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

KELLY HANNEL

Qualificação de Pesquisadores por Área da Ciência da Computação com Base em uma Ontologia de Perfil

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, março de 2008.

CIP – CATALOGAÇÃO NA PUBLICAÇÃO

Hannel, Kelly

Qualificação de Pesquisadores por Área da Ciência da Computação com Base em uma Ontologia de Perfil / Kelly Hannel – Porto Alegre: Programa de Pós-Graduação em Computação, 2008.

98 f.:il.

Dissertação (mestrado) – Universidade Federal do Rio Grande do Sul. Programa de Pós-Graduação em Computação. Porto Alegre, BR – RS, 2008. Orientador: José Valdeni de Lima.

1.Qualidade. 2.Competência de Pesquisadores. 3.Ontologia de Perfil. I. Lima, José Valdeni de. III. Título.

AGRADECIMENTOS

Apesar desses dois anos de mestrado terem passado rápido demais, foram grandes, demorados e sofridos os passos para chegar até aqui. E nessa caminhada existem muitos que merecem um agradecimento especial.

Primeiramente agradecer a Deus, pela vida, saúde e por me dar forças para terminar meu trabalho quando tudo parecia impossível. Aos meus amados pais, Flademir e Zaira pelo apoio incondicional e incansáveis palavras de incentivo. Aos meus irmãos, Laura e Júnior, muito obrigada pelo incentivo, carinho e amor de vocês. A minha avó querida, Dona Iole, obrigada pelo amor, e por me incluir em suas orações. À vocês, peço perdão pelas minhas ausências e obrigada por entenderem que foi por uma boa causa.

Ao meu amor Lúcio, fico sem palavras para agradecer todo o apoio em mais essa etapa da minha vida. Obrigada pelo teu amor, por acreditar em mim e também por me agüentar!

Ao meu orientador José Valdeni, obrigada pela confiança, paciência e conselhos.

À todos os professores do Instituto de Informática pelo conhecimento que me passaram ao longo desta etapa e pela dedicação com que ensinam. Também aos funcionários do II- UFRGS.

À Capes pela bolsa de estudos.

Aos colegas do grupo de pesquisa: Adriana Kampff, Carlos Morais, Rodrigo Rech, Tiago Telecken, Elmário Dutra, Dóris Reitz, Andrea Krob e Leonardo Daronco. E aos bolsistas de Iniciação Científica, Guilherme Haag Riback, Marcelo B. Anton e Maurício J. R. da Silva. Foi muito bom trabalhar com vocês!

Aos amigos Mário L. M. Machado, Luis H. G. Oliveira, Gabriel Simões, e às amigas que participaram destes 2 anos de mestrado, Renata Zanella e Rúbia Denardi e em especial a Mariusa Warpechowski que foi amiga, conselheira, parceria do chimarrão, das caminhadas, das teorias sobre ontologias e me fez acreditar que eu tinha uma dissertação. Vou levar vocês sempre no meu coração!

SUMÁRIO

LISTA DE ABREVIATURAS E SIGLAS

CNPq	Conselho Nacional de Desenvolvimento Científico e Tecnológico
CAPES	Coordenação de Aperfeiçoamento de Pessoal de Nível Superior
DEA	Análise por Envoltória de Dados, do inglês *Data Development Analysis*
C&T	Ciência e Tecnologia
Mo-DROP	Modelo para Determinação da Relevância da Opinião
RR	*Recommender's Rank*
CC	Coeficiente de competência
CC_c	Coeficiente de competência considerando indicadores do currículo
CC_b	Coeficiente de competência considerando indicadores da produção bibliográfica
CV	Curriculum Vitae
ISBN	*International Standard Book Number*
ISSN	*International Standard Serial Number*
UFRGS	Universidade Federal do Rio Grande do Sul
URL	*Universal Resource Locator*
XHTML	*eXtensible Hypertext Markup Language*
XML	*eXtensible Markup Language*
IEEE	*Institute of Electrical and Electronics Engineers*
MMS	Serviço de *MatchMaking*
OWL	*Ontology Web Language*
FOAF	*Friend Of A Friend*
RDF	*Resource Description Framework*
FOL	*First Logic Order*
W3C	*World Wide Web Consortium*
HTML	*HyperText Markup Language*
KIF	*Knowledge Interchange Format*
URI	*Uniform Resource Identifier*

SUMO	*Standard Upper Merged Ontology*
SHOE	*Simple HTML Ontology Extension*
DAML	*DARPA Agent Markup Language*
OIL	*Ontology Inference Layer*
DARPA	*Defence Advanced Research Projects Agency*
ISO	*International Organization for Standandariztion*
ACM	*Association for Computing Machinery*
PAL	*Protégé Axiomatic Language*
API	*Application Programming Interface*
XHTML	*eXtensible Hypertext Markup Language*
XPath	*XML Path Language*
XQuery	*XML Query Language*
XSLT	*eXtensible Stylesheet Language Transformations*
DOM	*Document Object Model*
SAX	*Simple API for XML*
MAUT	*Multi-Attribute Utility Theory*
RDQL	A *Query Language for RDF*
MDPREF	*Muldimensional Analysis of Preference Data*
SMART	*Simple Multi Attribute Rating Technique*
AHP	*Analytic Hierarchy Process*
CAH	*Agglomerative Hierarchical Clustering*

LISTA DE FIGURAS

LISTA DE TABELAS

RESUMO

A qualidade, tanto da produção científica quanto dos pesquisadores, tem sido foco de discussões e objeto de estudo, isto porque a busca pela excelência é constante no meio acadêmico. Sendo assim, conhecer e medir de forma sistematizada as competências dos pesquisadores constitui-se em uma importante ferramenta para identificar as melhores organizações e indivíduos em uma determinada área.

Esta dissertação buscou descobrir a qualificação dos pesquisadores nas áreas da Ciência da Computação. Para tal, foi desenvolvido um sistema Web (semi) automatizado. Este sistema é centrado na ontologia OntoResearcher, considera o reuso de outras ontologias, a extração de informações da Web e do currículo dos pesquisadores. A OntoResearcher foi modelada com características e indicadores de qualidade (quantitativos e qualitativos) que permitem mensurar as competências dos pesquisadores. O sistema desenvolvido utiliza as informações modeladas na OntoResearcher para automatizar o processo de avaliação dos pesquisadores e tem como diferencial a qualificação distribuída nas áreas da Ciência da Computação em que o pesquisador atua.

As principais contribuições desta dissertação são a definição do perfil de pesquisador, o desenvolvimento da ontologia OntoResearcher e a implementação do sistema de qualificação demonstrando a viabilidade das idéias propostas através dos testes realizados.

Palavras-Chave: Qualidade, qualificação dos pesquisadores, ontologia de perfil.

Researchers' Qualification by Computer Science Area Based on a Profile Ontology

ABSTRACT

The search for excellence is continuous in the academic field. So, the quality of scientific production and researchers has been focus of discussions and subject of study in the academic field. Thus, knowing and measuring the researcher's skills or qualifications in a systematized way is an important tool to identify the best organizations and individuals in a certain discipline.

