um processo de *parsing* que gera uma estrutura em forma de árvore (baseada no modelo DOM[12]) representando a hierarquia dos elementos do documento. Em seguida, as regras de extração são aplicadas nessa árvore e as informações extraídas são armazenadas em um formato interno chamado NSL (*Nested String List*). Finalmente, regras de mapeamento encapsulam as informações em um objeto Java[13] que pode ser manipulado por outras aplicações. Tanto as regras de recuperação, como as regras de extração e de mapeamento são explicitamente informadas pelo usuário em um arquivo de configuração declarativo. A ferramenta ainda fornece uma linguagem chamada HEL (*HTML Extraction Language*) que define as regras de extração.

2.4.1.3 XWRAP

A ferramenta XWRAP (LIU, PU e HAN, 2000) possibilita a construção semi-automática de *wrappers*. A principal contribuição da ferramenta foi a criação de uma interface interativa que facilita a elaboração das regras de extração das informações através do uso de heurísticas. Assim como na ferramenta W4F, o XWRAP também gera uma árvore sintática de *tokens* representando o documento HTML recuperado, conforme mostra a figura 2.8:

---

[12] http://www.w3.org/DOM/, acessado em mar. 2007.

[13] http://java.sun.com, acessado em mar. 2007.

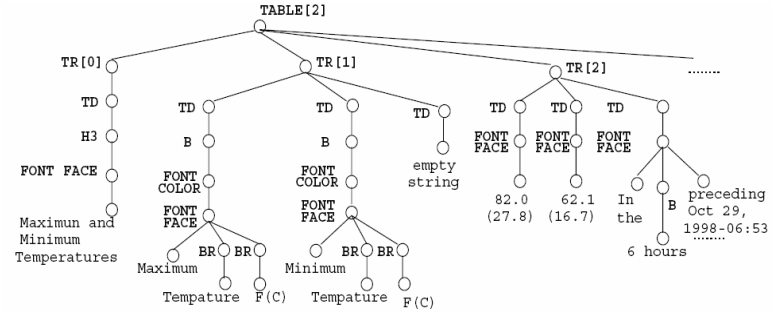

Figura 2.8: Exemplo de árvore sintática gerada no XWRAP (LIU, PU e HAN, 2000).

Após a geração da árvore, o componente de extração de informação do XWRAP realiza três tarefas principais: (i) identifica regiões de interesse em uma página Web, com o auxílio de uma interface que permite ao desenvolvedor do *wrapper* selecionar regiões como tabelas, parágrafos e listas; (ii) identifica *tokens* semânticos, através de um programa que permite ao desenvolvedor "caminhar" pela árvore e marcar os *tokens* de interesse; e (iii) determina a hierarquia do conteúdo das regiões de interesse, a partir do uso de heurísticas sobre marcações de *layout* HTML que indicam seções (por exemplo, <H3>, </H3>), tabelas (<TR>, <TD>), parágrafos (<P>), entre outras. Todas regras de extração são armazenadas em um arquivo de especificação declarativo no formato XML. Com base neste arquivo, ao final do processo a ferramenta gera o código do programa *wrapper*.

2.4.1.4 STALKER

A abordagem de Muslea, Minton e Knoblock (2001) teve como objetivo aplicar técnicas de aprendizado de máquina para indução de *wrappers* a partir de um conjunto de treinamento fornecido manualmente pelo usuário.

O algoritmo STALKER é baseado na extração de informações hierárquicas através de uma descrição da estrutura das páginas Web chamada ECT (*Embedded Catalog Tree*). A ECT possui o formato de uma árvore, sendo que as folhas representam os dados de interesse do usuário. O algoritmo modela as páginas *Web* como uma seqüência de *tokens* (palavras, números, marcações HTML, etc.). Esta seqüência é atribuída ao conteúdo do nodo raiz da estrutura ECT, sendo que cada um de seus nodos filhos recebe uma parte específica da seqüência. Com base em um conjunto de treinamento que identifica os limites das informações a serem extraídas, o algoritmo gera regras de extração que permitem percorrer a árvore ECT e localizar os nodos com os dados desejados.

2.4.1.5 DEByE

A proposta de Ribeiro-Neto, Laender e Silva (1999) solicita que os usuários informem exemplos dos objetos que desejam extrair das páginas Web. Assim, o sistema DEByE consegue produzir uma estratégia genérica para extração de dados em páginas com estrutura semelhante a dos exemplos coletados. A figura 2.9 exibe os componentes do sistema DEByE:

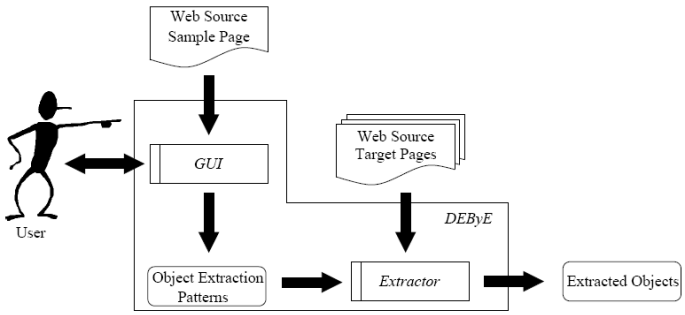

Figura 2.9: Sistema DEByE (RIBEIRO-NETO, LAENDER e SILVA, 1999).

Conforme mostra a figura 2.9, o sistema DEByE é composto por dois módulos principais. O módulo *GUI* (*Graphical User Interface*) permite que o usuário selecione os dados desejados em uma página Web de exemplo e os organize numa nova estrutura, de acordo com a sua percepção da hierarquia destes dados. Com base nesta estrutura, o sistema gera padrões genéricos para extração de novos objetos (*Object Extraction Patterns*). O módulo *Extractor* recebe estes padrões como entrada e os aplica nas demais páginas, identificando e extraindo os objetos que casem com a estrutura hierárquica definida pelos padrões.

2.4.1.6 RoadRunner

Outra importante ferramenta foi a RoadRunner (CRESCENZI, MECCA e MERIALDO, 2001). Sua proposta visa automatizar completamente a geração de *wrappers* pela comparação entre duas ou mais páginas HTML da mesma classe (por exemplo, de um *site* específico). A abordagem permite identificar as similaridades e diferenças entre as páginas e assim inferir um esquema para os dados contidos nas mesmas. Aqui, o problema de geração do *wrapper* consiste em construir uma expressão regular comum para as páginas comparadas. Como vantagens da proposta destacam-se: (i) o processo de geração do *wrapper* não necessita de qualquer interação com o usuário; e (ii) não é necessário qualquer conhecimento *a priori* sobre o conteúdo das páginas, isto é, sobre a hierarquia e organização dos dados nos documentos HTML.

2.4.1.7 ANDES

A proposta de Myllymaki (2002) utiliza tecnologias padrões para manipulação de documentos no formato XML na criação de um *framework* de extração chamado ANDES. Aqui, o processo de extração em primeiro lugar converte as páginas HTML recuperadas para o formato XHTML[14]. Esta etapa de conversão é fundamental, pois produz um documento que atende às especificações XML (em especial, garante a sintaxe bem-formada) e pode ser analisado por *parsers*. Em seguida, a ferramenta aplica diversas transformações XSLT[15] para produzir um documento XML de saída, conforme mostra a figura 2.10:

[14] http://www.w3.org/MarkUp/, acessado em mar. 2007.

[15] http://www.w3.org/TR/xslt, acessado em mar. 2007.

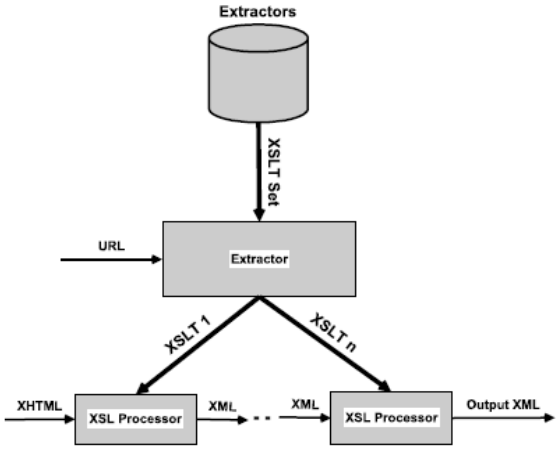

Figura 2.10: Processo de extração da ferramenta ANDES (MYLLYMAKI, 2002).

A figura 2.10 indica que um conjunto de extratores XSLT é selecionado de acordo com a URL de um documento XHTML. Este documento é processado em forma de *pipeline* pelos extratores XSLT, sendo que o resultado de um processamento é utilizado como entrada para outro. Ao final do processo é gerado um arquivo XML contendo as informações desejadas de forma estruturada.

2.4.1.8 Web-Harvest

A ferramenta Web-Harvest[16] fornece uma API que permite consultar servidores *Web*, obter a página HTML de resposta, transformá-la para XHTML e aplicar tecnologias de manipulação de documentos XML tais como XSLT, XPath[17] e XQuery[18] para extrair as informações desejadas, de forma semelhante ao sistema ANDES da seção anterior.

Todo procedimento de extração no Web-Harvest é definido pelo programador através de arquivos de configuração no formato XML. Cada arquivo de configuração descreve uma seqüência de processadores executando determinadas tarefas para atingir um objetivo final. Estes processadores rodam em forma de *pipeline*, ou seja, a saída de um processador é utilizada como entrada para outro. A figura 2.11 esquematiza o processo:

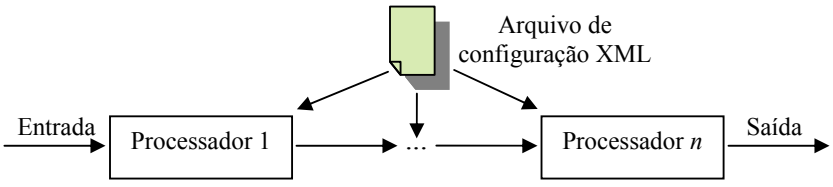

Figura 2.11: Processo de extração do Web-Harvest.

O arquivo de configuração apresentado na figura 2.12 exemplifica o processo:

---

[16] http://web-harvest.sourceforge.net, acessado em mar. 2007.

[17] http://www.w3.org/TR/xpath, acessado em mar. 2007.

[18] http://www.w3.org/XML/Query/, acessado em mar. 2007.

```
1 <?xml version="1.0" encoding="UTF-8"?>
  <config> 2
3 <file action="write" path="scholar.xml">
    <html-to-xml> 4
  5 <http url="http://scholar.google.com/scholar?q=valdeni&amp;hl=en&amp;lr=&amp;btnG=Search"/>
    </html-to-xml>
  </file>
</config>
```

Figura 2.12: Exemplo de arquivo de configuração do Web-Harvest.

O arquivo de configuração da figura 2.12 está marcado com números que servirão para elucidar as informações a seguir. O número 1 identifica que o arquivo é um documento XML. O elemento raiz do documento é `<config>` (número 2). O arquivo indica que o processo de extração utiliza três processadores, definidos pelos elementos: `<file>` (número 3), `<html-to-xml>` (número 4) e `<http>` (número 5). O fluxo de processamento ocorre da seguinte forma:

1. O processador `<http>` faz o *download* do conteúdo especificado pelo atributo `url`;
2. O processador `<html-to-xml>` realiza um processo de "limpeza" no código HTML retornado no passo anterior e produz código XHTML bem-formado;
3. O processador `<file>` grava o código XHTML gerado no passo anterior para o arquivo especificado no atributo `path`.

A figura 2.13 mostra uma parte do conteúdo do arquivo `scholar.xml` gerado ao final do processamento do arquivo de configuração de exemplo da figura 2.12:

```
<?xml version="1.0" standalone="yes"?>

<html>
  <head>
    <meta http-equiv="content-type"
      content="text/html; charset=UTF-8">
    </meta>
    <title>valdeni - Google Scholar</title>
  </head>
  <body bgcolor="#ffffff" onload="document.gs.reset()" topmargin="2"
    marginheight="2">
    <table border="0" cellpadding="0" cellspacing="0">
      <tr>
        <td valign="top">
          <a href="/scholar?hl=en&amp;lr=" target="_top">
            <img src="/intl/en/images/scholar_results.gif"
              width="150" height="58" alt="Scholar Home" border="0"
              vspace="12">
            </img>
          </a>
        </td>

        ...

</html>
```

Figura 2.13: Resultado do processamento do arquivo de configuração de exemplo.

A ferramenta Web-Harvest foi integrada no sistema desenvolvido nesta dissertação. As justificativas da escolha e os detalhes de implementação estão na Seção 4.3.2.

3 DESCRIÇÃO DO MODELO

Este capítulo descreve o modelo de pontuação na busca de competências acadêmicas de pesquisadores. Aqui, competência deve ser entendida como: "*S. f. 2.* Qualidade de quem é capaz de apreciar e resolver certo assunto, fazer determinada coisa; capacidade, habilidade, aptidão, idoneidade." (FERREIRA, 1999); "*S. f. 4.* Soma de conhecimentos ou habilidade. *5.* Indivíduo de grande autoridade num ramo do saber ou do fazer." (HOUAISS e VILLAR, 2001).

Medir a competência de pessoas é um processo extremamente complexo, porque envolve muitos aspectos subjetivos. A idéia do modelo é oferecer um *indicativo* da aptidão ou autoridade de um pesquisador através da análise de aspectos quantitativos da sua produção acadêmica. Utilizar apenas esta abordagem como critério de *determinação* da competência de um cientista é insuficiente. Portanto, o modelo não dispensa a adoção conjunta de outras formas de avaliação (como a revisão pelos pares) para oferecer uma visão mais adequada dos perfis dos pesquisadores.

O modelo é baseado na combinação de indicadores quantitativos que permitem mensurar a produtividade acadêmica dos cientistas. Um diferencial do modelo é a integração de indicadores quantitativos relacionados com a importância da produção bibliográfica dos pesquisadores. Estes indicadores podem oferecer uma percepção *quasi-qualitativa* da produção do cientista, considerando aspectos como repercussão na comunidade acadêmica e nível dos veículos de publicação. O modelo também proporciona calcular um coeficiente de competência (*CC*) que determina a pontuação obtida pelo pesquisador.

3.1 Indicadores Quantitativos

O modelo divide os indicadores em duas categorias principais:

1. Indicadores quantitativos relacionados ao currículo do pesquisador – são conhecidos como indicadores de produção, e quantificam o volume da produção do pesquisador em diversas dimensões: publicações (artigos em periódicos, trabalhos em anais de eventos, livros, etc.), produção técnica (softwares, relatórios e pareceres técnicos, etc.), orientações concluídas, participações em bancas e eventos, entre outras;
2. Indicadores quantitativos relacionados à produção bibliográfica do pesquisador – mensuram aspectos como o impacto ou repercussão dos trabalhos do pesquisador na comunidade acadêmica (através do número de citações), bem como a qualidade e alcance dos veículos de publicação nos quais o pesquisador possui trabalhos publicados.

3.1.1 Indicadores quantitativos relacionados ao currículo

Estes indicadores são extraídos a partir dos indicadores de produção do pesquisador, seguindo a abordagem proposta por Cazella (2006). Os seguintes indicadores foram considerados:

Tabela 3.1: Indicadores quantitativos relacionados ao currículo

Categoria: Produção Bibliográfica	
Indicador	*Sigla*
Artigos publicados em periódicos	APP
Livros publicados ou organizados	LPO
Capítulos de livros publicados	CLP
Trabalhos publicados em anais de eventos	TPAE
Textos em jornais ou revistas	TJR
Outros tipos de produção bibliográfica	OPB
Categoria: Produção Técnica	
Indicador	*Sigla*
Software	SW
Produtos tecnológicos	PTEC
Trabalhos técnicos	TT
Outros tipos de produção técnica (organização de eventos)	OPTOE
Outros tipos de produção técnica (relatórios de pesquisa)	OPTRP
Outros tipos de produção técnica (apresentação de trabalhos)	OPTAP
Categoria: Orientações Concluídas	
Indicador	*Sigla*
Orientações concluídas de doutorado	OCD
Orientações concluídas de mestrado	OCM
Outras orientações concluídas	OOC
Categoria: Informações Complementares	
Indicador	*Sigla*
Participações em bancas de trabalhos de conclusão de doutorado	PBTCD
Participações em bancas de trabalhos de conclusão de mestrado	PBTCM
Outras participações em bancas de trabalhos de conclusão	OPBTC
Participações em bancas de comissões julgadoras	PBCJ
Participações em eventos	PE

Orientações em andamento de doutorado	OAD
Orientações em andamento de mestrado	OAM
Outras orientações em andamento	OOA

Fonte: Adaptado de CAZELLA, 2006. p. 106-107

3.1.2 Indicadores quantitativos relacionados à produção bibliográfica

O objetivo destes indicadores é oferecer um mecanismo que permita diferenciar a produção bibliográfica entre os pesquisadores. A idéia básica é entendida através do seguinte exemplo: imagine uma situação onde dois pesquisadores possuem volume parecido de publicações, porém um deles apresenta quantidade maior de trabalhos em veículos de publicação com repercussão internacional, enquanto que o outro concentra suas publicações em veículos de menor alcance. Se o modelo utilizar apenas os indicadores quantitativos da Seção 3.1.1, torna-se inviável a identificação e diferenciação destas situações.