This work aimed to discover the researcher's qualification of Computer Science field. To accomplish this task, it was developed a Web system (semi) automatized. This system, which is centered on the OntoResearcher ontology, considers the ontology reuse, the information's extraction by the researcher's resume and by the Web. The OntoResearcher was modeled with indicators of scientific quality (quantitative and qualitative) which allows measuring the researcher's qualifications. The developed system uses the information from OntoResearcher to automatize the researcher's evaluation. The main differential of this work is the researcher's qualification distributed in the Computer Science fields on which the researcher has worked.

The main contributions of this work are: the researchers' profiles, the development of OntoResearcher and the development of qualification system demonstrating viability of the ideas through the experimentation.

Keywords: Quality, research qualification, profile ontology.

1 INTRODUÇÃO

A avaliação da atividade científica é uma prática comum na gestão de Ciência e Tecnologia (NIEDERAUER, 2002). Avaliar tanto a pesquisa acadêmica quanto industrial através de seus membros pode ajudar a identificar as melhores organizações e indivíduos em uma dada disciplina. Entretanto, avaliar pesquisadores é um processo complexo, pois envolve diversas variáveis, que muitas vezes são subjetivas e causam contestações (REN e TAYLOR, 2007).

No Brasil, a produção bibliográfica tem sido vista como a parte visível da atividade científica. "O conceito chave do processo é a qualidade e seu instrumento de legalidade, a quantificação. A avaliação, nesses moldes, é usada como instrumento de tomada de decisão e justificativa racional e objetiva na administração de recursos destinados à pesquisa." (GUIMARÃES, 1992, p. 15). A publicação dos resultados de uma pesquisa para o pesquisador visa: divulgar e disseminar suas descobertas científicas, proteger sua propriedade intelectual, obter o reconhecimento pelos pares e o reconhecimento da comunidade científica. Existem várias formas de divulgar o conhecimento científico. Velho (1997) levantou algumas evidências empíricas com relação à escolha dos canais de comunicação, como: à forma da publicação, o idioma e à localização geográfica das publicações para a veiculação dos resultados de pesquisa nas diversas áreas de conhecimento. Em especial, na área das ciências exatas e naturais, Velho afirma que os pesquisadores publicam muito em inglês e em revistas internacionais e que a pressão para garantir a prioridade da descoberta é um fator estimulador à corrida para a publicação.

Assim, a criação tanto de métricas quanto de instrumentos (que levem em consideração não apenas as publicações) que automatizem a descoberta das qualificações dos protagonistas da atividade científica são necessários. Como a qualidade de um pesquisador depende intrinsecamente de sua produtividade e de suas atividades acadêmicas é necessário obter essas informações sobre o pesquisador, e para isso utiliza-se a Web. A rede mundial de computadores compartilha um espaço virtual repleto de dados e informações. Tem-se muita informação disponível, entretanto, a estrutura na qual as informações são apresentadas não permite deduzir significados necessários para o processamento por aplicações computacionais. As informações disponíveis na Web não possuem semântica explícita, o seu significado é extraído por inferências, baseadas em conhecimento prévio, realizadas por pessoas. Por exemplo, uma máquina de busca consegue recuperar informações sobre pesquisadores, porém, não fornece o significado de tal informação, não sendo possível processar automaticamente estes dados para obter a competência ou a qualificação do pesquisador em determinada área. É necessário compreender e extrair as informações da Web, normalmente com apoio humano, para que seja possível processá-las e então obter o resultado.

Nesse contexto surgiu uma nova alternativa: a Web Semântica. Segundo Berners-Lee et al. (2001) os computadores necessitam ter acesso a dados e metadados e também precisam de conjuntos de regras de inferência que ajudem no processo de dedução automática para efetuar um raciocínio automatizado. As regras de inferência, além dos metadados, são especificadas através de ontologias. Com o uso de ontologias é possível elaborar uma rede de conhecimento humano em formato processável automaticamente, complementando o processamento da máquina e melhorando a qualidade dos serviços na Web. É no contexto da Web Semântica e da necessidade de certificar as competências dos pesquisadores que se desenvolveu uma abordagem ontológica para a definição do perfil do pesquisador. Salienta-se que, os termos "competência" e "qualificação" são usados neste trabalho como sinônimos, significando o quão apta uma pessoa é para desenvolver determinada tarefa, ou atuar em determinada área

A ontologia desenvolvida, denominada OntoResearcher, descreve o perfil acadêmico de pesquisadores da área da Ciência da Computação. Perfil, no contexto desta dissertação, é considerado um conjunto de características e indicadores de qualidade que identificam pesquisadores da Ciência da Computação. As características são informações consideradas pessoais. Já os indicadores de qualidade podem ser quantitativos ou qualitativos (ou de impacto).

Para o desenvolvimento da OntoResearcher, primeiramente foi definido um modelo com os indicadores que permitem obter a qualificação dos pesquisadores. Esse modelo foi definido através da análise de duas fontes de informações: a) o currículo Lattes do CNPq[1]; b) o documento do CNPq[2] que especifica os critérios para conceder bolsa de produtividade aos pesquisadores. A OntoResearcher é a base do protótipo desenvolvido para descobrir a qualificação dos pesquisadores por área da Ciência da Computação.

O modelo implementado envolve a obtenção das informações sobre os pesquisadores a partir de diferentes fontes, como: o Google Scholar[3], as áreas da CC definidas pela ACM[4] (*Association for Computing Machinery*), o currículo Lattes dos pesquisadores, e as ontologias OntoQualis (SOUTO et. al, 2007) e OntoDoc[5] (estas 2 ontologias, não foram desenvolvidas nesta dissertação).

A descoberta da qualificação do pesquisador é realizada no contexto do projeto DIGITEX (OLIVEIRA et al., 2005). O Projeto Digitex envolve a geração, indexação e busca personalizada de conteúdos digitais e tem por objetivo auxiliar no processo de criação e aperfeiçoamento de conhecimento através da revisão pelos pares e também indicar ou receber indicação de conhecimento relevante.

A flexibilidade do protótipo desenvolvido permite que possa ser utilizada em sistemas de recomendação de artigos acadêmicos; para criação de comunidades virtuais; em processos de seleção que necessitem saber quais pesquisadores são especialistas em

---

[1] http://lattes.cnpq.br/index.htm

[2] http://portal.cnpq.br/cas/ca-fr.htm

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

[4] http://portal.acm.org/ccs.cfm?part=author&coll=GUIDE&dl=GUIDE&CFID=22209137&CFTOKEN=19512755

[5] Dissertação de mestrado, em desenvolvimento no projeto DIGITEX, de Luis Henrique Gonçalves Oliveira sob orientação do Prof. Dr. José M. Palazzo de Oliveira.

determinada área; em disputas por recursos; na criação de *rankings* de pesquisadores utilizando diferentes critérios; sites comerciais para recomendação de livros na área dos pesquisadores, por exemplo.

1.1 Motivação e Definição do Problema

O Projeto DIGITEX (OLIVEIRA et al., 2005) sugere a criação de um sistema para editoração colaborativa de artigos científicos, com revisão interativa pelos pares de revisores do sistema, discussão dos artigos aberta à comunidade e processo de avaliação aberta. A Figura 1.1 apresenta a arquitetura para revisão aberta de documentos, a) representa a etapa de submissão do documento pelo autor. Caso o autor possua uma pontuação maior ou igual que um limiar (V>=a), esse documento passa para a etapa de avaliação aberta pelo público (representado por b); senão é necessário selecionar os revisores do sistema para avaliar o documento.

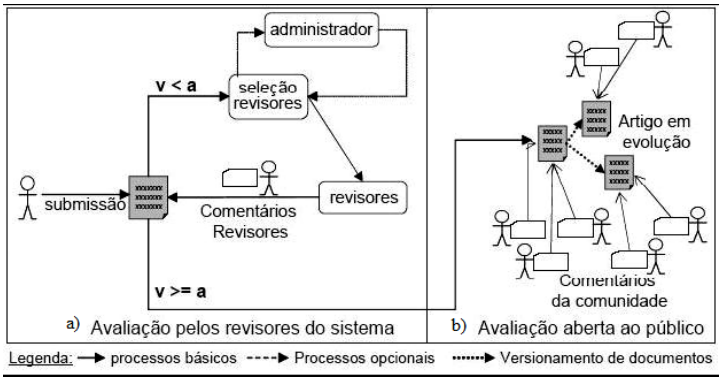

Figura 1.1: Arquitetura para revisão aberta de documentos (Modificado de OLIVEIRA et al., 2005).

Neste processo, identificar as competências acadêmicas dos membros do sistema, principalmente autores e revisores, em relação às diversas áreas da Ciência da Computação em que podem opinar, é imprescindível para o processo de avaliação aberta.

Nesse contexto, o principal problema deste trabalho é desenvolver um sistema capaz de encontrar e armazenar as informações acadêmicas dos pesquisadores, para que através destas informações seja possível definir as qualificações (competências) dos mesmos nas áreas da Ciência da Computação.

1.2 Detalhamento do Problema

A definição do problema pode ser melhor descrita com a decomposição em subproblemas, como segue:

1. É necessário identificar quais os indicadores que permitem qualificar os pesquisadores;
2. Definir quais as fontes de informações podem ser utilizadas para obter as informações necessárias;
3. Modelar o perfil do pesquisador;
4. Descobrir informações implícitas sobre o perfil;

5. Identificar formas de pontuar os indicadores de qualidade de forma diferenciada, isto porque, alguns indicadores são considerados mais importantes que outros;
6. Calcular a qualificação dos pesquisadores nas áreas em que o mesmo atua dentro da Ciência da Computação.

1.3 Objetivos da Dissertação

Com base no problema identificado, foram delimitados alguns objetivos para este trabalho, como:

- Definir os indicadores que permitam indicar o percentual de dedicação dos pesquisadores;
- Definir um perfil de pesquisador adequado para o problema identificado e descrever esse perfil em uma linguagem computacional que permita a descoberta de novos conhecimentos a cerca dos pesquisadores;
- Calcular a qualificação dos pesquisadores por área da Ciência da Computação;
- Exportar este perfil para o projeto DIGITEX;
- Desenvolver a qualificação do pesquisador em uma arquitetura flexível que permita que sistemas de recomendação possam utilizar os perfis descritos para melhorar o processo de recomendação.

1.4 Organização dos Capítulos

O restante desta dissertação está organizado da seguinte maneira: o capítulo dois aborda os trabalhos relacionados, sendo que estes foram subdivididos em duas subseções que tratam dos trabalhos relacionados à avaliação da pesquisa científica e os trabalhos que utilizam ontologias de forma similar à proposta nesta dissertação.

No capítulo três é apresentado um levantamento bibliográfico sobre ontologias bem como o seu uso para a criação de perfis. Além disso, são mostradas as classificações, os principais componentes, as principais linguagens e as ferramentas para o desenvolvimento das ontologias.

O capítulo quatro apresenta a arquitetura e principais funcionalidades do sistema, a OntoResearcher bem como as implementações realizadas e as ferramentas utilizadas nesta dissertação.

O capítulo cinco apresenta os experimentos realizados para descobrir a qualificação dos pesquisadores e os resultados obtidos. Também amostra as consultas realizadas na OntoResearcher para a descoberta de conhecimento e a criação de conglomerados. Já o sexto capítulo apresenta as conclusões e os trabalhos futuros que foram identificados ao longo desta dissertação.

2 TRABALHOS RELACIONADOS

Os trabalhos relacionados foram subdivididos em duas seções. A seção 2.1 trata dos trabalhos sobre análise e formas de avaliação da pesquisa e dos pesquisadores. Já a seção 2.2 apresenta trabalhos que utilizam ontologias de perfil. Não foram encontrados trabalhos sobre avaliação de pesquisadores que utilizassem uma abordagem ontológica dos perfis, por este motivo optou-se por realizar a divisão dos trabalhos correlatos.

2.1 Quanto à pesquisa acadêmica

Nesta seção encontram-se os trabalhos que abordam problemas de como medir, quantificar, avaliar, pontuar ou mesmo qualificar, tanto a pesquisa científica como os pesquisadores. São diferentes abordagens, focadas em aspectos particulares que foram analisadas e comparadas nesta dissertação.

2.1.1 Análise de Eficiência da Pesquisa Acadêmica

É uma abordagem sistemática para analisar a performance da pesquisa acadêmica em universidades e institutos de pesquisa. A análise é baseada em um conjunto de critérios. Nem todos os critérios são quantitativos, sendo que, os indicadores qualitativos são quantificados usando ferramentas analíticas apropriadas. A abordagem consiste dos seguintes passos (KORHONE et al., 2002): definição dos critérios e indicadores que serão utilizados para medir a performance da pesquisa; colecionar os dados das unidades de pesquisa apropriadas; e usando "*Data Development Analysis - DEA"* (ou Análise por envoltória dos Dados em português), calcular o valor da eficiência para cada unidade.

Os autores partem de um modelo de unidade de pesquisa. Uma unidade de pesquisa ideal é aquela em que seus membros continuamente produzem com alta qualidade, pesquisa inovadora e internacionalmente reconhecida. Além disso, os pesquisadores orientam alunos de doutorado e atuam ativamente na comunidade científica. Os conjuntos de critérios que servem para caracterizar o modelo de unidade de pesquisa são:

- Qualidade da pesquisa:
    1. Artigos em *journals* internacionais;
    2. Livros e capítulos de livros publicados internacionalmente;
    3. Citações.
- Atividades de pesquisa:
    1. Publicações com um mínimo padrão de qualidade (artigos em *journals* referenciados);
    2. *Paper* em *proceeding*;

3. Apresentações em conferências;

- Impacto da pesquisa;
    1. Citações de outros pesquisadores;
    2. Apresentações em conferências internacionais como pesquisador convidado;
    3. Número de co-autores estrangeiros nos artigos publicados;
- Atividade de orientação;
    1. Orientações de doutorado concluídas;
    2. Número de estudantes de doutorado orientados;
- Atividades na comunidade científica (não usado atualmente):
    1. Editor de livro;
    2. Organizador de conferências científicas, membro de comitê de programa.

No caso de todos os identificadores serem quantitativos o problema está em identificar uma função que agregue valores aos indicadores dentro de uma escala de critérios. Quando alguns indicadores são qualitativos, primeiramente é necessário quantificá-los usando ferramentas adequadas. Uma maneira bastante usada para agregar valores é usar somas ponderadas.