Os indicadores aqui considerados fornecem uma percepção *quasi-qualitativa* da produção acadêmica de um cientista. Eles possibilitam uma investigação da classificação dos veículos de publicação onde o pesquisador possui trabalhos publicados e também uma estimativa do impacto ou repercussão destes trabalhos na comunidade científica.

3.1.2.1 Classificação dos veículos de publicação

A importância de um artigo é diretamente proporcional a sua qualificação em função do nível do veículo de publicação. Ou seja, o modelo deve capturar a noção de que trabalhos publicados em periódicos ou eventos reconhecidos e com repercussão internacional possuem uma maior importância no cálculo da pontuação do pesquisador.

Assim sendo, o modelo possui 18 indicadores correspondentes às possíveis classificações dos veículos de publicação consideradas. Cada indicador representa o número de trabalhos do pesquisador que foram classificados na categoria específica, conforme mostra a tabela 3.2:

Tabela 3.2: Indicadores quantitativos da classificação dos veículos de publicação

Periódicos		**Eventos**	
Indicador	*Sigla*	*Indicador*	*Sigla*
Qualidade A e Repercussão Internacional	PQARI	Qualidade A e Repercussão Internacional	EQARI
Qualidade A e Repercussão Nacional	PQARN	Qualidade A e Repercussão Nacional	EQARN
Qualidade A e Repercussão Local	PQARL	Qualidade A e Repercussão Local	EQARL
Qualidade B e Repercussão Internacional	PQBRI	Qualidade B e Repercussão Internacional	EQBRI
Qualidade B e Repercussão Nacional	PQBRN	Qualidade B e Repercussão Nacional	EQBRN

Qualidade B e Repercussão Local	PQBRL	Qualidade B e Repercussão Local	EQBRL
Qualidade C e Repercussão Internacional	PQCRI	Qualidade C e Repercussão Internacional	EQCRI
Qualidade C e Repercussão Nacional	PQCRN	Qualidade C e Repercussão Nacional	EQCRN
Qualidade C e Repercussão Local	PQCRL	Qualidade C e Repercussão Local	EQCRL

3.1.2.2 Repercussão na comunidade acadêmica

A importância de um artigo também pode ser avaliada em termos de sua repercussão na comunidade acadêmica. Um trabalho citado por vários outros trabalhos provavelmente possui impacto e relevância superiores a um trabalho que não possui citações. Para representar esta característica, segundo Hirsch (2005), os seguintes critérios são usualmente adotados:

1. Número total de citações – Vantagens: mede o impacto total. Desvantagens: difícil de encontrar; pode estar inflacionado por citações de trabalhos em co-autoria ou auto-citações; não distingue os trabalhos com contribuições originais de pesquisa;
2. Citações por artigo, isto é, a razão entre o número total de citações pelo número total de artigos – Vantagens: permite a comparação de cientistas com idades científicas diferentes. Desvantagens: difícil de encontrar; recompensa baixa produtividade, penaliza alta produtividade;
3. Número de "artigos significantes", definido como o número de artigos com mais do que y citações. Vantagens: elimina as desvantagens dos critérios 1 e 2. Desvantagens: y é arbitrário;
4. Número de citações de cada um dos q artigos mais citados. Vantagens: elimina as desvantagens dos critérios 1 e 2. Desvantagens: não é um único número, dificultando sua obtenção e comparação; q é arbitrário.

Hirsch também propõe uma nova métrica chamada *h-index*, definida como o número de artigos com número de citações maior ou igual que h. Ela pode ser entendida intuitivamente da seguinte forma: após obter uma lista contendo os trabalhos e o número de citações de cada trabalho do pesquisador, crie um *ranking* destes trabalhos ordenando a lista pelo número de citações. Assim, na primeira posição do *ranking* estará o trabalho mais citado, e na última o menos citado. Percorra esta lista de cima para baixo até que o *ranking* do trabalho seja maior que o número de citações que ele possui. A posição anterior no *ranking* corresponde ao valor de h. Conforme Hirsch, o *h-index* mede o impacto geral dos trabalhos de um pesquisador, evitando todas as desvantagens dos critérios anteriores.

Nesta dissertação, os indicadores da tabela 3.3 foram considerados:

Tabela 3.3: Indicadores quantitativos da repercussão na comunidade acadêmica

Indicador	*Sigla*
Número total de citações	TOTCIT
Razão entre o número total de citações e o número total de trabalhos	RCIT
h-index	*h-index*

3.2 Cálculo do Coeficiente de Competência (*CC*)

O coeficiente de competência (*CC*) visa determinar a pontuação final do pesquisador, levando em consideração os indicadores quantitativos. O *CC* encontra-se na faixa de valores entre 0 a 10 pontos, onde 0 indica que o modelo não conseguiu identificar nenhuma competência do pesquisador e o valor 10 indica que o cientista possui "competência máxima" entre os pesquisadores avaliados.

A fim de proporcionar maior flexibilidade na realização dos experimentos e assim investigar em maiores detalhes a influência dos indicadores quantitativos no cálculo do *CC*, o modelo permite calcular um coeficiente de competência considerando apenas os indicadores quantitativos relacionados ao currículo (CC_c), e outro coeficiente analisando apenas os indicadores quantitativos relacionados à produção bibliográfica (CC_b). Posteriormente ambos os coeficientes podem ser combinados num único valor final, gerando o *CC*.

3.2.1 Normalização

Para manter a consistência da escala desejada (valores no intervalo [0,10]), todos os indicadores (aqui representados pelo símbolo *a*) são normalizados antes dos cálculos dos coeficientes de competência de acordo com a equação 14:

$$an = MinMax(a) = \frac{a - minA}{maxA - minA} \times (newmaxA - newminA) + newminA \qquad (14)$$

Segundo Cazella (2006), os indicadores apresentam diversas escalas de valores, e a equação 14 realiza uma transformação linear nos dados originais. A normalização *MinMax* procura mapear um valor *a* de *A* para um valor *an* na escala [*newminA*, *newmaxA*], sendo que *minA* e *maxA* são os valores mínimos e máximos de um indicador na escala de origem.

3.2.2 Coeficientes de competência

Os coeficientes de competência CC_c, CC_b e *CC* são calculados através da equação 15, que representa uma média aritmética ponderada dos indicadores quantitativos normalizados:

$$\frac{\sum_{i=1}^{n} (p_i * an_i)}{\sum_{i=1}^{n} p_i} \qquad (15)$$

Respectivamente, no cálculo do CC_c apenas os indicadores quantitativos relacionados ao currículo (Seção 3.1.1) são considerados. Já para o cálculo do CC_b,

apenas os indicadores quantitativos relacionados à produção bibliográfica (Seção 3.1.2) são considerados. Finalmente, para o cálculo final do *CC* é feita uma média aritmética ponderada dos resultados de CC_c e CC_b.

Uma questão que esta abordagem coloca é a determinação dos pesos de cada indicador quantitativo no cálculo das médias aritméticas ponderadas. Cada pessoa, unidade ou instituição de pesquisa, órgão de ciência e tecnologia, entre outros, tem uma percepção diferente da importância dos indicadores na avaliação da performance ou produtividade de um cientista. Diante disso, o modelo aqui proposto não pretende de forma alguma determinar quais seriam os pesos ótimos ou recomendáveis para cada indicador, deixando esta tarefa a critério dos responsáveis pela instanciação do modelo.

O próximo capítulo apresenta os detalhes da implementação do modelo em um sistema Web. O Capítulo 5 relata os experimentos efetuados para avaliar o modelo.

4 IMPLEMENTAÇÃO DO MODELO

O modelo descrito no Capítulo 3 foi implementado em um sistema visando a realização de experimentos para avaliação das proposições desta dissertação. Os seguintes requisitos de alto nível nortearam a construção do sistema:

- Automatização das tarefas – as principais funcionalidades da aplicação devem ser executadas com o mínimo de intervenção humana;
- Execução em ambiente Web;
- Utilização de tecnologias padrões de mercado;
- Utilização de tecnologias *open source*[19];
- Documentação da arquitetura e do código-fonte produzido;
- Arquitetura flexível para inclusão de novas funcionalidades e integração com pesquisas futuras;
- Disponibilização do sistema para a comunidade acadêmica no modelo *open source*.

O sistema desenvolvido permite que qualquer pesquisador avalie quantitativamente sua produção científica. Ele automatiza diversos aspectos relacionados à tarefa de avaliação, como a obtenção dos indicadores de produção e a integração das diferentes bases de informações, para oferecer uma visão mais completa da produtividade do cientista.

As próximas seções detalham a arquitetura e o funcionamento do sistema, ressaltando as dificuldades encontradas na sua implementação e as soluções adotadas para superá-las.

4.1 Descrição das Fontes de Coleta de Dados

O sistema utiliza informações oriundas de outros sistemas (externos) para capturar os indicadores quantitativos. Esta seção descreve as principais características das fontes de informação consideradas.

Os indicadores de produção foram obtidos da Plataforma Lattes[20] do CNPq, que oferece o sistema de currículos mais difundido no Brasil (CV-Lattes), definido pelo CNPq como (CNPQ, 2007):

---

[19] http://www.opensource.org, acessado em mar. 2007.

> Um sistema de informação curricular desenvolvido com o objetivo de avaliar a competência de candidatos à obtenção de bolsas e auxílios; selecionar consultores e membros de comitês e de grupos assessores; e subsidiar a avaliação da pesquisa e da pós-graduação brasileiras. Este sistema é utilizado pelo MCT, FINEP, CAPES/MEC, FAPESP, USP e outras instituições e também pela comunidade científica brasileira, a qual envolve pesquisadores, estudantes, gestores e profissionais. Todos os bolsistas de pesquisa, de mestrado, de doutorado e de iniciação científica, orientadores credenciados e outros membros da comunidade ligados ao CNPq devem ter um currículo cadastrado no CV-Lattes.

Já os indicadores quantitativos relacionados à produção bibliográfica foram obtidos a partir de duas bases de informações distintas. A primeira busca representar uma classificação dos veículos de publicação. O sistema Qualis-CAPES[21] foi escolhido para modelar este aspecto com base em trabalhos relacionados (CAZELLA, 2006; BERRETA, 2006; NIEDERAUER, 2002) e devido a crescente adoção do mesmo por diversos programas de pós-graduação no Brasil. Segundo a CAPES, o Qualis é (CAPES, 2007):

> ... o resultado do processo de classificação dos veículos utilizados pelos programas de pós-graduação para a divulgação da produção intelectual de seus docentes e alunos. A classificação é feita ou coordenada pelo representante de cada área e passa por processo anual de atualização. Os veículos de divulgação citados pelos programas de pós-graduação são enquadrados em categorias indicativas da qualidade - A, B ou C e do âmbito de circulação dos mesmos - local, nacional ou internacional. As combinações dessas categorias compõem nove alternativas indicativas da importância do veículo utilizado, e, por inferência, do próprio trabalho divulgado [...].

A segunda base visa capturar o impacto e repercussão dos trabalhos do pesquisador na comunidade científica. Neste caso, o serviço Google Scholar[22] foi selecionado como fonte de informação por ser um sistema público e transparente que recolhe na Web as publicações e computa automaticamente suas citações. Através dele, é possível identificar, para cada artigo indexado, o número de citações, quais são os trabalhos que fazem parte do conjunto de citações, quem são os autores, quais são os veículos de publicação, entre outras informações relevantes às funcionalidades do sistema aqui implementado.

No contexto deste trabalho é utilizado o conceito "referência" para designar as informações extraídas de um artigo indexado pelo Google Scholar.

4.2 Arquitetura e Funcionalidades

A figura 4.1 apresenta uma visão macro da arquitetura e das principais funcionalidades oferecidas pelo sistema:

---

[20] http://lattes.cnpq.br, acessado em mar. 2007.

[21] http://servicos.capes.gov.br/webqualis, acessado em mar. 2007.

[22] http://scholar.google.com, acessado em mar. 2007.

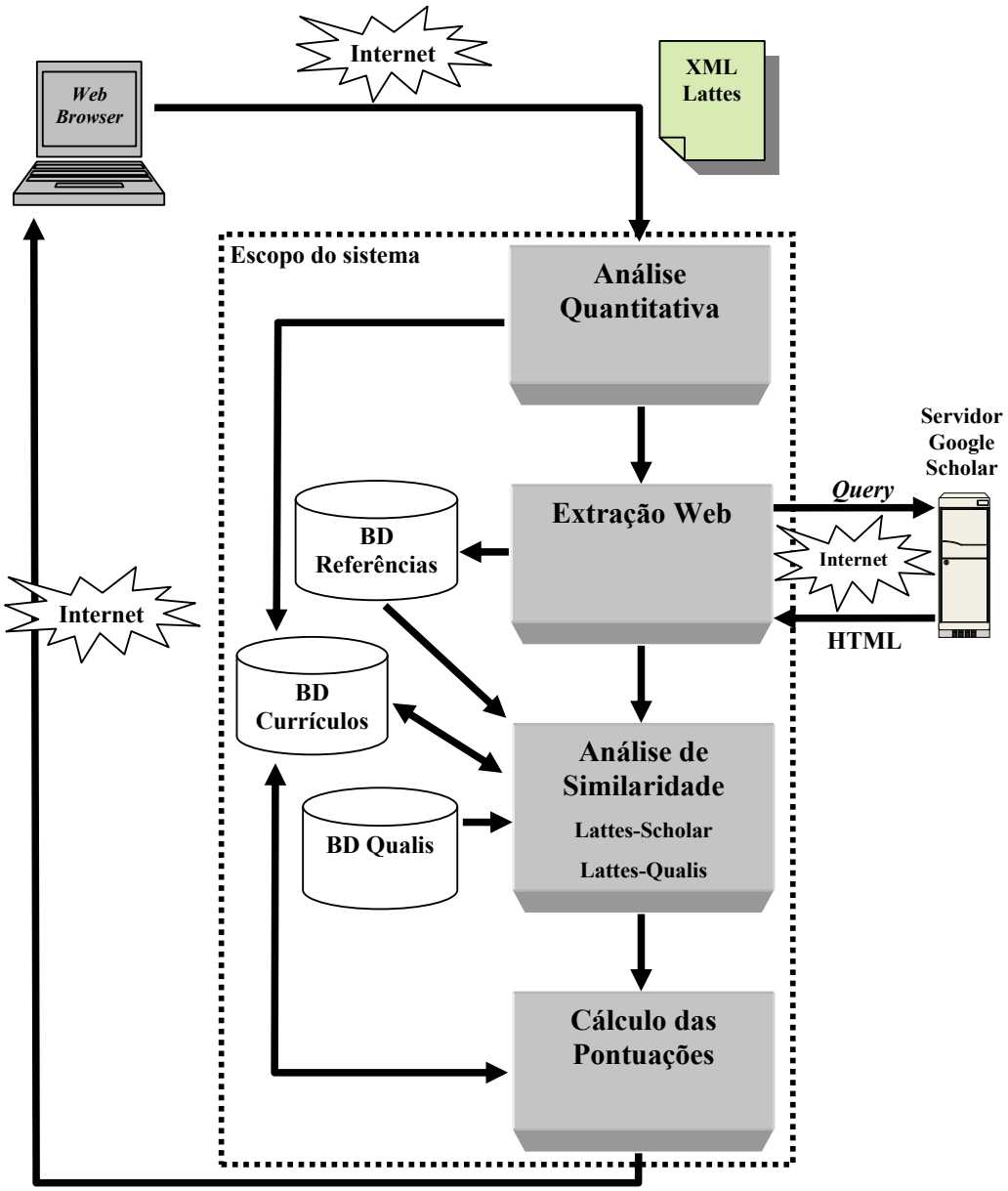

Figura 4.1: Arquitetura e funcionalidades do sistema.

Conforme mostra a figura 4.1, a entrada do sistema consiste apenas do arquivo XML do CV-Lattes do pesquisador. Não há necessidade de informar qualquer dado adicional. Optou-se por utilizar a versão em XML do currículo porque ela pode ser facilmente gerada pelo pesquisador na Plataforma Lattes do CNPq.

Imediatamente após o envio do currículo através de um navegador Web, o sistema executa o módulo chamado *Análise Quantitativa*. Este módulo extrai os indicadores quantitativos relacionados ao currículo descritos na Seção 3.1.1, bem como diversas informações necessárias aos demais processamentos da aplicação. O módulo *Extração Web* efetua consultas ao servidor Google Scholar para coleta das referências contendo o número de citações dos trabalhos do pesquisador. Já o módulo *Análise de Similaridade* possui duas funcionalidades principais: (i) a similaridade Lattes-Scholar tem por

objetivo garantir que o número de citações de um trabalho retornado pelo Google Scholar apenas será considerado quando este trabalho for efetivamente encontrado no CV-Lattes do pesquisador; e (ii) a similaridade Lattes-Qualis tem por objetivo verificar o nível dos veículos de publicação nos quais o pesquisador possui trabalhos publicados. Por sua vez, o módulo chamado *Cálculo das Pontuações* implementa a normalização dos indicadores e o cálculo dos coeficientes de competência descritos na Seção 3.2.

A figura 4.1 ainda distingue três bases de dados locais. A base *BD Currículos* armazena todos os indicadores quantitativos utilizados pelo sistema, bem como informações sobre as publicações cadastradas no CV-Lattes do pesquisador. A base *BD Referências* contém todas as referências extraídas do Google Scholar. Já a base *BD Qualis* representa uma cópia local da base de dados do sistema Qualis-CAPES.

A próxima seção apresenta em maiores detalhes as funcionalidades do sistema, destacando aspectos de implementação de cada módulo da aplicação.

4.3 Detalhamento das Funcionalidades e Aspectos de Implementação

4.3.1 Análise Quantitativa

Este módulo realiza a manipulação do arquivo XML do CV-Lattes do pesquisador para obtenção dos indicadores quantitativos relacionados ao currículo, descritos na Seção 3.1.1. Além disso, extrai diversas informações da produção bibliográfica que são necessárias aos demais módulos da aplicação.

A manipulação do arquivo XML foi efetuada através de expressões de caminho XPath, elaboradas com base na DTD do arquivo XML do CV-Lattes fornecida pela comunidade CONSCIENTIAS-LMPL[23].

Em primeiro lugar, foram identificados na DTD quais elementos correspondiam aos indicadores quantitativos desejados. Em conjunto, verificou-se a hierarquia destes elementos na estrutura XML. Um exemplo está na tabela 4.1:

Tabela 4.1: Análise da DTD do CV-Lattes

Indicador	*Elemento correspondente na DTD*	*Hierarquia do Elemento na DTD*
APP	ARTIGO-PUBLICADO	CURRICULO-VITAE | PRODUCAO-BIBLIOGRAFICA | ARTIGOS-PUBLICADOS | ARTIGO-PUBLICADO

Com base nesta hierarquia, elaborou-se uma expressão XPath para contar o número de ocorrências do elemento no documento XML, e assim extrair o indicador quantitativo correspondente, como mostra o exemplo da tabela 4.2:

[23] http://lmpl.cnpq.br/lmpl/, acessado em mar. 2007.

Tabela 4.2: Expressão XPath utilizada para obtenção do indicador quantitativo APP

Indicador	*Expressão XPath*
APP	`count(//PRODUCAO-BIBLIOGRAFICA/ARTIGOS-PUBLICADOS/ARTIGO-PUBLICADO)`

Além disso, o módulo de análise quantitativa é responsável pela obtenção de informações da produção bibliográfica do pesquisador necessárias aos demais módulos da aplicação. Estes dados são armazenados localmente no banco de dados do sistema. É importante ressaltar que todas as informações obtidas do CV-Lattes são de responsabilidade do pesquisador, e assume-se que as mesmas são verdadeiras. A tabela 4.3 detalha as informações extraídas:

Tabela 4.3: Informações extraídas da produção bibliográfica do CV-Lattes

Dados sobre Artigos Publicados em Periódicos	
Atributo	*Descrição*
Natureza	Representa a natureza do artigo, se é um artigo completo ou resumido
Título	Representa o título do artigo
Ano	Representa o ano de publicação do artigo
País	Representa o país onde o artigo foi publicado
Título do Periódico	Representa o título do periódico onde o artigo foi publicado
ISSN do Periódico	Representa o ISSN do periódico onde o artigo foi publicado
Dados sobre Trabalhos Publicados em Anais de Eventos	
Atributo	*Descrição*
Natureza	Representa a natureza do trabalho, se é um trabalho completo, resumido ou resumo-expandido
Título	Representa o título do trabalho
Ano	Representa o ano de publicação do trabalho
País	Representa o país onde o trabalho foi publicado
Nome do Evento	Representa o nome do evento onde o trabalho foi publicado
Classificação do Evento	Representa a classificação do evento nas categorias internacional, nacional, regional ou local
Título dos Anais ou *Proceedings*	Representa o título dos anais ou *proceedings* onde o trabalho foi publicado
ISBN dos Anais ou *Proceedings*	Representa o ISBN dos anais ou *proceedings*
Dados sobre Livros Publicados Ou Organizados	
Atributo	*Descrição*

Tipo	Representa o tipo do livro, se é livro publicado ou organizado
Título	Representa o título do livro
Ano	Representa o ano de publicação do livro
País	Representa o país onde o livro foi publicado
ISBN do Livro	Representa o ISBN do livro
Dados sobre Capítulos de Livros Publicados	
Atributo	*Descrição*
Título	Representa o título do capítulo de livro
Ano	Representa o ano de publicação do capítulo de livro
País	Representa o país onde o capítulo de livro foi publicado
Título do Livro	Representa o título do livro onde o capítulo foi publicado
ISBN do Livro	Representa o ISBN do livro onde o capítulo foi publicado

4.3.2 Extração Web

O objetivo deste módulo é extrair as referências contendo o número de citações dos trabalhos do pesquisador. Com esta informação, é possível calcular os indicadores quantitativos relacionados à repercussão acadêmica descritos na Seção 3.1.2.2.

O serviço Google Scholar foi escolhido como fonte de coleta das citações. Todavia, o Google Scholar não fornecia uma API para acesso aos seus dados de forma estruturada até a época de implementação deste módulo. Portanto, procedeu-se um estudo sobre técnicas de extração de dados na Web (ver Seção 2.4). Optou-se por integrar a ferramenta Web-Harvest no sistema, pois a mesma se adequou aos principais requisitos estabelecidos:

- Implementação na tecnologia Java, a linguagem de programação escolhida para a construção do sistema;
- Integração com o ambiente de execução Web, isto é, a ferramenta é capaz de rodar no mesmo servidor Web da aplicação;
- Suporte a tecnologias padrões para manipulação de documentos XML, como XSLT, XPath e XQuery. Com a rápida evolução e padronização destas tecnologias, elas se tornaram ferramentais valiosos para extração de informações;
- Performance em tempo real, ou seja, o módulo precisa ser executado imediatamente após o envio das informações pelo pesquisador, e os resultados devem ser apresentados na seqüência;
- Facilidade de configuração, tanto da ferramenta em si como das descrições dos processos de extração Web;
- Possibilidade de extensão e inclusão de novas funcionalidades, com código-fonte aberto (modelo *open source*);

- Principalmente, adaptar-se ao problema em questão, ou seja, extrair com sucesso as citações coletadas pelo serviço Google Scholar.

Com relação a este último requisito, faz-se necessário explorar algumas questões. Desde sua criação, em novembro de 2004, a interface do Google Scholar sofreu pequenas alterações, porém manteve a mesma estrutura lógica de apresentação das informações, conforme mostra a figura 4.2:

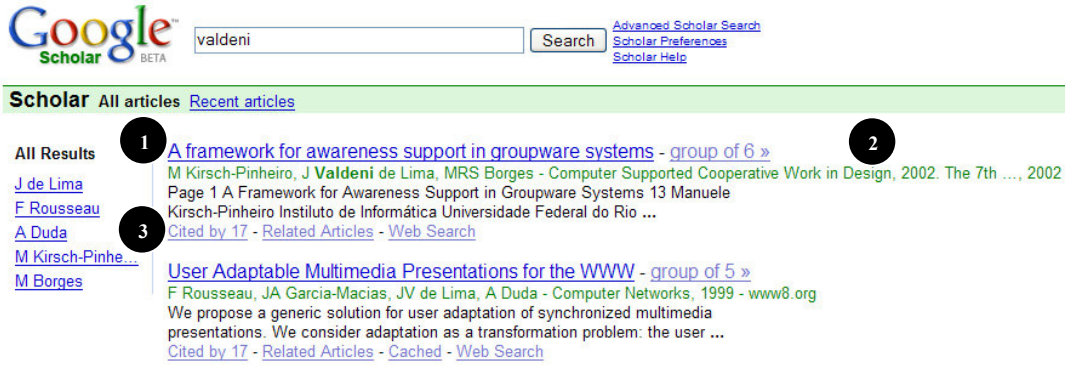

Figura 4.2: Interface do serviço Google Scholar.

Observa-se na figura 4.2 que a interface possui uma estrutura simples e bem organizada. Os artigos são apresentados em uma lista, sendo que ao lado direito do indicado pelo número 1 encontra-se o título do artigo. Logo abaixo do indicado pelo número 2 estão as informações sobre os autores, o veículo e o ano da publicação. Finalmente, a direita do indicado pelo número 3 encontra-se o número de citações que o artigo recebeu até o momento. Esta padronização facilita o mecanismo de extração adotado por ferramentas como o Web-Harvest, que consiste em "descobrir" a lógica de composição da página HTML e assim descrever o processo inverso, ou seja, como separar os dados desejados das marcações de formatação. Para uma descrição do funcionamento do Web-Harvest, consulte a Seção 2.4.1.8.

4.3.2.1 Definição do processo de extração para o Google Scholar

Em primeiro lugar, converteu-se uma página HTML de resposta do Google Scholar para o formato XHTML. Em seguida, identificou-se a estrutura lógica correspondente às informações desejadas (título do artigo, autores, veículo e ano de publicação, número de citações), conforme apresenta a figura 4.3:

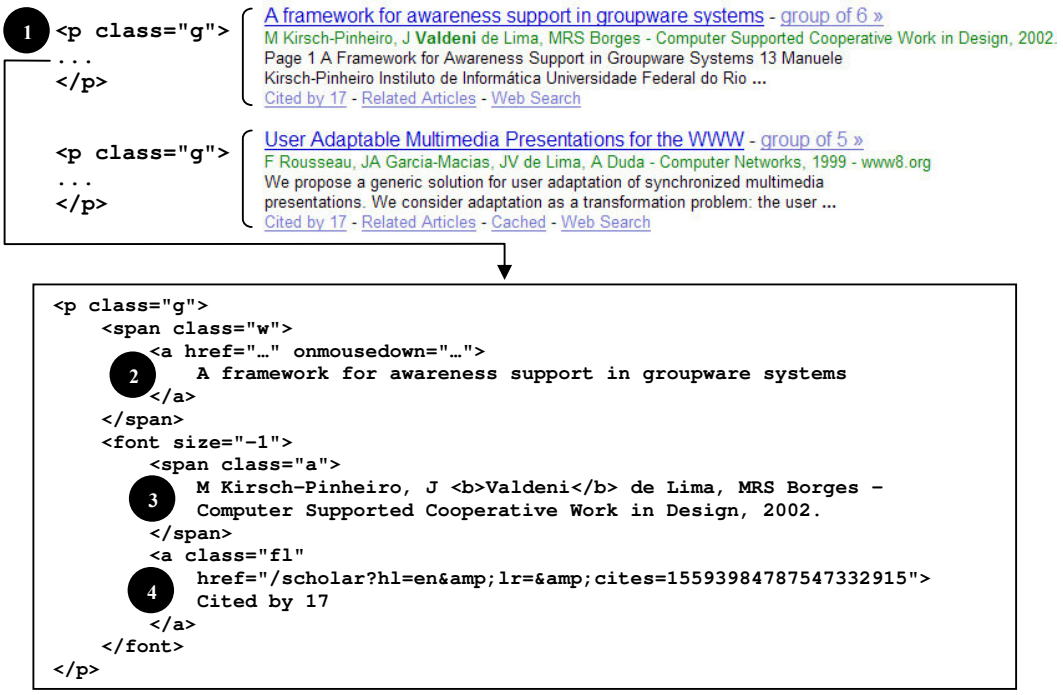

Figura 4.3: Estrutura lógica das informações desejadas.

A figura 4.3 mostra que cada artigo no documento XHTML está encapsulado pelo elemento `<p class="g">` (indicado pelo número 1 na figura). Este elemento possui outros elementos aninhados que por sua vez encapsulam as informações desejadas. O título do artigo, indicado na figura pelo número 2, é o texto do elemento `<a>` encapsulado pelo elemento `<span class="w">`. Já as informações sobre autores, veículo e ano de publicação estão no texto do elemento `<span class="a">` encapsulado pelo elemento `<font size="-1">`, conforme indica o número 3. Ainda, o número de citações pode ser extraído do texto do elemento `<a class="fl">` encapsulado pelo elemento `<font size="-1">`, de acordo com o indicado pelo número 4. Ou mais genericamente, o número de citações está no texto do elemento `<a>` cujo atributo `href` contém a string `"cites="`. A tabela 4.4 descreve possíveis expressões de caminho XPath que permitem localizar as informações desejadas, considerando como base o elemento `<p class="g">`:

Tabela 4.4: Expressões XPath para localizar as informações desejadas.

Infomação	*Expressão XPath*
Título	`//span[@class='w']/a`
Autores, veículo e ano de publicação	`//font[@size='-1']/span[@class='a']`
Número de citações	`//font[@size='-1']/a[contains(@href, 'cites=')]`

Partindo desta análise preliminar, foi descrito um processo de extração Web para o serviço Google Scholar da seguinte forma: para cada página HTML retornada como

resposta a uma consulta no serviço Google Scholar, aplique a transformação para o formato XHTML e extraia todos os elementos `<p class="g">` (bem como os elementos aninhados a este), armazenando-os em uma lista de elementos. Para cada elemento desta lista, aplique expressões XPath para localizar e extrair o conteúdo desejado. Ao final, gere um arquivo XML contendo as informações de forma estruturada.

4.3.2.2 Implementação do processo de extração para o Google Scholar

O processo de extração descrito na seção anterior foi implementado com a ferramenta Web-Harvest. O arquivo de configuração completo encontra-se no Anexo B desta dissertação. As figuras de 4.4 a 4.8 detalham aspectos específicos de cada etapa do processo descrito no arquivo de configuração.

```
<?xml version="1.0" encoding="UTF-8"?>

<config>

  <include path="functions.xml"/>  (1)

  <var-def name="url">  (3)
    <template>http://scholar.google.com/scholar?q=${qScholar}&amp;hl=en&amp;lr=&amp;btnG=Search</template>  (2)
  </var-def>
```

Figura 4.4: Arquivo de configuração do Web-Harvest (etapa inicial).

A figura 4.4 mostra que o arquivo de configuração utiliza outro arquivo (`functions.xml`), fornecido pela própria ferramenta, contendo um conjunto de funções pré-definidas (indicado pelo número 1 na figura). Estas funções encapsulam tarefas menores de extração úteis a maioria dos processos. A listagem deste arquivo está no Anexo B. A figura 4.4 também apresenta a definição da URL de consulta inicial ao servidor Google Scholar. O Web-Harvest foi configurado para receber a *string* de consulta via parâmetro. A expressão `${qScholar}` (indicada pelo número 2) é substituída em tempo de execução pelo valor fornecido ao sistema (informado pelo usuário através de um formulário Web). A URL é então armazenada internamente pelo Web-Harvest na variável `url` (indicada pelo número 3).

```
(6) <var-def name="referencias">
      <call name="download-multipage-list">  (1)
      (2) <call-param name="pageUrl"><var name="url"/></call-param>
      (3) <call-param name="nextXPath">//td[.='Next']/a/@href</call-param>
      (4) <call-param name="itemXPath">//p[@class="g"]</call-param>
      (5) <call-param name="maxloops"><var name="qtdePaginas"/></call-param>
      </call>
  </var-def>
```

Figura 4.5: Arquivo de configuração do Web-Harvest (etapa intermediária).

A figura 4.5 apresenta a etapa intermediária do processo de extração. Ela é realizada por uma função disponível no Web-Harvest chamada `download-multipage-list` (indicada pelo número 1 na figura). Ela percorre todas as páginas HTML de resposta à consulta inicial, transforma-as para documentos XHTML e gera uma lista contendo todos os elementos extraídos de cada documento. A função recebe quatro parâmetros de entrada. O primeiro parâmetro corresponde a URL inicial de consulta (indicado pelo

número 2), que no caso é a variável `url` gerada na etapa anterior. O segundo parâmetro (indicado pelo número 3) corresponde a uma expressão XPath que permite localizar nos documentos XHTML o elemento que contém a próxima URL a ser processada (veja a figura 4.6). O terceiro parâmetro (indicado pelo número 4) corresponde a uma expressão XPath que especifica quais elementos (com seus respectivos elementos aninhados) devem ser extraídos de cada documento XHTML e adicionados na lista de resultados final. O quarto parâmetro (indicado pelo número 5) especifica o número de páginas que devem ser processadas. O valor deste parâmetro é informado pelo sistema em tempo de execução (digitado pelo usuário em um formulário Web). Ao final do processamento da função a lista de resultados é armazenada internamente pelo Web-Harvest na variável `referencias` (indicada pelo número 6).