2.1.2 Sistema ETHOS

O sistema Ethos (NIEDERAUER, 2002) mede a produtividade relativa de pesquisadores candidatos à Bolsa de Produtividade em Pesquisa do CNPq, baseando-se nos dados do currículo Lattes dos pesquisadores, deixando a cargo do próprio pesquisador a escolha dos indicadores de ciência e tecnologia (C&T) com os quais deseja ser comparado.

Para o julgamento das Bolsas de Produtividade em Pesquisa o CNPq possui critérios definidos em norma específica. Além disso, cada área de conhecimento possui um Comitê de Acessoramento que é responsável pela definição dos seus critérios de julgamento específicos. De fato, os comitês definem uma espécie de "pesquisador padrão", ao determinar quais os indicadores de C&T que consideram relevantes e, em alguns casos, dando pesos a cada um deles (NIEDERAUER, 2002). Os indicadores utilizados na construção do sistema Ethos, são:

- Produção Bibliográfica: artigos completos publicados em periódicos nacionais; artigos completos publicados em periódicos estrangeiros; trabalhos completos publicados em eventos nacionais; trabalhos completos publicados em eventos estrangeiros; livros publicados e capítulos de livros publicados;
- Produção Técnica: processos tecnológicos, com ou sem registro ou patente; produtos tecnológicos, com ou sem registro ou patente; software, com ou sem registro ou patente e trabalhos técnicos realizados;
- Indicadores de Formação de Recursos Humanos: dissertações de mestrado orientadas e concluídas, e teses de doutorado orientadas e concluídas;

Outros indicadores, como orientação de trabalhos de graduação e especialização foram descartados. Como o sistema foi desenvolvido para o julgamento de Bolsas de Produtividade, apenas os pesquisadores doutores são analisados.

O método utilizado para medir a produtividade dos pesquisadores é a Análise por Envoltória dos Dados. A medição é quantitativa, pois o modelo proposto somente considera dados quantitativos referentes à produção científica, tecnológica e técnica dos

pesquisadores. O sistema Ethos não promove nenhuma avaliação sobre o projeto de pesquisa que os candidatos à Bolsa de Produtividade em Pesquisa apresentam ao CNPq.

2.1.3 Relevância de Opinião de Usuários

A relevância de opinião de um usuário pode ser deduzida de suas competências (CAZELLA, 2006). 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.

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.

O sistema calcula o RR (valor entre 0 – nenhum *expertise*, e 10 – *expertise* máximo) para cada área de interesse do usuário, conforme a Equação 1.

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

O autor 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 experimento. Neste, 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 atributos selecionados por Cazella estão descritos no Anexo A.

2.1.4 Mineração de Competências para Criação de Comunidades Virtuais

Rodrigues e Oliveira (2004) propuseram uma técnica para criar/sugerir comunidades científicas baseadas nas competências dos cientistas. As competências são identificadas usando as publicações científicas e considerando que uma possível indicação para a participação de alguém em uma comunidade depende de seu conhecimento publicado e grau de *expertise*[6]. Os autores afirmam que comunidades têm o principal propósito de aquisição, troca e disseminação de conhecimento científico em certo domínio de pesquisa, encorajando a colaboração científica (RODRIGUES; OLIVEIRA, 2004).

Para mapear a competência dos pesquisadores através da análise de suas publicações, os autores utilizam métodos e técnicas de descoberta de conhecimento em textos (*text mining*). Cada pesquisador possui um diretório onde todas suas publicações são armazenadas. Para cada publicação é atribuído uma chave de identificação (*identification key*). Essa chave é armazenada no diretório do pesquisador correspondente, evitando que seja necessário rodar o algoritmo novamente para este mesmo texto.

[6] Medida da relevância de uma palavra, segundo Rodrigues e Oliveira (2004).

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 os radicais.

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* são armazenados no banco de dados.

Após a fase de identificação de competências o sistema procura se existem comunidades sobre determinado tópico, caso não, ele procura pessoas com interesses similares e propõem a criação de uma comunidade sobre este tópico. Além disso, sugere a participação de pessoas em comunidades que casam com o seu perfil.

2.1.5 Modelo de Pontuação na Busca de Competências Acadêmicas

Rech (2007) descreve um modelo para descobrir e pontuar competências acadêmicas de pesquisadores, baseado na combinação de 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. São eles: 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 outros;
2. Indicadores quantitativos relacionados à produção bibliográfica do pesquisador: mensuram aspectos como o impacto ou repercussão dos trabalhos (através do número de citações), bem como a qualidade e alcance dos veículos de publicação nos quais o pesquisador possui trabalhos publicados.

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 nenhuma competência do pesquisador e o valor 10 indica "competência máxima" entre os pesquisadores avaliados. 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.

Os indicadores de produção foram obtidos da Plataforma Lattes do CNPq. Já os indicadores quantitativos relacionados à produção bibliográfica foram obtidos a partir de duas bases de informações distintas. Uma para classificação dos veículos de publicação (Qualis-Capes[7]) e outra para capturar o impacto e repercussão dos trabalhos do pesquisador na comunidade científica (Google Scholar).

A entrada do sistema consiste do arquivo XML (*eXtensible Markup Language*) do Currículo Lattes do pesquisador. O sistema extrai os indicadores quantitativos relacionados ao currículo, bem como diversas informações necessárias aos demais processamentos da aplicação. Efetua consultas ao site Google Scholar para coleta das

[7] http://servicos.capes.gov.br/webqualis

referências contendo o número de citações dos trabalhos do pesquisador. Aplica técnicas de similaridade para verificar a similaridade entre as informações do Lattes e do Scholar (para 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 entre as informações do Lattes e do Qualis (serve para verificar o nível dos veículos de publicação nos quais o pesquisador possui trabalhos publicados). Para o cálculo das pontuações o autor implementa a normalização dos indicadores e posterior cálculo dos coeficientes de competência.

Os indicadores utilizados pelo autor foram os mesmo apresentados por Cazella (2006), descritos no Anexo A, com a inclusão de indicadores quantitativos relacionados com a importância da produção bibliográfica dos pesquisadores, também descritos no Anexo A.

2.1.6 Identificação Automática de *Expertise*

Borges et al. (2004) desenvolveram uma ferramenta para identificação automática da *expertise*, baseada na extração de informações do currículo XML gerado pela Plataforma Lattes. Os perfis são armazenados em uma base de dados que mantém informações sobre os membros da comunidade que utilizam o sistema. Estas informações correspondem a dados cadastrais (nome, e-mail, etc.) e também a informações sobre o seu grau de interesse e/ou conhecimento em determinados assuntos, representados por conceitos presentes na ontologia. A base de perfis do sistema é construída visando aprimorar o processo de recomendação, apresentando para o usuário somente itens que forem do seu interesse e adequados ao seu nível de conhecimento.

Utilizando técnicas de mineração de textos comparam as palavras extraídas com os conceitos presentes em uma ontologia de domínio da área da Ciência da Computação. A ontologia foi implementada como uma estrutura hierárquica, contendo um conjunto de conceitos. Cada conceito possui associado a si uma lista de termos e seus respectivos pesos, que ajudam a identificar o conceito presente nos textos. A abordagem representa currículos e conceitos através de vetores de termos e 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 abordagem 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-chave, áreas de atuação, entre outros).