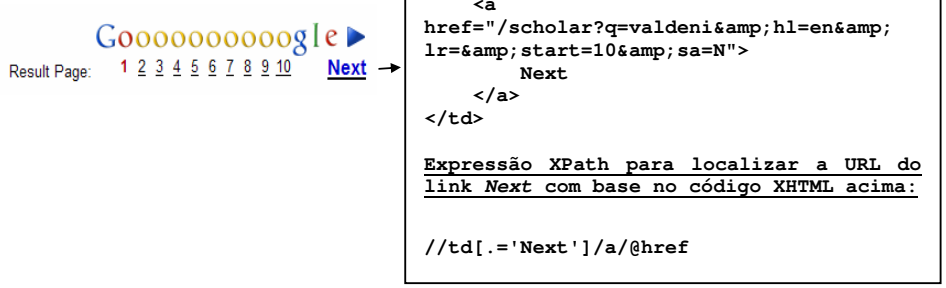

Figura 4.6: Definição da expressão XPath para localizar a próxima URL de consulta.

A etapa final do processo de extração para o serviço Google Scholar é mostrada na figura 4.7:

```
<file action="write" path="referencias_scholar.xml" charset="UTF-8">
  <![CDATA[ <?xml version="1.0" encoding="UTF-8"?>
            <scholar>
  ]]>
  <loop item="item" index="i">
    <list><var name="referencias"/></list>
    <body>
      <xquery>
        <xq-param name="item"><var name="item"/></xq-param>
        <xq-expression><![CDATA[

          let $titulo := data($item//span[@class='w']/a)
          let $autores_veiculo_ano := data($item//font[@size='-1']/span[@class='a'])
          let $citacoes := data($item//font[@size='-1']/a[contains(@href, 'cites=')])

          return
            <referencia>
              <titulo>{normalize-space($titulo)}</titulo>
              <autores_veiculo_ano>{normalize-space($autores_veiculo_ano)}</autores_veiculo_ano>
              <citacoes>{normalize-space($citacoes)}</citacoes>
            </referencia>

        ]]></xq-expression>
      </xquery>
    </body>
  </loop>
  <![CDATA[ </scholar>
  ]]>
</file>

</config>
```

Figura 4.7: Arquivo de configuração do Web-Harvest (etapa final).

A etapa final mostrada na figura 4.7 consiste em percorrer a lista de resultados `referencias` obtida na etapa anterior (utilizando o processador `<loop>` indicado pelo número 1) e, para cada elemento da lista, aplicar a tecnologia XQuery em conjunto

com expressões XPath para localizar e extrair as informações desejadas (conforme indicado pelo número 2). As expressões de caminho XPath foram construídas com base na tabela 4.4. As informações são estruturadas em um novo elemento `<referencia>`, que contém os elementos `<titulo>`, `<autores_veiculo_ano>` e `<citacoes>`, conforme indica o número 3. Após percorrer toda a lista de resultados, um novo arquivo XML é salvo com todas as referências extraídas (indicado pelo número 4) e o processo de extração está terminado. A figura 4.8 apresenta um trecho do arquivo XML de saída do processo de extração:

```
<?xml version="1.0" encoding="UTF-8"?>
<scholar>
  <referencia>
    <titulo>
      A framework for awareness support in groupware systems
    </titulo>
    <autores_veiculo_ano>
      M Kirsch-Pinheiro, J Valdeni de Lima, MRS Borges - Computer
      Supported Cooperative Work in Design, 2002. The 7th …, 2002
      - ieeexplore.ieee.org
    </autores_veiculo_ano>
    <citacoes>Cited by 17</citacoes>
  </referencia>
  <referencia>
    <titulo>
      User Adaptable Multimedia Presentations for the WWW
    </titulo>
    <autores_veiculo_ano>
      F Rousseau, JA Garcia-Macias, JV de Lima, A Duda - Computer
      Networks, 1999 - www8.org
    </autores_veiculo_ano>
    <citacoes>Cited by 17</citacoes>
  </referencia>
  ...
</scholar>
```

Figura 4.8: Arquivo XML gerado ao final do processo de extração.

O documento XML da figura 4.8 passa por um processo de *parsing* e suas informações são armazenadas no banco de dados local do sistema, conforme mostra a tabela 4.5:

Tabela 4.5: Informações das referências extraídas do Google Scholar

Dados sobre Referências	
Atributo	*Descrição*
Título	Representa o título do artigo extraído
Autores_Veículo_Ano	Representa os autores, veículo e ano de publicação do artigo extraído. Armazenados como uma única *string*
Número de Citações	Representa o número de citações do artigo extraído

4.3.3 Análise de Similaridade

O objetivo deste módulo é aplicar funções de similaridade entre *strings* necessárias para o processo de extração dos indicadores quantitativos relacionados à produção bibliográfica descritos na Seção 3.1.2.

O módulo Análise de Similaridade possui duas funcionalidades principais: a similaridade Lattes-Scholar e a similaridade Lattes-Qualis.

4.3.3.1 Similaridade Lattes-Scholar

A similaridade Lattes-Scholar visa garantir que o número de citações de uma referência extraída do Google Scholar apenas será considerado no cálculo dos indicadores quantitativos da Seção 3.1.2.2 quando o título desta referência for efetivamente encontrado entre os títulos dos trabalhos cadastrados no CV-Lattes do pesquisador. Em outras palavras, o objetivo deste processamento é evitar que citações de trabalhos que não são da autoria do pesquisador sejam indevidamente atribuídas ao mesmo.

O processamento da similaridade Lattes-Scholar é essencial já que as consultas efetuadas ao serviço Google Scholar podem gerar respostas com ruído. Por ruído entende-se que no conjunto de respostas estão artigos que não são da autoria do pesquisador consultado. Sem um processamento adicional, este ruído propaga-se aos indicadores quantitativos comprometendo o resultado final do modelo.

Decidiu-se por utilizar uma abordagem baseada em funções de similaridade pois as bases de dados envolvidas no processo (CV-Lattes e Google Scholar) não possuem relacionamentos por chave primária nem restrições conjuntas de integridade. Assim sendo, é possível que os títulos dos trabalhos do pesquisador não sejam 100% similares nas duas bases de dados. Isto ocorre porque o pesquisador cadastra manualmente as informações de um trabalho de sua autoria no CV-Lattes, e erros de digitação são comuns de acontecer. Também, o processo de coleta e extração automatizada do serviço Google Scholar nem sempre gera resultados precisos. Observe os exemplos na tabela 4.6:

Tabela 4.6: Exemplos de inconsistências de títulos no serviço Google Scholar

Título Original do Trabalho	*Título do Trabalho no Google Scholar*
User Adaptable Multimedia Presentations for the WWW	José Valdeni & DUDA, Andrzej. User Adaptable Multimedia Presentations for the WWW
Functional Variability of the Human Cortical Motor Map: Electrical Stimulation Findings in Perirolandic Epilepsy Surgery	Functional Variability of the Human Cortical Motor Map: Electrical Stimulation Findings in …

A tabela 4.6 mostra duas situações de inconsistências de títulos de trabalhos coletados pelo serviço Google Scholar. Na primeira situação, o Google Scholar incluiu informações sobre os autores no título do trabalho. Já na segunda situação, o Google Scholar disponibilizou o título de forma abreviada (cortando a parte final). Nesta dissertação, não foram efetuadas tarefas de pré-processamento adicionais para identificar e tratar estas situações. As tarefas ficaram sob responsabilidade das funções de similaridade.

4.3.3.2 Similaridade Lattes-Qualis

A similaridade Lattes-Qualis visa possibilitar a extração dos 18 indicadores quantitativos relacionados com as classificações dos veículos de publicação, descritos na Seção 3.1.2.1.

Para cada periódico ou anais de evento cadastrado no CV-Lattes de um pesquisador, é preciso identificar a classificação do mesmo na base local contendo as classificações do sistema Qualis-CAPES. Assim como ocorre na similaridade Lattes-Scholar, as bases CV-Lattes e Qualis-CAPES não possuem relacionamentos por chave-primária nem restrições conjuntas de integridade. Portanto, é possível que os títulos dos periódicos e dos anais de eventos não sejam 100% similares nas duas bases. A tabela 4.7 mostra alguns exemplos:

Tabela 4.7: Exemplos de inconsistências nos títulos de periódicos e anais de eventos

Periódicos	
Título no CV-Lattes	*Título na Base Qualis-CAPES*
Revista Novas Tecnologias na Educação	RENOTE. Revista Novas Tecnologias na Educação
Anais de Eventos	
Título no CV-Lattes	*Título na Base Qualis-CAPES*
SBBD'2004 - 19 Simpósio Brasileiro de Banco de Dados	- 19 Simposio Brasileiro de Bancos de Dados - 19. Simpósio Brasileiro de Banco de Dados - 19º Simpósio Brasileiro de Banco de Dados - Anais do 19 SBBD - Anais do 19 Simpósio Brasileiro de Banco de Dados - Anais do 19o Simpósio Brasileiro de Banco de Dados - SBBD 2004 - Anais do SBBD´2004 - Anais do XIX SBBD

A tabela 4.7 indica que não existe uma padronização na entrada dos títulos dos periódicos e anais de eventos nas bases CV-Lattes e Qualis-CAPES. Na base CV-Lattes, o pesquisador possui total liberdade para digitar o título da maneira que desejar. Na base Qualis-CAPES a situação não é diferente. Inclusive, percebe-se que o mesmo evento está cadastrado repetidas vezes com diversos títulos diferentes. Portanto, para que seja possível integrar estas bases de dados e assim extrair a informação desejada (as classificações dos periódicos ou anais de eventos), optou-se por utilizar a abordagem baseada em funções de similaridade nesta dissertação.

A base local contendo as classificações dos periódicos e anais de eventos foi criada a partir das listas completas em formato Microsoft Excel[24] disponibilizadas no site do sistema Qualis-CAPES. Estas listas representavam a última classificação dos periódicos e anais de eventos para o triênio 2004-2006. As seguintes informações foram extraídas e armazenadas na base de dados local do sistema:

[24] http://office.microsoft.com/excel, acessado em mar. 2007.

Tabela 4.8: Informações extraídas do sistema Qualis-CAPES

Dados sobre Periódicos e Anais de Eventos	
Atributo	*Descrição*
Título	Representa o título do periódico ou dos anais de evento
Qualidade	Representa a qualidade do periódico ou dos anais de evento nas categorias A, B ou C
Circulação	Representa a circulação do periódico ou dos anais de evento nas categorias I (Internacional), N (Nacional) ou L (Local)

4.3.3.3 Aplicação das funções de similaridade

Os atributos selecionados para o processamento da similaridade Lattes-Scholar foram: o Título das referências extraídas do serviço Google Scholar (ver tabela 4.5) e os Títulos dos trabalhos cadastrados na categoria produção bibliográfica dos CV-Lattes dos pesquisadores (ver tabela 4.3). Os demais atributos disponíveis (por exemplo, autores e veículo de publicação) não foram utilizados no processamento.

Por sua vez, os atributos selecionados no processamento da similaridade Lattes-Qualis para periódicos foram: Título do Periódico na categoria Artigos Publicados em Periódicos do CV-Lattes (ver tabela 4.3) e Título do Periódico na base local Qualis (ver tabela 4.8). Já no processamento da similaridade Lattes-Qualis para anais de eventos, selecionaram-se os atributos Nome do Evento na categoria Trabalhos Publicados em Anais de Eventos do CV-Lattes (ver tabela 4.3) e Título dos Anais de Eventos armazenados na base local Qualis (ver tabela 4.8).

Os atributos envolvidos na análise da similaridade passaram por um processo de pré-processamento. Este processo substitui por espaços em branco todos os caracteres que não são letras ou números, conforme exibe a figura 4.9:

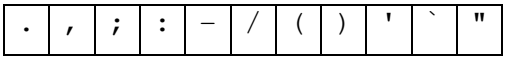

Figura 4.9: Caracteres substituídos por espaços em branco.

Ainda, o pré-processamento substitui todos os caracteres com acentuação pelos respectivos caracteres não-acentuados. Por exemplo: Á é substituído por A. Finalmente, todos os caracteres são transformados para maiúsculo. A tabela 4.9 mostra exemplos do pré-processamento dos atributos:

Tabela 4.9: Exemplos de pré-processamento dos atributos

Atributo antes do Pré-Processamento	*Atributo após Pré-Processamento*
AvalWeb-Sistema interativo para gerência de questões e aplicação de avaliações na Web	AVALWEB SISTEMA INTERATIVO PARA GERENCIA DE QUESTOES E APLICACAO DE AVALIACOES NA WEB
Document Numérique (Paris)	DOCUMENT NUMERIQUE PARIS

O processamento das funções de similaridade foi efetuado de forma exaustiva. Por exemplo, caso um pesquisador possua *n* periódicos cadastrados em seu CV-Lattes e na base local Qualis existam *m* periódicos cadastrados, no total serão *n*m* aplicações da função de similaridade.

Os atributos adotados no processamento da similaridade foram considerados similares quando o *score* retornado pela função é igual ou maior que o *threshold* avaliado.

No caso da similaridade Lattes-Scholar, quando o título de uma referência extraída do Google Scholar é similar ao título de um trabalho cadastrado no CV-Lattes do pesquisador, o número de citações da referência é considerado no cálculo dos indicadores quantitativos descritos na Seção 3.1.2.2.

Na similaridade Lattes-Qualis para periódicos foi necessário adotar um critério adicional, pois o mesmo periódico podia estar cadastrado em diferentes áreas de conhecimento na base local Qualis (sendo que em cada área o periódico podia estar com uma classificação diferente). Em primeiro lugar, é processada a similaridade considerando apenas os periódicos classificados na área do pesquisador. Caso seja encontrado um periódico similar, sua classificação é utilizada no cálculo dos indicadores quantitativos descritos na Seção 3.1.2.1. Caso contrário, é realizado um processamento adicional para todas as demais áreas de conhecimento, e a maior classificação obtida por um periódico similar, se o mesmo for encontrado, é adotada.

Finalmente, a similaridade Lattes-Qualis para anais de eventos restringiu-se apenas para eventos da área da Ciência da Computação. Quando o título de um evento cadastrado no CV-Lattes de um pesquisador é similar ao título de um evento armazenado na base local Qualis, a classificação deste evento na base é considerada no cálculo dos indicadores quantitativos descritos na Seção 3.1.2.1. Caso dois ou mais eventos cadastrados na base local Qualis atinjam o mesmo *score* de similaridade (observe na tabela 4.7 que o mesmo evento pode estar cadastrado diversas vezes), a maior classificação do evento é considerada.

O Capítulo 5 apresenta os resultados de um experimento cujo objetivo foi investigar as funções de similaridade e definir valores de *threshold* adequados aos dados processados pelo módulo de Análise de Similaridade.

4.3.4 Cálculo das Pontuações

Este módulo apenas implementa as equações da Seção 3.2 para normalização dos indicadores e cálculo dos coeficientes de competência.

4.3.5 Tecnologias

A tabela 4.10 sumariza as tecnologias adotadas na construção do sistema:

Tabela 4.10: Tecnologias adotadas na construção do sistema

Categoria	*Tecnologia*	*URL (acessada em mar. 2007)*
Linguagem de Programação	Java JDK 5	http://java.sun.com
Framework Web	Apache Struts 1.2.9 e JSTL 1.1	http://struts.apache.org http://java.sun.com/products/jsp/jstl/
Biblioteca para Processamento das Funções de Similaridade	SimMetrics 1.6.0	http://www.dcs.shef.ac.uk/~sam/simmetrics.html
Framework para Extração de Dados na Web	Web-Harvest 0.3	http://web-harvest.sourceforge.net
Manipulação de Arquivos XML	Apache Xalan-Java 2.7.0	http://xml.apache.org/xalan-j/
Manipulação de Arquivos Microsoft Excel	Jakarta POI 2.5.1	http://jakarta.apache.org/poi/
Framework para Mapeamento Objeto-Relacional	Hibernate 3.2.0	http://www.hibernate.org
Servidor de Aplicação	JBoss 4.0.5	http://www.jboss.com
Banco de Dados	PostgreSQL 8.1	http://www.postgresql.org
Ambiente de Desenvolvimento	Eclipse 3.2.1	http://www.eclipse.org

4.4 Dificuldades e Limitações

Durante o período de implementação do modelo as seguintes limitações e dificuldades foram identificadas:

- O processo de extração de dados na *Web* adotado pela ferramenta Web-Harvest geralmente é dependente da lógica de composição das páginas HTML. Ou seja, caso esta lógica mude, muito provavelmente será necessário ajustar o arquivo de configuração da ferramenta para refletir as alterações. No caso do processo de extração para o serviço Google Scholar, durante todo o desenvolvimento desta dissertação foi necessário ajustar apenas uma única vez o arquivo de configuração, devido a pequenas mudanças efetuadas na interface do serviço;

- As bases disponibilizadas pelo sistema Qualis-CAPES (principalmente a base que contém as classificações dos anais de eventos) possuíam pouca padronização (algumas vezes o mesmo evento chegava a ser classificado mais de 20 vezes com títulos diferentes). Esta falta de qualidade dificultou a implementação do módulo de Análise de Similaridade e limitou o processamento ao triênio 2004-2006;
- O serviço Google Scholar bloqueia por algumas horas o acesso aos seus dados quando vários processos de extração são realizados em seqüência. Esta política é adotada pelo Google Scholar com a intenção de evitar ataques de negação de serviço.

5 EXPERIMENTOS E RESULTADOS

Este capítulo apresenta e discute os experimentos realizados nesta dissertação, que tiveram dois objetivos principais:

1. Verificar, dentre as funções de similaridade selecionadas, quais são as mais adequadas aos problemas de similaridade enfrentados;
2. Verificar o uso do modelo de pontuação na busca de competências acadêmicas em uma aplicação real.

5.1 Definição do Conjunto de Dados

Os experimentos foram realizados a partir de um conjunto de dados contendo no total 12 CV-Lattes na versão XML de pesquisadores doutores da área da Ciência da Computação. Este conjunto foi obtido através do envio de um e-mail para listas de discussões da área na Universidade Federal do Rio Grande do Sul. No e-mail foram explicados os objetivos da dissertação e dos experimentos. Ressaltou-se que todos os participantes não seriam identificados, preservando assim a privacidade dos mesmos.

Para cada currículo do conjunto foi efetuada uma consulta ao servidor Google Scholar a fim de extrair as referências contendo o número de citações dos trabalhos do pesquisador. A consulta consistiu das iniciais do nome do pesquisador seguido pelo sobrenome, unida como a informação obtida a partir do campo `NOME-EM-CITACOES-BIBLIOGRAFICAS` do CV-Lattes. Por exemplo: `"R O RECH" OR "RECH, R. O."`. As primeiras 40 páginas de resposta foram processadas. Não foi implementado um processamento adicional para verificar auto-citações e co-autoria (este processamento é tema de trabalhos futuros da dissertação). A tabela 5.1 sumariza o volume de dados analisado:

Tabela 5.1: Volume de dados analisado

Pesquisador	*Número de Publicações no Currículo (APP + TPAE + LPO + CLP)*	*Número de Referências Extraídas*	*Número de Citações Extraídas*
P1	21	35	80
P2	15	40	41
P3	55	160	583
P4	168	230	476
P5	113	313	1742
P6	43	91	314

P7	25	285	2687
P8	98	248	873
P9	34	104	139
P10	123	160	218
P11	193	400	825
P12	35	58	1750
Total	**923**	**2124**	**9728**

No tocante aos dados oriundos da base do sistema Qualis-CAPES, totalizaram-se 26.694 periódicos (considerando todas as áreas de conhecimento definidas pela CAPES) e 824 anais de eventos (considerando apenas a área da Ciência da Computação), correspondentes à última classificação disponível[25] dos veículos de publicação para o triênio 2004-2006.

5.2 Primeiro Experimento: Análise das Funções de Similaridade

Neste experimento, as funções de similaridade *Levenshtein*, *Smith-Waterman*, *Jaro* e *Cosine Similarity* foram investigadas em relação a dois aspectos: (i) quais funções produzem melhores resultados no conjunto de dados analisado e (ii) quais valores de *threshold* devem ser utilizados.

As referidas funções foram selecionadas a partir de trabalhos correlatos que também aplicaram as funções em bases de dados similares (SILVA et al., 2007; BERRETA, 2006). Deve-se ressaltar que não foi objetivo desta dissertação fazer uma análise exaustiva de todas as funções de similaridade existentes. Trabalhos futuros irão concentrar seus esforços em identificar as soluções ótimas.

Adotou-se a abordagem proposta por Silva et al. (2007) para atingir os objetivos do experimento, descrita em detalhes na Seção 2.3.4. Para cada função de similaridade, os valores de intervalo de *thresholds* ótimos, f_{max} e *discernabilty* foram calculados de acordo com as equações 9, 12 e 13. No cálculo da *discernability*, os valores dos coeficientes foram $c1 = c2 = 1$.

5.2.1 Similaridade Lattes-Scholar

Uma amostra contendo 45 referências distintas extraídas do Google Scholar foi selecionada aleatoriamente. Processou-se então a similaridade Lattes-Scholar entre os títulos destas 45 referências com os títulos dos trabalhos cadastrados nos CV-Lattes dos pesquisadores. Para cada referência da amostra, extraiu-se manualmente o menor *score* de um trabalho relevante (similar) e o maior *score* de um trabalho irrelevante (não similar). Logo, para cada função de similaridade processada foi gerada uma tabela contendo os referidos *scores* das 45 referências, conforme mostra o Anexo C.

As tabelas do Anexo C foram utilizadas como entrada nos cálculos dos valores de intervalo de *thresholds* ótimos, f_{max} e *discernability* das respectivas funções de similaridade. A tabela 5.2 apresenta os resultados obtidos:

[25] Obtida do *site* http://servicos.capes.gov.br/webqualis em 15 jan. 2007.

Tabela 5.2: Resultados das funções na similaridade Lattes-Scholar

Função	$f_{\max}$	*discernability*	$\left[t_{best}^{\min}, t_{best}^{\max}\right]$
Smith-Waterman	90	0,5715	[0,742, 0,886]
Levenshtein	82	0,5650	[0,489, 0,709]
Cosine Similarity	84	0,5211	[0,645, 0,755]
Jaro	82	0,4600	[0,782, 0,792]

A tabela 5.2 está ordenada pelo valor de *discernability*. Este valor indica que a função *Smith-Waterman* proporcionou o melhor resultado geral para o conjunto de dados analisado. Inclusive, ela atingiu o valor máximo possível para $f_{\max}$, ou seja, a função conseguiu separar todos os itens relevantes dos irrelevantes.

Valores de $f_{\max}$ próximos ao máximo já eram esperados, devido às características dos dados envolvidos na similaridade Lattes-Scholar. Basicamente, o título de uma referência extraída do Google Scholar será pouco *ou* bastante similar em relação ao título de um trabalho cadastrado no CV-Lattes do pesquisador, o que facilita a separação dos itens relevantes e irrelevantes. Isto ocorre porque o título de uma referência é extraído automaticamente pelo Google Scholar a partir do próprio título do trabalho do pesquisador encontrado na Web. Como este processo é automatizado, a ocorrência de erros geralmente é pequena. Ainda, é pouco provável que dois pesquisadores possuam nomes similares e ao mesmo tempo trabalhos com títulos muito semelhantes. Caso isto ocorra, recomenda-se incluir outras informações (além do título) no processamento da similaridade (por exemplo, dados sobre autores e veículos de publicação). Este processamento adicional é tema dos trabalhos futuros desta dissertação.

Verificou-se também que a função *Smith-Waterman* conseguiu identificar corretamente os problemas de inconsistências de títulos do serviço Google Scholar descritos na tabela 4.6. Por utilizar o conceito de melhor alinhamento local (ver Seção 2.3.2.2), a função considera que as *strings s="USER ADAPTABLE MULTIMEDIA PRESENTATIONS FOR THE WWW"* e *t="JOSE VALDENI & DUDA ANDRZEJ USER ADAPTABLE MULTIMEDIA PRESENTATIONS FOR THE WWW"* são similares, já que *s* está inteiramente contida em *t*.

Portanto, a função *Smith-Waterman* foi selecionada como padrão para o processamento da similaridade Lattes-Scholar. O *threshold* padrão adotado foi o valor 0,814, que corresponde à média entre o menor e o maior valor do intervalo de *thresholds* ótimo da função indicado na tabela 5.2. A tabela 5.3 mostra os resultados finais do processamento da similaridade Lattes-Scholar para os 12 pesquisadores do conjunto de dados:

Tabela 5.3: Similaridade Lattes-Scholar (*Smith-Waterman*, 0,814)

Pesquisador	*Referências Extraídas*	*Citações Extraídas*	*Referências sem Citações*	*Referências sem Trabalhos similares no CV-Lattes (< 0,814)*	*Referências com Trabalhos similares no CV-Lattes (≥ 0,814)*	*Citações Desprezadas*	*Citações Consideradas*
P1	35	80	25	6	4	52	28
P2	40	41	27	6	7	13	28
P3	160	583	86	48	26	414	169
P4	230	476	131	49	50	239	237
P5	313	1742	157	127	29	1633	109
P6	91	314	38	35	18	221	93
P7	285	2687	88	189	8	2666	21
P8	248	873	143	58	47	549	324
P9	104	139	53	34	17	110	29
P10	160	218	91	25	44	68	150
P11	400	825	192	99	109	320	505
P12	58	1750	38	8	12	1698	52
					Total	**7983**	**1745**

Na tabela 5.3 percebe-se a importância da similaridade Lattes-Scholar. Para os pesquisadores P5 e P7, por exemplo, fica claro que a grande maioria das referências extraídas (e das citações correspondentes) pertencem a outros pesquisadores com nomes semelhantes. No caso do pesquisador P7, das 2.666 citações extraídas, apenas 21 de fato correspondiam a trabalhos do pesquisador. Ou seja, a consulta efetuada ao servidor Google Scholar retornou grande quantidade de ruído, que foi minimizado com o processamento da similaridade.

5.2.2 Similaridade Lattes-Qualis

5.2.2.1 Periódicos

Selecionou-se aleatoriamente uma amostra contendo 20 títulos distintos de periódicos cadastrados na categoria Artigos Publicados em Periódicos (APP) nos CV-Lattes dos pesquisadores. Processou-se então a similaridade Lattes-Qualis entre os títulos dos periódicos da amostra com os títulos dos periódicos cadastrados na base local contendo a classificação do sistema Qualis-CAPES. Para cada periódico da amostra, extraiu-se manualmente o menor *score* de um periódico relevante (similar) e o maior *score* de um periódico irrelevante (não-similar) na base local Qualis. As tabelas com os *scores* extraídos para cada função de similaridade estão no Anexo C. A tabela 5.4 exibe os resultados obtidos:

Tabela 5.4: Resultados das funções na similaridade Lattes-Qualis para periódicos

Função	$f_{\max}$	*discernability*	$\left[t_{best}^{\min}, t_{best}^{\max}\right]$
Smith-Waterman	38	0,477	[0,875, 0,880]
Levenshtein	32	0,408	[0,833, 0,851]
Cosine Similarity	32	0,407	[0,800, 0,816]
Jaro	28	0,355	[0,901, 0,913]

A tabela 5.4, ordenada pelo valor de *discernability*, mostra que as funções de similaridade obtiveram excelente performance em separar periódicos similares de não similares no conjunto de dados analisado. Isto pode ser explicado pela boa padronização dos títulos dos periódicos nas duas bases. No caso, a função *Smith-Waterman* atingiu um valor de $f_{\max}$ igual a 38 (o máximo possível seria 40). Este valor indica que para apenas 1 periódico (do total de 20 da amostra) não foi possível determinar se o periódico correspondente na base local Qualis era relevante ou não, considerando qualquer valor de *threshold* no intervalo ótimo de *thresholds* da função.

A função *Smith-Waterman* com valor de *threshold* 0,877 (média entre os valores mínimo e máximo do intervalo de *thresholds* ótimo da função na tabela 5.4) foi adotada como padrão no processamento da similaridade Lattes-Qualis para periódicos. A tabela 5.5 apresenta os resultados do processamento para todos os pesquisadores do conjunto de dados:

Tabela 5.5: Similaridade Lattes-Qualis – periódicos (*Smith-Waterman*, 0,877)

Pesquisador	*Periódicos no Indicador APP* *(Triênio 2004-2006)*	*Periódicos sem similares na Base Qualis* *(< 0,877)*	*Periódicos com similares na Base Qualis* *(≥ 0,877)*
P1	3	1	2
P2	0	0	0
P3	2	1	1
P4	3	1	2
P5	3	0	3
P6	0	0	0
P7	1	0	1
P8	1	1	0
P9	0	0	0
P10	2	1	1
P11	14	2	12
P12	2	0	2
Total	**31**	**7**	**24**

5.2.2.2 Anais de Eventos

O mesmo procedimento descrito na seção anterior foi adotado. Selecionou-se aleatoriamente uma amostra contendo 20 títulos distintos de eventos cadastrados na

categoria Trabalhos Publicados em Anais de Eventos (TPAE) nos CV-Lattes dos pesquisadores. Processou-se então a similaridade Lattes-Qualis entre os títulos dos eventos da amostra com os títulos dos eventos cadastrados na base local Qualis. Para cada evento da amostra, extraiu-se o menor *score* de um evento relevante (similar) e o maior *score* de um evento irrelevante (não-similar) na base local Qualis. As tabelas com os *scores* extraídos para cada uma das funções de similaridade analisadas estão no Anexo C. A tabela 5.6 exibe os resultados obtidos:

Tabela 5.6: Resultado das funções na similaridade Lattes-Qualis em anais de eventos

Função	$f_{\max}$	*discernability*	$\left[t_{best}^{\min}, t_{best}^{\max}\right]$
Smith-Waterman	28	0,352	[0,806, 0,812]
Cosine Similarity	26	0,327	[0,572, 0,577]
Levenshtein	18	0,271	[0,611, 0,704]
Jaro	2	0,026	[0,811, 0,815]

Os dados da tabela 5.6 indicam que a aplicação das funções de similaridade neste caso não produziu resultados tão expressivos como na similaridade Lattes-Qualis para periódicos. Este fato já era esperado, devido à grande falta de padronização nos títulos dos anais de eventos na base local Qualis (ver tabela 4.7). Novamente, a função *Smith-Waterman* produziu o melhor resultado geral final. Destaca-se a boa performance da função *Cosine Similarity*, pois em diversas situações apenas a ordem dos termos estava trocada nas *strings* de consulta. Por exemplo: *s="SBBD 2004 19 SIMPOSIO BRASILEIRO DE BANCO DE DADOS"* e *t="19 SIMPOSIO BRASILEIRO DE BANCO DE DADOS SBBD 2004"*.

A função *Smith-Waterman* com valor de *threshold* 0,809 (média entre os valores mínimo e máximo do intervalo de *thresholds* ótimo da função na tabela 5.6) foi adotada como padrão no processamento da similaridade Lattes-Qualis para anais de eventos. A tabela 5.7 exibe os resultados do processamento para os 12 pesquisadores do conjunto de dados:

Tabela 5.7: Similaridade Lattes-Qualis – anais de eventos (*Smith-Waterman*, 0,809)

Pesquisador	*Anais de Eventos no Indicador TPAE* *(Triênio 2004-2006)*	*Anais de Eventos sem similares na Base Qualis (< 0,809)*	*Anais de Eventos com similares na Base Qualis (≥ 0,809)*
P1	11	10	1
P2	3	2	1
P3	17	6	11
P4	24	18	6
P5	15	9	6
P6	13	13	0
P7	7	4	3
P8	9	4	5
P9	15	6	9

P10	22	11	11
P11	47	26	21
P12	8	6	2
Total	**191**	**115**	**76**

5.3 Segundo Experimento: Aplicação do Modelo

O objetivo deste experimento foi apresentar uma aplicação prática do modelo de pontuação na busca de competências acadêmicas de pesquisadores descrito no Capítulo 3. O modelo foi instanciado com as informações disponíveis nos 12 CV-Lattes de pesquisadores doutores da área da Ciência da Computação.

5.3.1 Cálculo do coeficiente de competência CC_c

Para o cálculo do CC_c, foram considerados todos os indicadores quantitativos relacionados ao currículo descritos na Seção 3.1.1.