2.1.7 h-Index

Hirsch (2007) propõe uma métrica chamada *h-Index*, definida como o número de artigos com número de citações maior ou igual que *h*. Ela pode ser entendida como segue: 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.

2.1.8 Considerações

O problema de identificar as competências dos pesquisadores é tratado em diferentes abordagens. E, por ser um processo subjetivo, dependente do ponto de vista de quem julga e do objetivo do julgamento (se é em disputa por recursos ou alocação de vagas, por exemplo) freqüentemente é um processo incompleto. Por esta razão, quanto menos indicadores de qualidade um processo de identificação de competências possui, menos confiável ele é considerado.

Segundo Parnas (2007) está se alastrando a prática de medir os pesquisadores pelo número de publicações, sem ao menos ler e julgar tais publicações, o que encoraja práticas como: pesquisas superficiais, alguns grupos de pesquisa colocam como autores pessoas que não participaram da construção do artigo, repetição de artigos, além de estudos insignificantes e mal planejados. Algumas avaliações agregam indicadores como o número de citações, mas estas também podem ser questionadas, pois muitas citações são incluídas nos artigos apenas para mostrar que os autores conhecem a literatura e o baixo número de citações pode significar que o autor não é conhecido na comunidade acadêmica e não que o trabalho é ruim e por isso pouco citado. Além deste problema, há também o das auto-citações, ou amigos que se citam (KORHONE et al., 2002), o que indica que apenas o critério do número de citações não é confiável.

Nesta dissertação, foram encontradas diferentes abordagem para medir a competência dos pesquisadores. Esta seção apresenta as principais diferenças entre a abordagem proposta nesta dissertação e as abordagens descritas ao longo da seção 2.1. São elas:

- A abordagem apresentada por Korhone et al. (2002) foi desenvolvida para análise de unidades de pesquisa e não de pesquisadores individualmente. A caracterização das unidades de pesquisa é semelhante à adotada nesta dissertação, entretanto possui menos critérios e não trabalha com perfis de pesquisadores.
- O sistema Ethos (NIEDERAUER, 2002) utiliza algumas informações do currículo Lattes como fonte de dados, entretanto não utiliza o número de citações das publicações e nem dados qualitativos como o Qualis dos veículos de publicação. Além disso, o processo de avaliação é focado em pesquisadores doutores, candidatos a bolsa de produtividade em pesquisa do CNPq e não trabalha com a modelagem dos pesquisadores, ao contrário do trabalho aqui proposto.
- O trabalho de Cazella (2006) se difere do proposto nesta dissertação por não incluir dados como repercussão das publicações na comunidade científica e nem o nível dos veículos de publicação. Além disso, solicita informações explicitamente ao usuário, enquanto neste trabalho as informações são obtidas do Lattes e de outras fontes automaticamente. O trabalho de Cazella não objetiva encontrar as áreas de atuação dos pesquisadores, esta informação é solicitada ao pesquisador e é identificada como área de interesse (o que significa que o pesquisador tem interesse na área, entretanto, não identifica se ele efetivamente atua nela). O experimento desenvolvido por Cazella, no qual pesquisadores identificaram e ponderaram os indicadores de produção acadêmica que consideravam importantes para a definição da relevância de opinião de um pesquisador, é utilizado nesta dissertação como base para a definição dos pesos dos indicadores.

- No trabalho de Rodrigues e Oliveira (2004) as competências são identificadas usando apenas as publicações e considerando que uma possível indicação para a participação de alguém em uma comunidade depende de seu conhecimento publicado e grau de *expertise*, sendo que a descoberta das áreas de interesse é feita através de técnicas de *text mining*. A técnica de descoberta da área de interesse poderia ser utilizada na presente dissertação, entretanto, somente são analisadas as publicações e não foi desenvolvida com o intuito de avaliar as competências dos pesquisadores.
- O trabalho apresentado nesta dissertação inclui alguns dos indicadores apresentados por Rech (2007) e também utiliza como dados de entrada as informações do XML do Currículo Lattes dos pesquisadores. Entretanto, o trabalho de Rech não modela o perfil dos pesquisadores nem distingue as áreas de atuação dos mesmos. Além disso, na abordagem de Rech os pesquisadores têm seu "índice de competência" normalizado em relação aos demais pesquisadores do sistema. O trabalho de Rech utiliza os valores dos pesos definidos em Cazella (2006) e define os pesos dos indicadores qualitativos (os quais não foram utilizados por Cazella). Os pesos dos indicadores qualitativos, definidos por Rech e por Cazella, serão utilizados como base para a definição dos pesos utilizados nesta dissertação.
- A abordagem de Borges et al. (2004) refinada por Ribeiro Junior et al. (2005) é adequada no processo de descoberta de áreas de experiência ou de interesse. Os autores utilizam uma ontologia, também baseada nas áreas da ACM, para a definição dos termos que permitem identificar as áreas. Entretanto, os autores identificam as áreas de atuação dos pesquisadores levando em conta apenas as suas publicações, enquanto a presente dissertação utiliza vários outros indicadores, tais como: qualidade e repercussão científica, dentre outros.
- O *h-Index* dá uma noção restrita de competência, pois considera apenas o número de citações das publicações de um pesquisador, desprezando outros indicadores de qualidade. Este modelo, apesar de apresentar resultados interessantes é reducionista. Por ter essa restrição, considera-se que esta abordagem não é adequada em um processo de qualificação abrangente como o proposto nesta dissertação, ele poderia ser utilizado como um indicador de qualidade. Porém, como nenhuma das fontes de informações analisadas (currículo Lattes, os critérios do CNPq para conceder a bolsa de produtividade em pesquisa) considera o h-Index, o mesmo não foi utilizado como indicador.

2.2 Quanto ao Uso de Ontologias

Esta seção apresenta os trabalhos que utilizam ontologias em abordagens similares à apresentada nessa dissertação. A maioria dos trabalhos utiliza ontologias para descrição de perfis (mesmo que com objetivos diferentes), exceto no Mesur (RODRIGUEZ et al., 2007) onde a ontologia representa as relações existentes em uma comunidade acadêmica.

2.2.1 MMS

O serviço de *matchmaking* (MMS) foi desenvolvido para auxiliar encontros entre pessoas portadoras de dispositivos móveis que estejam geograficamente próximas e que tenham perfis de interesses similares (MACHADO, 2005). Os perfis dos usuários são descritos como ontologias na linguagem OWL (*Ontology Web Language*). O usuário deve fornecer seus níveis de interesse, e estes são utilizados como dados de entrada em um algoritmo para avaliação de correlação linear, chamado Pearson R, que calcula a similaridade entre estes níveis de interesses para cada localização.

A modelagem dos perfis de usuários foi desenvolvida usando a estratégia Top-Down onde os conceitos e suas relações foram identificados partindo-se dos cenários de uso e dos requisitos. A idéia principal dessa ontologia é que os usuários disponibilizem interesses sobre os quais estariam dispostos a interagir com outros usuários. Os interesses podem depender da localização do usuário, ou seja, existem locais em que um usuário deseja e outros em que ele não deseja interagir. Da mesma forma, podem existir usuários que queiram informar níveis de interesses diferenciados para cada localização. Por exemplo, um usuário pode indicar que seu interesse por futebol é mais relevante (alto) em um local de lazer do que num local de trabalho.