Os pesos de cada indicador foram determinados a partir da análise efetuada por Cazella (2006), que está no Anexo A desta dissertação. Na referida análise, Cazella solicitou a 25 pesquisadores doutores da área da Ciência da Computação que identificassem a importância dos indicadores de produção disponíveis no CV-Lattes. Por exemplo, conforme mostra o Anexo A, o indicador *trabalhos em anais de eventos* (TPAE) foi avaliado com uma importância individual de 28% na sua categoria (que possui uma importância de categoria igual a 46%). Para determinar o peso do indicador TPAE, calcula-se a relação do mesmo com todos os demais indicadores avaliados: TPAE=28*46/100=12,88. A tabela 5.8 sumariza os indicadores adotados com seus respectivos pesos no cálculo do CC_c:

Tabela 5.8: Indicadores quantitativos do currículo e seus respectivos pesos

Indicador	*Sigla*	*Peso*
Artigos publicados em periódicos	APP	18,56
Livros publicados ou organizados	LPO	9
Capítulos de livros publicados	CLP	4,34
Trabalhos publicados em anais de eventos	TPAE	12,88
Textos em jornais ou revistas	TJR	1,84
Outros tipos de produção bibliográfica	OPB	1,38
Software	SW	2,55
Produtos tecnológicos	PTEC	2,7
Trabalhos técnicos	TT	3
Outros tipos de produção técnica (organização de eventos)	OPTOE	3,3
Outros tipos de produção técnica (relatórios de pesquisa)	OPTRP	2,1
Outros tipos de produção técnica (apresentação de trabalhos)	OPTAP	1,35

Orientações concluídas de doutorado	OCD	13,05
Orientações concluídas de mestrado	OCM	7,54
Outras orientações concluídas	OOC	8,41
Participações em bancas de trabalhos de conclusão de doutorado	PBTCD	1,5
Participações em bancas de trabalhos de conclusão de mestrado	PBTCM	0,5
Outras participações em bancas de trabalhos de conclusão	OPBTC	0,24
Participações em bancas de comissões julgadoras	PBCJ	2,96
Participações em eventos	PE	1,52
Orientações em andamento de doutorado	OAD	0,72
Orientações em andamento de mestrado	OAM	0,36
Outras orientações em andamento	OOA	0,2
	Total	**100**

Fonte: Adaptado de CAZELLA, 2006. p. 106-107

Os resultados da extração dos indicadores estão na tabela 5.9. Esta tabela indica, por exemplo, que o pesquisador P1 possui 3 artigos publicados em periódicos e 18 trabalhos publicados em anais de eventos. Já a tabela 5.10 apresenta a normalização dos dados da tabela 5.9 para a escala [0, 10], de acordo com a equação 14 da Seção 3.2.1.

Tabela 5.9: Indicadores extraídos para o cálculo do CC_c

Pesq.	*Indicadores*																						
	APP	LPO	CLP	TPAE	TJR	OPB	SW	PTEC	TT	OPTOE	OPTRP	OPTAP	OCD	OCM	OOC	PBTCD	PBTCM	OPBTC	PBCJ	PE	OAD	OAM	OOA
P1	3	0	0	18	1	3	0	0	31	4	0	11	0	0	4	0	3	0	1	23	0	2	1
P2	2	0	1	12	0	3	0	0	2	0	0	0	0	0	1	0	6	17	4	2	0	1	1
P3	6	3	2	44	0	12	2	0	0	15	0	9	0	9	16	3	38	18	6	29	4	4	1
P4	15	4	6	143	5	17	0	0	14	2	0	11	9	44	39	12	60	9	1	15	2	3	0
P5	9	2	2	100	1	1	2	0	1	0	0	0	4	40	1	2	61	5	8	1	5	3	0
P6	3	0	3	37	0	5	1	0	12	0	0	2	0	0	13	1	4	25	0	23	0	0	2
P7	1	0	0	24	0	7	0	0	7	6	0	0	0	1	26	0	3	44	7	12	2	4	4
P8	7	6	0	85	0	0	1	0	0	0	0	0	5	35	0	5	8	2	8	28	2	0	0
P9	2	0	0	32	0	0	1	14	18	3	0	1	0	0	66	0	6	2	10	0	0	0	5
P10	11	0	2	110	0	28	1	0	9	4	0	1	5	29	44	0	0	0	0	0	4	2	0
P11	24	2	7	160	2	12	2	1	2	0	0	0	3	12	21	13	31	17	4	5	5	3	0
P12	3	0	2	30	1	0	1	1	0	0	0	4	0	5	39	0	6	8	4	24	1	4	2

Tabela 5.10: Indicadores normalizados na escala [0, 10] para o cálculo do CC_c

Pesq.	*Indicadores*																						
	APP	LPO	CLP	TPAE	TJR	OPB	SW	PTEC	TT	OPTOE	OPTRP	OPTAP	OCD	OCM	OOC	PBTCD	PBTCM	OPBTC	PBCJ	PE	OAD	OAM	OOA
P1	0,87	0,00	0,00	0,41	2,00	1,07	0,00	0,00	10,00	2,67	0,00	10,00	0,00	0,00	0,61	0,00	0,49	0,00	1,00	7,93	0,00	5,00	2,00
P2	0,43	0,00	1,43	0,00	0,00	1,07	0,00	0,00	0,65	0,00	0,00	0,00	0,00	0,00	0,15	0,00	0,98	3,86	4,00	0,69	0,00	2,50	2,00
P3	2,17	5,00	2,86	2,16	0,00	4,29	10,00	0,00	0,00	10,00	0,00	8,18	0,00	2,05	2,42	2,31	6,23	4,09	6,00	10,00	8,00	10,00	2,00
P4	6,09	6,67	8,57	8,85	10,00	6,07	0,00	0,00	4,52	1,33	0,00	10,00	10,00	10,00	5,91	9,23	9,84	2,05	1,00	5,17	4,00	7,50	0,00
P5	3,48	3,33	2,86	5,95	2,00	0,36	10,00	0,00	0,32	0,00	0,00	0,00	4,44	9,09	0,15	1,54	10,00	1,14	8,00	0,34	10,00	7,50	0,00
P6	0,87	0,00	4,29	1,69	0,00	1,79	5,00	0,00	3,87	0,00	0,00	1,82	0,00	0,00	1,97	0,77	0,66	5,68	0,00	7,93	0,00	0,00	4,00
P7	0,00	0,00	0,00	0,81	0,00	2,50	0,00	0,00	2,26	4,00	0,00	0,00	0,00	0,23	3,94	0,00	0,49	10,00	7,00	4,14	4,00	10,00	8,00
P8	2,61	10,00	0,00	4,93	0,00	0,00	5,00	0,00	0,00	0,00	0,00	0,00	5,56	7,95	0,00	3,85	1,31	0,45	8,00	9,66	4,00	0,00	0,00
P9	0,43	0,00	0,00	1,35	0,00	0,00	5,00	10,00	5,81	2,00	0,00	0,91	0,00	0,00	10,00	0,00	0,98	0,45	10,00	0,00	0,00	0,00	10,00
P10	4,35	0,00	2,86	6,62	0,00	10,00	5,00	0,00	2,90	2,67	0,00	0,91	5,56	6,59	6,67	0,00	0,00	0,00	0,00	0,00	8,00	5,00	0,00
P11	10,00	3,33	10,00	10,00	4,00	4,29	10,00	0,71	0,65	0,00	0,00	0,00	3,33	2,73	3,18	10,00	5,08	3,86	4,00	1,72	10,00	7,50	0,00
P12	0,87	0,00	2,86	1,22	2,00	0,00	5,00	0,71	0,00	0,00	0,00	3,64	0,00	1,14	5,91	0,00	0,98	1,82	4,00	8,28	2,00	10,00	4,00

O coeficiente CC_c foi calculado através da média aritmética ponderada (ver equação 15 da Seção 3.2.2) entre os indicadores normalizados da tabela 5.10 com seus respectivos pesos descritos na tabela 5.8. Os resultados estão na tabela 5.11:

Tabela 5.11: Resultados do cálculo do CC_c

Pesquisador	CC_c
P1	1,01
P2	0,35
P3	2,87
P4	6,74
P5	3,84
P6	1,18
P7	1,06
P8	3,95
P9	2,07
P10	4,09
P11	5,64
P12	1,57

Os dados da tabela 5.11 indicam as pontuações obtidas pelos pesquisadores. O coeficiente CC_c permite mensurar em termos quantitativos a produtividade acadêmica do cientista. No caso, o pesquisador P4 atingiu a maior pontuação neste quesito, seguido pelo pesquisador P11. Observando os indicadores de ambos na tabela 5.9, percebe-se que embora o pesquisador P11 possua maior quantidade de publicações, na categoria orientações concluídas seu desempenho é inferior ao pesquisador P4, o que explica sua menor pontuação. Devido à flexibilidade na atribuição dos pesos, o cálculo do CC_c permite capturar diferentes noções de produtividade, ficando a critério dos responsáveis pela instanciação do modelo a definição da importância dos indicadores.

Com base nas informações da tabela 5.11, é possível gerar um *ranking* dos pesquisadores ordenando-os pelo valor de CC_c, conforme mostra a tabela 5.12. Nesta dissertação o *ranking* foi gerado apenas para possibilitar uma comparação entre os diferentes coeficientes de competência. Ele não deve ser adotado diretamente como classificação da competência dos pesquisadores porque o modelo não possui informações temporais sobre a "idade acadêmica" dos cientistas. O coeficiente CC_c tende a privilegiar pesquisadores com mais tempo de doutorado, pois teoricamente o volume da produção de um cientista doutor aumenta com o passar do tempo. Assim sendo, faz-se necessário normalizar os resultados do coeficiente CC_c de acordo com a idade acadêmica do pesquisador. Este processo de normalização é tema de trabalhos futuros da pesquisa.

Tabela 5.12: Ranking dos pesquisadores em termos do coeficiente CC_c

Posição	*Pesquisador*
1	P4
2	P11
3	P10
4	P8
5	P5
6	P3
7	P9
8	P12
9	P6
10	P7
11	P1
12	P2

5.3.2 Cálculo do coeficiente de competência CC_b

Para o cálculo do CC_b, foram considerados todos os indicadores quantitativos relacionados à produção bibliográfica descritos na Seção 3.1.2.

Os pesos de cada indicador foram determinados empiricamente de acordo com a tabela 5.13, que apresenta a importância atribuída a cada categoria e a importância relativa de cada indicador. O cálculo do peso do indicador é feito através da relação do indicador com todos os demais indicadores avaliados. Por exemplo: PQARI=40*50/100=20.

Tabela 5.13: Importância dos indicadores quantitativos da produção bibliográfica

Categoria: Classificação dos Veículos de Publicação – Periódicos			**% Imp. da Cat.**
Indicador	*% Imp. do Indicador*	*Peso Calculado*	50
PQARI	40	20	
PQARN	15	7,50	
PQARL	0	0	
PQBRI	20	10	
PQBRN	7,5	3,75	
PQBRL	0	0	

PQCRI	15	7,50	
PQCRN	2,5	1,25	
PQCRL	0	0	
Categoria: Classificação dos Veículos de Publicação – Anais de Eventos			**% Imp. da Cat.**
Indicador	*% Imp. do Indicador*	*Peso Calculado*	
EQARI	40	10	25
EQARN	15	3,75	
EQARL	0	0	
EQBRI	20	5	
EQBRN	7,5	1,875	
EQBRL	0	0	
EQCRI	15	3,75	
EQCRN	2,5	0,625	
EQCRL	0	0	
Categoria: Repercussão e Impacto na Comunidade Acadêmica			**% Imp. da Cat.**
Indicador	*% Imp. do Indicador*	*Peso Calculado*	
TOTCIT	25	6,25	25
RCIT	25	6,25	
h-index	50	12,50	

Os resultados da extração dos indicadores estão na tabela 5.14. No caso dos indicadores relacionados à classificação dos veículos de publicação, a extração limitou-se ao triênio 2004-2006. A tabela 5.15 apresenta a normalização dos dados da tabela 5.14 para a escala [0, 10], de acordo com a equação 14 descrita na Seção 3.2.1.

Tabela 5.14: Indicadores extraídos para o cálculo do CC_b

Pesq.	*Indicadores*																				
	PQARI	PQARN	PQARL	PQBRI	PQBRN	PQBRL	PQCRI	PQCRN	PQCRL	EQARI	EQARN	EQARL	EQBRI	EQBRN	EQBRL	EQCRI	EQCRN	EQCRL	TOTCIT	RCIT	h-index
P1	0	0	0	0	1	0	1	0	0	1	0	0	0	0	0	0	0	0	28	1,33	2
P2	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	28	1,87	4
P3	1	0	0	0	0	0	0	0	0	0	0	0	1	10	0	0	0	0	169	3,07	6
P4	1	1	0	0	0	0	0	0	0	1	2	0	0	2	0	0	1	0	237	1,41	7
P5	0	0	0	0	0	0	0	3	0	1	4	0	0	0	0	1	0	0	109	0,96	5
P6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	93	2,16	4
P7	1	0	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	0	21	0,84	3
P8	0	0	0	0	0	0	0	0	0	3	2	0	0	0	0	0	0	0	324	3,31	9
P9	0	0	0	0	0	0	0	0	0	0	9	0	0	0	0	0	0	0	29	0,85	3
P10	0	0	0	0	0	0	1	0	0	0	2	0	3	6	0	0	0	0	150	1,22	6
P11	3	1	0	7	1	0	0	0	0	9	5	0	1	0	0	6	0	0	505	2,62	11
P12	2	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	52	1,48	5

Tabela 5.15: Indicadores normalizados na escala [0, 10] para o cálculo do CC_b

Pesq.	*Indicadores*																				
	PQARI	PQARN	PQARL	PQBRI	PQBRN	PQBRL	PQCRI	PQCRN	PQCRL	EQARI	EQARN	EQARL	EQBRI	EQBRN	EQBRL	EQCRI	EQCRN	EQCRL	TOTCIT	RCIT	h-index
P1	0,00	0,00	0,00	0,00	10,00	0,00	10,00	0,00	0,00	1,11	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,14	1,98	0,00
P2	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	1,11	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,14	4,17	2,22
P3	3,33	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	3,33	10,00	0,00	0,00	0,00	0,00	3,06	9,03	4,44
P4	3,33	10,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	1,11	2,22	0,00	0,00	2,00	0,00	0,00	10,00	0,00	4,46	2,31	5,56
P5	0,00	0,00	0,00	0,00	0,00	0,00	0,00	10,00	0,00	1,11	4,44	0,00	0,00	0,00	0,00	1,67	0,00	0,00	1,82	0,49	3,33
P6	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	1,49	5,34	2,22
P7	3,33	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	3,33	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	1,11
P8	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	3,33	2,22	0,00	0,00	0,00	0,00	0,00	0,00	0,00	6,26	10,00	7,78
P9	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	10,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,17	0,04	1,11
P10	0,00	0,00	0,00	0,00	0,00	0,00	10,00	0,00	0,00	0,00	2,22	0,00	10,00	6,00	0,00	0,00	0,00	0,00	2,67	1,54	4,44
P11	10,00	10,00	0,00	10,00	10,00	0,00	0,00	0,00	0,00	10,00	5,56	0,00	3,33	0,00	0,00	10,00	0,00	0,00	10,00	7,21	10,00
P12	6,67	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,00	1,11	1,11	0,00	0,00	0,00	0,00	0,00	0,00	0,00	0,64	2,59	3,33

O coeficiente CC_b foi calculado através da média aritmética ponderada (ver equação 15 da Seção 3.2.2) entre os indicadores normalizados da tabela 5.15 com seus respectivos pesos descritos na tabela 5.13. Os resultados estão na tabela 5.16:

Tabela 5.16: Resultados do cálculo do CC_b

Pesquisador	CC_b
P1	1,37
P2	0,59
P3	2,33
P4	2,83
P5	1,03
P6	0,70
P7	0,93
P8	2,41
P9	0,53
P10	2,26
P11	8,20
P12	2,10

Os dados da tabela 5.16 indicam as pontuações obtidas pelos pesquisadores. O coeficiente CC_b permite mensurar em termos quantitativos a importância dos trabalhos do pesquisador, considerando a classificação dos veículos de publicação e a repercussão ou impacto dos trabalhos na comunidade acadêmica. De imediato, percebe-se que neste quesito o pesquisador P11 apresentou resultado bastante superior. Conforme mostram os indicadores da tabela 5.14, no triênio 2004-2006 o pesquisador P11 registrou a maior quantidade de trabalhos publicados em periódicos ou eventos com qualidade A e repercussão internacional. Ainda, considerando toda a produção, o pesquisador P11 acumulou a maior quantidade de citações bem como o maior *h-index*, o que justifica sua alta pontuação. Novamente, é importante ressaltar que o modelo é flexível na atribuição dos pesos aos indicadores. Por exemplo, para considerar apenas o indicador *h-index* no cálculo do coeficiente de competência CC_b, basta zerar os pesos dos demais indicadores. Assim sendo, verifica-se que o modelo permite capturar diferentes noções de importância dos trabalhos de um cientista.

Com base nas informações da tabela 5.16, é possível gerar um *ranking* dos pesquisadores ordenando-os pelo valor de CC_b, conforme mostra a tabela 5.17. Nesta dissertação o *ranking* foi gerado apenas para possibilitar uma comparação entre os diferentes coeficientes de competência.

Tabela 5.17: Ranking dos pesquisadores em termos do coeficiente CC_b

Posição	*Pesquisador*
1	P11
2	P4
3	P8
4	P3
5	P10
6	P12
7	P1
8	P5
9	P7
10	P6
11	P2
12	P9

Observe que o *ranking* mostrado na tabela 5.17 é diferente do ranking da tabela 5.12, indicando que o modelo capturou diferentes noções de competência. Na tabela 5.12, a competência foi medida em termos da produtividade acadêmica dos pesquisadores. Já na tabela 5.17, mediu-se a importância da produção bibliográfica dos mesmos. Esta flexibilidade constitui-se na contribuição principal do modelo.