Os principais conceitos da ontologia são: pessoa (*Person*), interesse (*Interest*), nível de interesse (*InterestLevel*) e localização (*Location*). Esses e outros conceitos são descritos na Tabela 2.1.

Tabela 2.1: Conceitos presentes na ontologia do MMS (MACHADO, 2005).

Conceito	Comentário
InstantMessenger	Classe que instancia relacionamentos (identificação de usuários) entre instâncias *IMSoftware* e *Person*.
IMSoftware	Classe que instancia softwares de *Instant Messenger*.
Interest	Classe que instancia interesses de usuários.
InterestLevel	Classe que relaciona um interesse a um valor de interesse em uma localização, mensurado por uma escala e com restrições de tempo.
KnowList	Classe que instancia uma lista de amigos de um usuário proveniente de interações passadas.
Language	Classe que instancia tipos de idiomas.
Proficiency	Classe que instancia relacionamentos entre graus de proficiência de usuários em um idioma
Location	Classe que instancia regiões simbólicas.
Person	Classe que instancia dados sobre os usuários do serviço.
Scale	Classe que instancia escalas que irão mensurar níveis de interesses.

Uma idéia que pode ser reusada do trabalho de Machado é a utilização do algoritmo de *matching* para encontrar pessoas com perfis similares e a partir disso seria possível fazer que pesquisadores interajam, ou mesmo recomendar a criação de comunidades virtuais.

2.2.2 FOAF

O projeto FOAF- Friend of a Friend- iniciou em 1999, com o objetivo de criar uma Web com sites que sejam inteligíveis por máquinas e que descrevam pessoas, *link*s entre indivíduos e coisas que eles podem criar e fazer. A partir de 2003, com o desenvolvimento da Web Semântica, FOAF começou a ser notado. A ontologia FOAF foi descrita utilizando OWL (o código OWL pode ser obtido em: www.mindwasp.org/2003/owl/foaf). Qualquer pessoa pode gerar seu perfil FOAF, publicá-lo na Web e o seu perfil pode ser adicionado à ontologia (LUGANO, 2005).

O vocabulário FOAF possui 12 classes e 52 propriedades. As classes são:

- *Agent*: pessoa, grupo, software;
- *Document*: um documento físico ou eletrônico;
- *Group*: uma classe de agentes;
- *Image*: uma imagem; essa classe é subclasse de *Document*;
- *OnlineAccount*: uma conta *online*, correspondendo a previsão de um serviço;
- *OnlineChatAccount*: uma conta de *chat* (*online*);
- *OnlineEcommerceAccount*: uma conta *online* de *e-commerce*;
- *OnlineGamingAccount*: uma conta de jogos *online*;
- *Organization*: uma organização, correspondente a instituições sociais, companhias, associações;
- *Person*: Classe que representa pessoas (vivas, mortas, reais ou imaginárias). É subclasse de *Agent*;
- P*ersonalProfileDocument*: descreve propriedades da pessoa que é autor de um documento;
- *Project*: classe de entidades que representam um projeto. Pode ser informal, formal, individual ou coletiva.

A descrição das 52 propriedades pode ser encontrada em Lugano (2005). A Figura 2.1 apresenta o exemplo de construção de um bloco FOAF para descrever uma pessoa.

```
<rdf:RDF
      xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
      xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
      xmlns:foaf="http://xmlns.com/foaf/0.1/">
<foaf:Person>
      <foaf:name>Giuseppe Lugano</foaf:name>
      <foaf:mbox rdf:resource="mailto:gl@localhost.org" />
      <foaf:depiction rdf:resource="http://localhost.org/me.jpg"/>
</foaf:Person>
</rdf:RDF>
```

Figura 2.1: Descrição de uma pessoa (LUGANO, 2005).

O exemplo da Figura 2.1 inicia com: “existe uma pessoa chamada Giuseppe Lugano e que tem o endereço de e-mail g@localhost.org e uma foto dele está disponível em http://localhost.org/me.jpg”. Com essa descrição fica simples de perceber que FOAF é simplesmente um vocabulário RDF (*Resource Description Framework*), e que o RDF é usado para codificar as descrições FOAF. Usando RDF, o FOAF ganha um poderoso mecanismo de extensibilidade, permitindo que as descrições baseadas em FOAF sejam adicionadas a qualquer outro vocabulário RDF. Além disso, permite que o FOAF considere apenas um vocabulário específico de uma pessoa sem ter que negociar também com outros conceitos, como dados geográficos. Diferentes vocabulários podem ser usados em conjunto de uma maneira simples: RDF provê um modelo base de objetos, com seus atributos e relacionamentos. Um exemplo de objeto (classe) é, `foaf:Person`, enquanto que `foaf:Knows` e `foaf:Name` são, respectivamente, exemplos de relacionamento e atributo da classe *Person*.

Usando as várias propriedades que o FOAF oferece para expressar informações pessoais, é possível construir redes sociais. O componente mais importante de um documento FOAF é o vocabulário, que é identificado pelo *namespace* URI (*Uniform Resource Identifier*): http://xmlns.com/foaf/0.1.

Desta abordagem, pode-se reusar a idéia de que os sites dos pesquisadores concordem com a OntoResearcher, e assim aplicações computacionais poderiam extrair informações sobre os pesquisadores diretamente de seus sites pessoais.

2.2.3 Foxtrot

Middleton et al. (2004) utilizam uma abordagem de perfis descritos em ontologias para auxiliar no problema de recomendar artigos acadêmicos de forma *online*. Os perfis são criados de forma não obstrusiva através do monitoramento do comportamento e relevância do *feedback*, representando os perfis em forma de artigos acadêmicos nos tópicos da ontologia. O sistema de recomendação desenvolvido- Foxtrot- é híbrido, suportando recomendação colaborativa e baseada em conteúdo, além disso, consulta diretamente a base de dados de artigos.

A ontologia é baseada em uma biblioteca digital que classifica os tópicos da Ciência da Computação com exemplos de artigos para cada tópico. Os relacionamentos entre os tópicos de interesse são usados para inferir novos interesses que não estão explícitos. A Figura 2.2 apresenta a seção de um tópico de pesquisa da ontologia utilizada no Foxtrot.

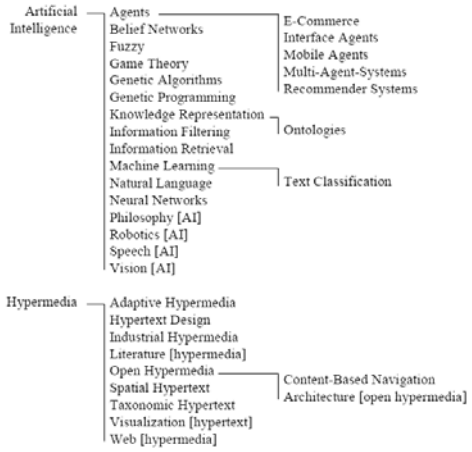

Figura 2.2: Parte da ontologia de tópicos da Ciência da Computação (MIDDLETON et. al, 2004).

Os artigos são representados como vetores de termos com a freqüência de cada termo, o número total de termos usados para o peso dos termos e os termos que representam palavras simples no texto do artigo. A etapa de classificação é feita utilizando um algoritmo que permite treinar exemplos, os quais são adicionados ao vetor de termos e que retornam as vizinhanças mais relevantes. A proximidade de um vetor não classificado com um vetor de termos de sua vizinha é o que determina sua classificação. Com este trabalho, os autores demonstraram que:

- Inferência ontológica pode ser utilizada para incrementar a precisão dos perfis;
- Conhecimento ontológico externo (isto é, informações adicionais sobre os pesquisadores) pode ser empregado para minimizar o problema de *cold-start* dos sistemas de recomendação;
- Visualização gráfica e manipulação dos perfis pelo próprio usuário (*profile feedback*) permitem aumentar a precisão dos perfis.

2.2.4 Mesur

Os autores apresentam uma ontologia construída para representar a comunidade acadêmica, incluindo dados bibliográficos, citações, sendo que os dados usados foram coletados de quem publicou o documento (*publishers*) e repositórios na Web (instanciação na ordem de 50 milhões de artigos e seus objetos, por exemplo: autores e *journal* onde foi publicado). Essa abordagem foca na representação das relações entre quem faz o artigo (chamado de *Agent*, e pode ser tanto uma pessoa quanto uma organização); o artigo em si (chamado de *Document*, e pode ser: artigo, livro, *proceedings*, *journal* e livro editado) e também o contexto (chamado *Context*, serve para interação entre documentos e agentes). A Figura 2.3 mostra as três principais classes da ontologia Mesur em destaque.

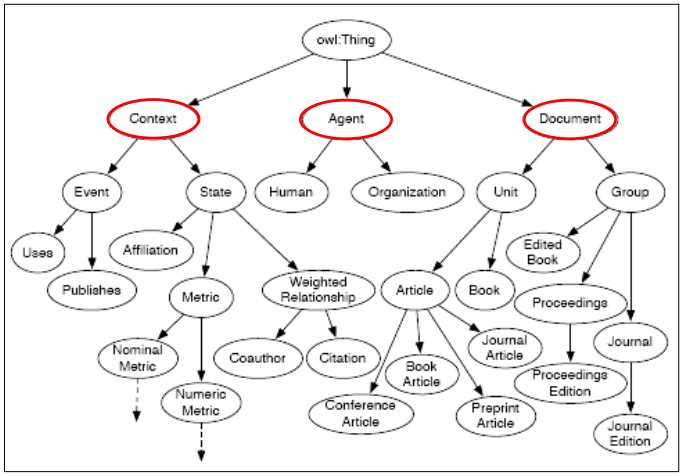

Figura 2.3: Taxonomia da Mesur (RODRIGUEZ et al., 2007).

Um agente (classe *Agent*) pode ser tanto um humano (subclasse *Human*) como uma organização (subclasse *Organization*). Tendo o documento é possível descobrir, por inferência, quem são os autores (propriedade *authored*) e quem publicou o documento (propriedade *published*). A classe sobre o contexto (*Context*) tem duas subclasses, evento (*Event*) e estado (*State*). *Event* é uma medição feita por um provedor em um período de tempo, por exemplo, as subclasses de *Event*, *Publishes* e *Uses* são gravadas pelas propriedades *publishes* e *repositories* no mesmo período de tempo. Já

State é uma medição que pode ocorrer sobre determinado momento, e é usada para representar relacionamentos complexos entre artefatos ou como uma forma de anexar metadados a um objeto.

São realizadas algumas inferências muito interessantes na ontologia Mesur, como: *citation* (infere que agente citou e qual foi citado); *coauthor* (infere quem trabalhou junto, e dá um peso referente ao número de vezes em que trabalharam juntos); *impactFactor* (por exemplo, o do JCDL de 2007 é definido como o número de citações de qualquer artigo publicado em 2007 para artigos publicados nos JCDL de 2005 e 2006 normalizados pelo total de artigos publicados no JCDL de 2005 e 2006).

2.2.5 Considerações

Buscou-se analisar abordagens que pudessem auxiliar na definição do perfil do pesquisador. Os trabalhos correlatos descritos ao longo da seção 2.2 foram resumidos na Tabela 2.2. Nenhuma das abordagens estudadas contempla, no todo, o objetivo de encontrar a qualificação dos pesquisadores por áreas da Ciência da Computação. Por esta razão, optou-se pela descrição de um novo perfil e definição da ontologia OntoResearcher.

Foram estudadas também, abordagens para descrição de perfil de aluno, como a de Chen e Mizoguchi (1999), Dolog (2003) e Musa (2006). Entretanto tais abordagens não foram selecionadas para trabalhos correlatos por serem abordagens focadas em conceitos exclusivos de alunos (estilos de aprendizagem, preferências de idioma, dispositivo, de recursos, etc.).

Tabela 2.2: Resumo dos trabalhos sobre perfil de usuário.

	Perfil	Uso de Ontologias	Competência de Pesquisadores	Principais conceitos da Ontologia
MMS	Perfis armazenam interesses dos usuários.	Identificação de perfis similares.	Não trata esse problema.	*Person, Interest, InterestLevel* e *Location.*
FOAF	Descreve pessoas, links entre indivíduos e coisas que eles podem fazer ou criar.	Usando as propriedades é possível criar redes sociais.	Não trata esse problema.	*Agent; Document; Group; Image; OnlineAccount; OnlineChatAccount; Person; OnlineEcommerceAccount; OnlineGamingAccount; Project; Organization; PersonalProfileDocument;.*
Foxtrot	Os perfis são criados através de monitoramento das ações e relevância do feedback.	Os perfis são representados como artigos acadêmicos nos tópicos da ontologia.	Não trata esse problema.	A ontologia é baseada em uma biblioteca digital que classifica os tópicos da Computação com exemplos de artigos em cada um.
Mesur	Não trata de perfis.	Usada para descrever a comunidade acadêmica.	Aborda as relações acadêmicas e as métricas que permitem avaliar as publicações.	*Context*, *Agent* e *Document*.

3 FUNDAMENTAÇÃO TEÓRICA

Este capítulo visa apresentar os conceitos que nortearam o desenvolvimento da ontologia de perfil, chamada OntoResearcher. São apresentados conceitos de ontologias, algumas linguagens estudadas, as ferramentas para trabalhar com ontologias bem como as razões para a descrição de perfis como ontologias.

3.1 Ontologia

O termo ontologia deriva do grego "onto", ser, e "logia", discurso escrito ou falado. É também, originário da filosofia, onde é usado para representar uma visão do mundo em um sistema de categorias. Segundo Sowa (2000), o assunto ontologias é o estudo das categorias de coisas que existem ou podem existir em algum domínio. O resultado deste estudo, então denominado de ontologia, é um catálogo dos tipos de coisas supostas a existir em um domínio de interesse, na perspectiva de uma pessoa. Esse catálogo é expresso através de uma linguagem para ontologias.