5.3.3 Junção dos coeficientes de competência

Finalmente, o modelo permite unir os coeficientes de competência CC_c e CC_b em um único coeficiente de competência geral. Para isso, atribui-se um peso para cada um dos coeficientes e aplica-se novamente uma média aritmética ponderada. Assim, os responsáveis pela instanciação do modelo podem decidir qual coeficiente deve ser priorizado no cálculo da pontuação final. Neste experimento, ambos coeficientes receberam um peso de 50%.

As tabelas 5.18 e 5.19 mostram os resultados finais:

Tabela 5.18: Resultados do cálculo do *CC*

Pesquisador	*CC*
P1	1,19
P2	0,47
P3	2,60
P4	4,78
P5	2,43
P6	0,94

P7	1,00
P8	3,18
P9	1,30
P10	3,18
P11	6,92
P12	1,84

Tabela 5.19: Ranking dos pesquisadores em termos do coeficiente *CC*

Posição	*Pesquisador*
1	P11
2	P4
3	P8
4	P10
5	P3
6	P5
7	P12
8	P9
9	P1
10	P7
11	P6
12	P2

5.3.4 Conclusões e limitações

Os dois experimentos realizados foram de fundamental importância para finalização desta dissertação, pois permitiram:

1. Verificar a viabilidade da abordagem baseada em funções de similaridade na integração das diferentes fontes de informação. Este experimento demonstrou que em duas situações (similaridade Lattes-Scholar e similaridade Lattes-Qualis para periódicos) os resultados foram altamente expressivos. Já na similaridade Lattes-Qualis para anais de eventos, considera-se o resultado satisfatório, sendo que trabalhos futuros irão concentrar seus esforços em aprimorar a performance das funções de similaridade;
2. Aplicar o modelo de pontuação na busca de competências acadêmicas em uma situação real. O resultado demonstrou que o modelo possibilita capturar diferentes noções de competência, relacionadas com a produtividade acadêmica e com a importância da produção bibliográfica dos pesquisadores.

Uma limitação dos experimentos é que os mesmos foram realizados com amostras sem valor estatístico. A abordagem de Silva et al. 2007, adotada no primeiro experimento, não especifica diretamente o tamanho da amostra ideal para os cálculos de intervalos de *thresholds* ótimos, f_{max} e *discernability*. Além disso, a referida abordagem exige um grande esforço manual na determinação da relevância dos elementos da amostra. O segundo experimento também foi realizado com uma amostra restrita de currículos disponíveis, o que torna os resultados interessantes, porém não generalizáveis.

6 CONCLUSÃO

Nesta dissertação foi apresentado um modelo para descobrir e pontuar competências acadêmicas de pesquisadores, baseado na combinação de indicadores quantitativos que permitem mensurar a produção acadêmica dos cientistas.

A partir dos resultados dos experimentos realizados evidenciou-se que o modelo é capaz de capturar diferentes noções de competência, pois separa a avaliação quantitativa da produção acadêmica em dois critérios:

1. Produtividade – utiliza indicadores de produção para avaliar o volume da produção acadêmica dos pesquisadores. Ressalta-se que o modelo suporta múltiplas categorias de indicadores (produção bibliográfica, produção técnica, orientações concluídas e informações complementares), descritas em detalhes no Capítulo 3, a fim de representar várias dimensões da produtividade de um cientista;
2. Importância da produção bibliográfica – este critério destaca-se como diferencial do modelo, e emprega indicadores quantitativos relacionados com a importância da produção bibliográfica dos pesquisadores. Estes indicadores avaliam a produção em termos da repercussão ou impacto dos trabalhos do pesquisador na comunidade científica, bem como através da análise da qualidade e alcance dos veículos de publicação nos quais o pesquisador possui trabalhos publicados.

Além do referido modelo, esta dissertação contribuiu com a especificação de uma arquitetura flexível e extensível fundamentada em técnicas de extração de dados na Web e casamento aproximado de dados (através de funções de similaridade). A arquitetura foi implementada em um sistema Web cuja principal característica é a integração de diversas tecnologias *open source*. O sistema desenvolvido permite que qualquer pesquisador avalie quantitativamente sua produção científica, automatizando diversos aspectos relacionados à tarefa de avaliação, como a obtenção dos indicadores e a integração das diferentes bases de informações.

Os resultados dos experimentos também demonstraram que a abordagem baseada em funções de similaridade é viável na integração das diferentes fontes de informação consideradas (CV-Lattes, Google Scholar e Qualis-CAPES). Desta forma, evidenciou-se a aplicabilidade da abordagem para minimizar o impacto de informações inconsistentes (no caso, evitar que citações de trabalhos que não são da autoria do pesquisador sejam indevidamente atribuídas ao mesmo), e recuperar informações a partir de bases de dados construídas sem relacionamentos por chaves primárias e restrições conjuntas de integridade.

A pesquisa realizada nesta dissertação identifica os seguintes trabalhos futuros, visando aprimorar os resultados já obtidos:

1. Disponibilizar o sistema Web para a comunidade acadêmica, permitindo que os pesquisadores avaliem sua produção científica de acordo com os critérios suportados pelo modelo descrito no Capítulo 3. Com isso, será possível obter um volume de dados expressivo e realizar experimentos com base estatística;
2. Conduzir um estudo exaustivo sobre funções de similaridade, a fim de identificar as abordagens ótimas aos problemas de integração das fontes de informações aqui consideradas. Este estudo deve focar-se em aspectos como performance e qualidade de recuperação e também investigar formas de otimização do processamento da similaridade (em termos de tempo de execução e consumo de recursos computacionais);
3. Incluir informações para modelar as diferentes áreas (e subáreas) de conhecimento. No presente trabalho, o modelo limita-se a calcular os coeficientes de competência de uma forma geral. Porém, um pesquisador da área geral Ciência da Computação, por exemplo, na realidade possui competências distribuídas em subáreas específicas (Inteligência Artificial, Engenharia de Software, Banco de Dados, Sistemas de Recomendação, etc.). Além disso, faz-se necessário incluir informações temporais no modelo, com o objetivo de acompanhar a evolução temporal dos coeficientes de competência e identificar mudanças de áreas de interesse dos pesquisadores. Assim sendo, está em andamento uma dissertação de mestrado cujo objetivo é construir uma ontologia de pesquisadores, que suportará informações temporais e modelará a granularidade das áreas de conhecimento;
4. Integrar os coeficientes de competência em outras aplicações, como por exemplo: sistemas de recomendação, sistemas baseados em redes sociais e comunidades virtuais, sistemas de gestão de competências, entre outros. Está em andamento a integração do modelo em um sistema para edição colaborativa de artigos científicos com processo de avaliação aberta (OLIVEIRA et al., 2005). Os usuários deste sistema serão classificados em categorias específicas (leitores, revisores, avaliadores) de acordo com as pontuações indicadas pelos coeficientes de competência modelados na presente dissertação;
5. Estudar as características dos diferentes *rankings* gerados com base nos coeficientes de competência descritos na presente pesquisa. Este estudo já está em andamento, e adota uma abordagem baseada em fusão de dados (DUTRA e LIMA, 2006);
6. Investigar em maiores detalhes aspectos relacionados às fontes de informação utilizadas: (i) verificar a cobertura das informações disponibilizadas pelo serviço Google Scholar, bem como efetuar uma análise sobre auto-citações e co-autoria de artigos; (ii) identificar e considerar outras fontes para obtenção dos indicadores (e analisar a viabilidade de incluir outros indicadores no modelo);
7. Generalizar o modelo, pois o mesmo está adaptado à realidade brasileira em virtude das fontes de informações consideradas.

REFERÊNCIAS

ARLOTTA, L. et al. Automatic annotation of data extracted from large web sites. In: INT. WORKSHOP ON WEB AND DATABASES, WEBDB, 6., 2003, San Diego, CA. **Proceedings...** [S.l.: s.n.], 2003.

BAEZA-YATES, R.; RIBEIRO-NETO, B. **Modern Information Retrieval**. New York: ACM Press, 1999.

BERNERS-LEE, T.; HENDLER, J.; LASSILA, O. The Semantic Web. **Scientific American**, [S.l.], May 2001.

BERRETA, E. C. **Uma Aplicação de Integração de Dados por Similaridade**. 2006. 48 f. Projeto de Diplomação (Bacharelado em Ciência da Computação) – Instituto de Informática, UFRGS, Porto Alegre.

BILENKO, M. et al. Adaptive name matching in information integration. **IEEE Intelligent Systems**, [S.l.], v.18, n.5, p. 16-23, 2003.

BORGES, T. B. et al. Identificação Automática de Expertise Analisando Currículos no Formato Lattes. In: SIMPÓSIO BRASILEIRO DE SISTEMAS DE INFORMAÇÃO, 1., 2004, Porto Alegre. **Anais...** Porto Alegre: PUCRS, 2004. p. 127-134.

CAPES. **Sistema Qualis-CAPES**. Disponível em: <http://servicos.capes.gov.br/webqualis/>. Acesso em: mar. 2007.

CAZELLA, S. C.; ALVARES, L. O. C. **W-RECMAS:** Um sistema de recomendação híbrido baseado em sistemas multiagentes para recomendação de artigos científicos. Porto Alegre: Instituto de Informática, UFRGS, 2003. 86f. (RP – 339).

CAZELLA, S. C. **Aplicando a Relevância da Opinião de Usuários em Sistema de Recomendação para Pesquisadores**. 2006. 180 f. Tese (Doutorado em Ciência da Computação) – Instituto de Informática, UFRGS, Porto Alegre.

CNPQ. **Sistema CV-Lattes**. Disponível em: <http://plsql1.cnpq.br/images/ajuda/conhecendo-o-sistema.html>. Acesso em: mar. 2007.

COHEN, W. W.; RAVIKUMAR, P; FIENBERG, S. E. A comparison of string distance metrics for name-matching tasks. In: INTERNATIONAL JOINT CONFERENCE ON ARTIFICIAL INTELLIGENCE, IJCAI, 18., 2003, Acapulco. **Proceedings...** San Francisco, CA: Morgan Kaufmann, 2003.

CRESCENZI, V.; MECCA, G.; MERIALDO, P. RoadRunner: Towards automatic data extraction from large Web sites. In: INTERNATIONAL CONFERENCE ON VERY LARGE DATA BASES, 26., 2001, Rome, Italy. **Proceedings...** [S.l.: s.n.], 2001.

DAVYT, A.; VELHO, L. The evaluation of science and peer review: past and present. What will the future be like? **História, Ciências, Saúde-Manguinhos**, Rio de Janeiro, v. 7, n. 1, 2000. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0104-59702000000200005&lng=en&nrm=iso>. Acesso em: mar. 2007.

DORNELES, C. F. **Uma estratégia genérica para casamento aproximado de instâncias**. 2006. 107 f. Tese (Doutorado em Ciência da Computação) – Instituto de Informática, UFRGS, Porto Alegre.

DUTRA, E. G.; LIMA, J. V. Completamento de Ranks Parciais para Fusão de Dados. In: SIMPÓSIO BRASILEIRO DE SISTEMAS MULTIMÍDIA E WEB, WEBMEDIA, 12., 2006, Natal. **Anais...** Porto Alegre, RS: Sociedade Brasileira de Computação, 2006.

FERREIRA, A. B. H. **Novo Aurélio Século XXI:** o dicionário da língua portuguesa. 3. ed. Rio de Janeiro: Nova Fronteira, 1999.

GROSSI, M. A. **Avaliando a Performance das Funções de Similaridade através da Revocação e Precisão**. 2005. 61 f. Projeto de Diplomação (Bacharelado em Ciência da Computação) – Instituto de Informática, UFRGS, Porto Alegre.

GUSFIELD, D. **Algorithms on Strings, Trees and Sequences**. Cambridge: Cambridge University Press, 1997.

HAMMER, J. et al. Extracting semistructured information from the web. In: WORKSHOP ON MANAGEMENT OF SEMISTRUCTURED DATA, 1997. **Proceedings...** [S.l.: s.n.], 1997.

HANDSCHUH, S.; STAAB, S.; CIRAVEGNA, F. S-CREAM – Semi-automatic CREAtion of Metadata. In: INTERNATIONAL CONFERENCE ON KNOWLEDGE ENGINEERING AND KNOWLEDGE MANAGEMENT, EKAW, 13., 2002. **Proceedings...** Berlim: Springer-Verlag, 2002, p. 358-372.

HIRSCH, J. E. **An index to quantify an individual's scientific research output**. Disponível em: <http://xxx.arxiv.org/abs/physics/0508025>. Acesso em: mar. 2007.

HOUAISS, A.; VILLAR, M. S. **Dicionário Houaiss da Língua Portuguesa**. Rio de Janeiro: Objetiva, 2001.

JARO, M. A. Advances in record-linkage methodology as applied to matching the 1985 census of Tampa, Florida. **Journal of the American Statistical Association**, [S.l.], v.84, p. 414–420, 1989.

KOSTOFF, R. N. The use and misuse of citation analysis in research evaluation. **Scientometrics**, [S.l.], v.43, n.1, p. 27–43, Sept. 1998.

LAENDER, A. A Brief Survey of Web Data Extraction Tools. **SIGMOD Record**, New York, v.31, n.2, p. 84-92, June 2002.

LEVENSHTEIN, V. I. Binary codes capable of correcting deletions, insertions and reversals. **Soviet Physics Doklady**, [S.l.], v. 10, n. 8, p. 707-710, 1966.

LIU, L.; PU, C.; HAN, W. XWRAP: An XML-enabled wrapper construction system for Web information sources. In: INTERNATIONAL CONFERENCE ON DATA ENGINEERING, 16., 2000, San Diego, CA. **Proceedings...** [S.l.: s.n.], 2000.

MACIAS-CHAPULA, C. A. The role of informetrics and scientometrics in the national and international perspective. **Ciência da Informação**, Brasília, v. 27, n. 2, 1998. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0100-19651998000200005&lng=en&nrm=iso>. Acesso em: mar. 2007.

MUSLEA, I.; MINTON, S.; KNOBLOCK, C. Hierarchical wrapper induction for semistructured information sources. **Autonomous Agents and Multi-Agent Systems**, [S.l.], v.1, n.2, p. 93-114, 2001.

MYLLYMAKI, J. Effective Web data extraction with standard XML technologies. In: THE INTERNATIONAL CONFERENCE ON WORLD WIDE WEB, 10., 2001, Hong Kong. **Proceedings...** [S.l.: s.n.], 2001.

NEEDLEMAN, S. B.; WUNSCH, C. D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. **J. Mol. Biol.**, [S.l.], v.48, p. 443-453, 1970.

NIEDERAUER, C. A. P. **Ethos:** Um Modelo para Medir a Produtividade Relativa de Pesquisadores Baseado na Análise por Envoltória de Dados. 2002. 146 f. Tese (Doutorado em Engenharia de Produção) – Programa de Pós-Graduação em Engenharia de Produção, UFSC, Florianópolis.

NSF. National Science Foundation – Science and Engineering Indicators. Disponível em: < http://www.nsf.gov/statistics/seind06/ >. Acesso em: mar. 2007.

OLIVEIRA, J. P. M. et al. Uma Proposta para Editoração, Indexação e Busca de Documentos Científicos em um Processo de Avaliação Aberta. In: WORKSHOP EM BIBLIOTECAS DIGITAIS, WDL, 1., 2005, Uberlândia, MG. **Anais...** [S.l.: s.n.], 2005.

PRICE, D. **Little science, big science**. New York: Columbia University Press, 1963.

RIBEIRO JUNIOR, L. C. et al. Identificação de Áreas de Interesse a partir da Extração de Informações de Currículos Lattes/XML. In: ESCOLA REGIONAL DE BANCO DE DADOS, 1., 2005, Porto Alegre. Disponível em: <http://www.inf.ufrgs.br/~erbd2005/Artigos/7866.pdf>. Acesso em: set. 2006.

RIBEIRO-NETO, B.; LAENDER, A. H. F.; SILVA, A. S. Extracting semi-structured data through examples. In: ACM CIKM INTERNATIONAL CONFERENCE ON INFORMATION AND KNOWLEDGE MANAGEMENT, 1999, Kansas City, MO. **Proceedings...** [S.l.: s.n.], 1999.

RODRIGUES, S. et al. Competence mining for virtual scientific community creation. **International Journal of Web Based Communities**, [S.l.], v.1, n.1, p. 90-102, July 2004.

ROUSSEAU, R. Bibliometric and econometric indicators for the evaluation of scientific institutions. **Ciência da Informação**, Brasília, v. 27, n. 2, 1998. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0100-19651998000200007&lng=en&nrm=iso>. Acesso em: mar. 2007.

SAHUGUET, A.; AZAVANT, F. Building Light-Weight Wrappers for Legacy Web Data-Sources Using W4F. In: INTERNATIONAL CONFERENCE ON VERY LARGE DATA BASES, 25., 1999. **Proceedings...** [S.l.: s.n.], 1999.

SALTON, G.; MCGILL, M. J. **Introduction to Modern Information Retrieval**. New York: McGraw-Hill Book, 1983.

SALTON, G.; BUCKLEY, C. Term-weighting approaches in automatic retrieval. **Information Processing & Management**, [S.l.], v.24, n.5, p. 513-523, 1988.

SILVA, R. et al. Measuring quality of similarity functions in approximate data matching. **Journal of Informetrics**, [S.l.], v.1, n.1, p. 35–46, Jan. 2007.

SMITH, T. F.; WATERMAN, M. S. Identification of common molecular subsequences. **J. Mol. Biol.**, [S.l.], v.147, p. 195-197, 1981.

SPINAK, E. Scientometric indicators. **Ciência da Informação**, Brasília, v. 27, n. 2, 1998. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S0100-19651998000200006&lng=en&nrm=iso>. Acesso em: mar. 2007.

TICONA, W. G. C. **Aplicação de Algoritmos Genéticos Multi-Objetivo para Alinhamento de Seqüências Biológicas**. 2003. 112 f. Dissertação (Mestrado em Ciências da Computação e Matemática Computacional) – Instituto de Ciências Matemáticas e de Computação, USP, São Paulo.

WIVES, L. K. **Um estudo sobre Agrupamento de Documentos Textuais em Processamento de Informações não Estruturadas Usando Técnicas de "Clustering"**. 1999. 102 f. Dissertação (Mestrado em Ciência da Computação) – Instituto de Informática, UFRGS, Porto Alegre.

ANEXO A IMPORTÂNCIA DOS INDICADORES DE PRODUÇÃO

Neste Anexo é apresentada a importância dos indicadores de produção, conforme descrito em Cazella (2006). A coluna "% Imp. da Cat." indica o percentual de importância da categoria de indicador de produção. A coluna "% Imp. do Ind." indica o percentual de importância de um indicador específico dentro de sua categoria. E a coluna "% Consid. Ind. Rel." indica o percentual de pesquisadores que consideraram o indicador relevante para determinação da relevância de opinião de um cientista. Segundo Cazella, 25 pesquisadores doutores da área da Ciência da Computação contribuíram para a definição da importância dos indicadores de produção.

Categoria: Produção Bibliográfica			%Imp.da Cat.
Indicadores de Produção	% Consid. Ind. Rel.	% Imp.do Ind.	
1) Artigos publicados em periódicos	100	36	
2) Trabalhos em anais de eventos	100	28	46
3) Livros ou Capítulos de livros	100	29	
4) Textos em jornais ou revistas	48	4	
5) Demais tipos de produção bibliográfica	40	3	
TOTAL:		100	
Categoria: Produção Técnica			%Imp.da Cat.
Indicadores de Produção	% Consid. Ind. Rel.	% Imp.do Ind.	
1) Software	95	17	
2) Produtos tecnológicos	85	18	
3) Trabalhos técnicos	93	20	
4) Demais tipos de produção técnica (organização de eventos)	93	22	15
5) Demais tipos de produção técnica (relatórios de pesquisa)	93	14	
6) Demais tipos de produção técnica (apresentações de trabalhos)	85	9	
TOTAL:		100	
Categoria: Orientação Concluída			%Imp.da Cat.
Indicadores de Produção	% Consid. Ind. Rel.	% Imp.do Ind.	
1) Tese doutorado	100	45	
2) Dissertação de mestrado	100	26	29
3) Trabalho de conclusão	96	12	
4) Especialização /Aperfeiçoamento	93	8	
5) Iniciação científica	93	9	
TOTAL:		100	
Categoria: Informações Complementares			%Imp.da Cat.
Indicadores de Produção	% Consid. Ind. Rel.	% Imp.do Ind.	
1) Participações em banca de trabalhos de conclusão	85	28	
2) Participações em eventos	78	19	08
3) Participações em banca de comissões julgadoras	93	37	
4) Orientações em andamento	93	16	
TOTAL:		100	
Categoria: Demais Ttrabalhos Relevantes			%Imp.da Cat.
Indicadores de Produção	% Consid. Ind. Rel.	% Imp.do Ind.	
Demais trabalhos relevantes	19	100	02
TOTAL:		100	
TOTAL Geral:			100

ANEXO B ARQUIVOS DE CONFIGURAÇÃO DO WEB-HARVEST

Este Anexo apresenta os arquivos de configuração utilizados no processo de extração efetuado ao serviço Google Scholar com a ferramenta Web-Harvest versão 0.3. O arquivo `functions.xml` é fornecido pela própria ferramenta. Já o arquivo `scholar.xml` foi desenvolvido nesta dissertação.

Arquivo functions.xml

```
<?xml version="1.0" encoding="UTF-8"?>

<config>
  <function name="download-multipage-list">
    <return>
      <while condition="${pageUrl.toString().trim() != ''}" maxloops="${maxloops}"
index="i">
        <empty>
          <var-def name="content">
            <html-to-xml>
              <http url="${pageUrl}"/>
            </html-to-xml>
          </var-def>

          <var-def name="nextLinkUrl">
            <xpath expression="${nextXpath}">
              <var name="content"/>
            </xpath>
          </var-def>

          <var-def name="pageUrl">
            <template>${sys.fullUrl(pageUrl, nextLinkUrl)}</template>
          </var-def>
        </empty>

        <xpath expression="${itemXPath}">
          <var name="content"/>
        </xpath>
      </while>
    </return>
  </function>
</config>
```

Arquivo ***scholar.xml***

```
<?xml version="1.0" encoding="UTF-8"?>

<config charset="UTF-8">

  <include path="functions.xml"/>

  <var-def name="url">
    <template>http://scholar.google.com/scholar?q=${qScholar}&amp;hl=en&amp;lr=&amp;btnG=Search</template>
  </var-def>

  <var-def name="referencias">
    <call name="download-multipage-list">
      <call-param name="pageUrl"><var name="url"/></call-param>
      <call-param name="nextXPath">//td[.='Next']/a/@href</call-param>
      <call-param name="itemXPath">//p[@class="g"]</call-param>
      <call-param name="maxloops"><var name="qtdePaginas"/></call-param>
    </call>
  </var-def>

  <file action="write" path="referencias_scholar.xml" charset="UTF-8">
    <![CDATA[ <?xml version="1.0" encoding="UTF-8"?>
              <scholar>
    ]]>
    <loop item="item" index="i">
      <list><var name="referencias"/></list>
      <body>
        <xquery>
          <xq-param name="item"><var name="item"/></xq-param>
          <xq-expression><![CDATA[
            let $titulo := data($item//span[@class='w']/a)
            let $autores_veiculo_ano := data($item//font[@size='-1']/span[@class='a'])
            let $citacoes := data($item//font[@size='-1']/a[contains(@href, 'cites=')])
            return
              <referencia>
                <titulo>{normalize-space($titulo)}</titulo>
                <autores_veiculo_ano>{normalize-space($autores_veiculo_ano)}</autores_veiculo_ano>
                <citacoes>{normalize-space($citacoes)}</citacoes>
              </referencia>
          ]]></xq-expression>
        </xquery>
      </body>
    </loop>
    <![CDATA[
              </scholar> ]]>
  </file>

</config>
```

ANEXO C *SCORES* DAS FUNÇÕES DE SIMILARIDADE

Neste Anexo estão os *scores* das funções de similaridade para os experimentos realizados na Seção 5.2 desta dissertação. Nas tabelas C.1 até C.3, s_{rel} representa o menor *score* de um item relevante (similar) e s_{irrel} representa o maior *score* de um item irrelevante (não-similar).

Tabela C.1: Similaridade Lattes-Scholar

	Levenshtein		**Smith-Waterman**		**Jaro**		**Cosine Similarity**	
Item	s_{rel}	s_{irrel}	s_{rel}	s_{irrel}	s_{rel}	s_{irrel}	s_{rel}	s_{irrel}
1	1.0	0.4324	1.0	0.4400	1.0	0.7376	1.0	0.4558
2	1.0	0.3804	1.0	0.4219	1.0	0.7692	1.0	0.4020
3	1.0	0.4314	1.0	0.3921	1.0	0.7291	1.0	0.4364
4	1.0	0.3333	1.0	0.5000	1.0	0.6497	1.0	0.3162
5	0.8750	0.3636	0.9286	0.2557	1.0	0.7347	1.0	0.3857
6	1.0	0.4545	1.0	0.7424	1.0	0.6733	1.0	0.6455
7	0.7901	0.4444	0.9831	0.3820	0.8015	0.7770	0.8216	0.5262
8	0.8070	0.4319	1.0	0.4210	0.8125	0.7093	0.9129	0.3651
9	0.8594	0.4794	0.8947	0.3684	0.8384	0.7536	0.3750	0.2357
10	0.9250	0.3050	0.9594	0.2857	0.9750	0.6906	0.9258	0.3651
11	0.8611	0.3333	0.9677	0.5278	0.7924	0.6526	0.6708	0.2828
12	0.8226	0.4032	0.9608	0.3571	0.9409	0.7342	0.7559	0.3849
13	0.9892	0.4260	0.9945	0.3600	0.9964	0.7507	0.8895	0.4160
14	0.7373	0.3146	0.9831	0.2714	0.8806	0.7264	0.8199	0.2182
15	0.7200	0.3667	1.0	0.2818	0.9067	0.7516	0.8864	0.4029
16	1.0	0.3833	1.0	0.5278	1.0	0.7335	1.0	0.3777
17	1.0	0.3492	1.0	0.7073	1.0	0.6853	1.0	0.5164
18	1.0	0.4333	1.0	0.2700	1.0	0.7139	1.0	0.3086
19	1.0	0.4255	1.0	0.4074	1.0	0.8100	1.0	0.6671
20	1.0	0.6491	1.0	0.5893	1.0	0.8700	1.0	0.5345
21	0.8701	0.4193	0.9920	0.3095	0.9248	0.7489	0.8249	0.5773
22	0.9678	0.3333	1.0	0.4042	0.8485	0.7243	0.9608	0.2264
23	0.9239	0.4891	0.9588	0.6000	0.9119	0.8139	0.9534	0.5393
24	0.9399	0.4787	1.0	0.5328	0.9800	0.7724	0.9636	0.5547
25	0.9839	0.4355	0.9918	0.5370	0.9345	0.7197	0.8333	0.4629
26	0.9878	0.4878	0.9756	0.4153	0.9919	0.7297	0.8889	0.3333
27	0.6552	0.3474	1.0	0.3000	0.8850	0.7340	1.0	0.3145
28	0.7384	0.3999	1.0	0.6500	0.9128	0.7407	0.8165	0.2887

29	0.7097	0.3118	0.8870	0.6000	0.7792	0.7523	0.7877	0.4297
30	0.7101	0.3367	1.0	0.3000	0.9034	0.7448	0.8563	0.2010
31	1.0	0.4419	1.0	0.2877	1.0	0.7286	1.0	0.3813
32	1.0	0.3571	1.0	0.2466	1.0	0.7252	1.0	0.2000
33	1.0	0.3875	1.0	0.4434	1.0	0.7439	1.0	0.3015
34	1.0	0.7333	1.0	0.5431	1.0	0.7825	1.0	0.7303
35	1.0	0.3232	1.0	0.2273	1.0	0.6684	1.0	0.2835
36	1.0	0.3196	1.0	0.3830	1.0	0.7322	1.0	0.3698
37	1.0	0.3721	1.0	0.2301	1.0	0.7133	1.0	0.2860
38	1.0	0.3750	1.0	0.3594	1.0	0.7111	1.0	0.2279
39	0.8362	0.3402	1.0	0.3457	0.9454	0.7615	0.9258	0.4564
40	0.9873	0.3718	0.9936	0.4239	0.9573	0.6992	0.8704	0.4029
41	0.8393	0.3548	1.0	0.2400	0.9464	0.7458	0.9258	0.2611
42	0.7661	0.3578	1.0	0.3500	0.9220	0.7244	0.8864	0.2727
43	0.8167	0.6724	1.0	0.5431	1.0	0.7825	0.9129	0.6000
44	0.9818	0.4074	0.9907	0.2870	0.9630	0.7370	0.8750	0.2673
45	0.5760	0.3131	0.9727	0.3000	0.7932	0.7090	0.6614	0.2857

Tabela C.2: Similaridade Lattes-Qualis para periódicos

	Levenshtein		**Smith-Waterman**		**Jaro**		**Cosine Similarity**	
Item	s_{rel}	s_{irrel}	s_{rel}	s_{irrel}	s_{rel}	s_{irrel}	s_{rel}	s_{irrel}
1	1.0	0.6562	1.0	0.6999	1.0	0.7333	1.0	0.5000
2	0.8636	0.5714	0.8809	0.7143	0.8871	0.7592	0.6666	0.4082
3	1.0	0.8222	1.0	0.8378	1.0	0.8870	1.0	0.8164
4	0.7916	0.8421	1.0	0.8684	0.9305	0.8947	0.8660	0.7745
5	0.8409	0.6216	1.0	0.8333	0.8028	0.7877	0.9128	0.5477
6	1.0	0.6097	1.0	0.7857	1.0	0.8330	1.0	0.6666
7	1.0	0.8000	1.0	0.8333	1.0	0.9019	1.0	0.8660
8	0.7500	0.5000	1.0	0.5555	0.9166	0.7407	0.8164	0.3162
9	1.0	0.6666	1.0	0.6428	1.0	0.8419	1.0	0.6666
10	0.8888	0.8888	1.0	0.9375	0.8587	0.9212	0.8944	0.7500
11	1.0	0.6097	1.0	0.7500	1.0	0.7835	1.0	0.5773
12	0.8571	0.7142	0.9166	0.7826	0.9246	0.8534	0.8164	0.6666
13	0.9836	0.7049	0.9916	0.8500	0.9445	0.8733	0.8571	0.6761
14	1.0	0.6190	1.0	0.8333	1.0	0.8362	1.0	0.8333
15	1.0	0.8108	1.0	0.8750	1.0	0.8888	1.0	0.6708
16	0.8518	0.5555	1.0	0.8750	0.8056	0.7731	0.8164	0.3333
17	1.0	0.7500	1.0	0.7608	1.0	0.8325	1.0	0.5000
18	0.8769	0.5692	0.9745	0.7500	0.8505	0.7752	0.8249	0.4472
19	0.8787	0.6000	0.9310	0.6999	0.9136	0.7663	0.9128	0.4472
20	1.0	0.8333	1.0	0.8750	1.0	0.8943	1.0	0.8000

Tabela C.3: Similaridade Lattes-Qualis para anais de eventos

	Levenshtein		**Smith-Waterman**		**Jaro**		**Cosine Similarity**	
Item	s_{rel}	s_{irrel}	s_{rel}	s_{irrel}	s_{rel}	s_{irrel}	s_{rel}	s_{irrel}
1	0.6461	0.6111	0.8703	0.8500	0.8154	0.8080	0.6187	0.5345
2	0.3461	0.6923	0.5384	0.7500	0.0	0.7794	0.4472	0.5000
3	0.3999	0.6999	1.0	0.8000	0.0	0.8962	0.7071	0.5000
4	0.5344	0.5294	0.7375	0.6607	0.7393	0.7765	0.5773	0.4082
5	0.7582	0.5657	0.8684	0.6612	0.7770	0.7556	0.6030	0.4714
6	0.7129	0.6499	0.8625	0.8064	0.7507	0.7464	0.6396	0.6681
7	0.7049	0.7017	0.7599	0.6250	0.7231	0.8016	0.5892	0.5345
8	0.8227	0.5662	0.8618	0.6666	0.8177	0.7276	0.7833	0.5773
9	0.7051	0.6406	0.8125	0.6999	0.7093	0.7258	0.6674	0.4780
10	0.5000	0.4864	0.8593	0.5937	0.7013	0.6908	0.5070	0.4000
11	0.5365	0.5365	0.8135	0.6374	0.7837	0.7774	0.5270	0.5025
12	0.7142	0.5249	1.0	0.9347	0.7505	0.8116	0.8401	0.4082
13	0.6301	0.4545	0.8787	0.8000	0.7547	0.7471	0.5270	0.3535
14	0.7192	0.5520	0.8426	0.6999	0.7387	0.7496	0.7453	0.4811
15	0.6571	0.3448	1.0	0.5000	0.0	0.7002	0.8660	0.2886
16	0.6716	0.4404	0.7500	0.5000	0.7254	0.7131	0.5773	0.5555
17	0.8674	0.6041	0.9662	0.6999	0.7986	0.7508	0.8432	0.5720
18	0.5079	0.4693	0.8750	0.5625	0.7318	0.7191	0.5962	0.4000
19	0.6153	0.4848	0.8874	0.5124	0.7217	0.7578	0.5962	0.4743
20	0.7101	0.5616	0.8275	0.6250	0.7575	0.7638	0.7071	0.6681