Em geral, humanos usam a liguagem para se comunicar e criar modelos de mundo. Mas, a linguagem natural não é adequada para construir modelos para a Ciência da Computação, por ser muito ambígua. Entretanto, linguagens formais são usadas para especificar modelos de mundo. Uma linguagem formal bem conhecida é a lógica de primeira ordem (FOL- *first logic order*). A principal função das ontologias é fazer uma ponte entre a representação sintática da informação e como sua conceitualização é realizada. Compartilhar ou reusar conhecimento entre os sistemas é complicado, já que diferentes sistemas utilizam termos diferentes para descrever informação. A visão de linguagens formais foi unida à visão das ontologias, no campo da ciência da computação, para gerar uma definição formal de ontologias e assim facilitar a interoperabilidade de sistemas (MAEDCHE, 2002).

Maedche (2002) apresenta a descrição de ontologias em uma estrutura de 5-tupla composta pelos elementos primitivos de uma ontologia, isto é, conceitos, relacionamentos, hierarquia de conceitos, função que relaciona conceitos e um conjunto de axiomas. Ontologias que usam essa estrutura podem ser mapeadas para a maioria das linguagens de descrição de ontologias conhecidas. Essa 5-tupla, O:= $\{C, R, H^c, rel, A^o\}$, consiste em:

− C (conceitos/classes) e R (relacionamentos) são dois conjuntos disjuntos;

− H^C é uma relação direcionada $H^c \rightarrow C \; x \; C$ que é chamada hierarquia de conceitos ou taxonomia. Por exemplo, H^c *(C1, C2)* significa que *C1* é um subconceito de *C2*;

− *rel* é uma função *rel:* $R \rightarrow C \; x \; C$ que relaciona não taxonomicamente conceitos;

− A^o é um conjunto de axiomas expresso em linguagem lógica apropriada.

No campo da Ciência da Computação, a definição mais comum de ontologia é encontrada em Gruber:

> Ontologia é uma especificação formal e explícita de uma conceitualização, o que existe é aquilo que pode ser representado [...] Quando o conhecimento de um domínio é apresentado num formalismo declarado, o conjunto de objetos que podem ser representados são chamados de universo do discurso. Esse conjunto de objetos, e o relacionamento descritivo entre eles, são refletidos num vocabulário representacional com o qual um programa de conhecimento de base representa o conhecimento. Mas, no contexto de Inteligência Artificial, nós podemos descrever a ontologia de um programa pela definição de um conjunto de termos representacionais. Nesse tipo de ontologia, definições associam os nomes de entidades no universo do discurso (por exemplo, classes, relações, funções ou outros objetos) com textos legíveis descrevendo o que os nomes significam, e axiomas formais que limitam a interpretação e o uso bem formado desses termos. Formalmente, uma ontologia é uma afirmação da lógica teórica. (1995, p. 907-928)

De maneira mais sucinta, o W3C (*World Wide Web Consortium*) coloca que ontologias devem prover descrição para os seguintes tipos de conceito (KOIVUNEM, 2003):

- Classes (ou "coisas") nos vários domínios de interesse;
- Relacionamentos entre essas classes;
- Propriedades (ou atributos) que essas coisas devem possuir.

Nesta dissertação, seguindo a definição de Gruber (1995), o conceito adotado é que uma ontologia para modelar perfil dos pesquisadores é a representação de termos, definições e indicadores de qualidade do conceito pesquisador.

Independente da definição escolhida, as ontologias vêm sendo usadas para descrever diferentes "coisas" com variados graus de estruturação e diferentes propósitos. A variação vai desde simples taxonomias, como a proposta pelo Yahoo, até as representações para metadados, como o Dublin Core[8], chegando a modelos descritos em Lógica (BREITMAN, 2005). Segundo Guizzardi:

> À medida que tem crescido o interesse por ontologias pela comunidade de Ciência da Computação, elas têm sido utilizadas de diferentes maneiras. Muitas vezes são usadas para descrever domínios já consagrados, como Medicina, Engenharia e Direito, a fim de promover consenso entre a comunidade de agentes interessada no domínio em questão. Outras vezes, para promover integração entre bases de conhecimento de Sistemas Baseados em Conhecimento distintos. De forma geral, ontologias constituem uma ferramenta poderosa para suportar a especificação e a implementação de sistemas computacionais de qualquer complexidade (2000, p. 51).

Devido a existência de trabalhos que usam taxonomia e tesauro, é importante distinguir esses termos de ontologia, para evitar interpretações errôneas. Segundo Breitman (2005), resumidamente, uma taxonomia serve para classificar informação em uma hierarquia (árvore), utilizando o relacionamento pai-filho (generalização ou "tipo-de"). A autora, também afirma que um tesauro pode ser definido como uma taxonomia adicionada de um conjunto de relacionamentos semânticos (equivalência, associação, entre outros) entre seus termos.

[8] http://dublincore.org/

3.2 Classificação de ontologias

As ontologias não apresentam sempre a mesma estrutura, dependem sempre da proposta de cada uma. Mas existem algumas características e componentes básicos comuns que são encontrados em muitas ontologias, fazendo com que possuam semelhanças entre suas funções (FELICÍSSIMO, 2004). Os tópicos a seguir apresentam resumidamente as classificações.

3.2.1 Quanto à generalidade

Guarino e Giaretta (1995) definiram quatro diferentes tipos de ontologia, de acordo com a sua generalidade. São eles:

- Ontologias de alto nível: descrevem conceitos de forma bem geral como espaço, tempo, material, objeto, evento, ação, etc., os quais são independentes de um problema ou domínio particular.
- Ontologias de domínio: descrevem o vocabulário relacionado a um domínio genérico (como medicina ou automóveis)
- Ontologias de tarefas: descrevem uma tarefa genérica (como diagnóstico ou vendas).
- Ontologias de aplicação: descrevem conceitos dependendo do domínio e de tarefas particulares. Estes conceitos, freqüentemente, correspondem a papéis desempenhados por entidades do domínio, quando na realização de certas tarefas. A OntoResearcher se enquadra neste tipo.

Ainda segundo Guarino (1995), a figura 2.2 mostra a relação entre estas ontologias. Os conceitos de uma ontologia de domínio ou de tarefa devem ser especializações dos termos introduzidos por uma ontologia genérica (alto nível). Enquanto que os conceitos de uma ontologia de aplicação devem ser especializações dos termos das ontologias de domínio ou de tarefa.

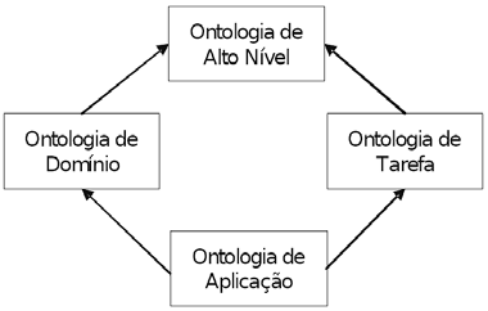

Figura 3.1: Tipos de ontologias, segundo seu nível de dependência em relação à uma tarefa ou ponto de vista particular (GUARINO, 1995).

3.2.2 Quanto ao tipo de informação que representam

Gómez- Pérez et al. (2004) sugerem uma classificação de ontologias baseando-se no tipo de informação a ser modelado. O autor identificou os seguintes tipos:

- Ontologias para representação do conhecimento: capturam primitivas de representação do conhecimento, fornecendo as primitivas para a modelagem

das linguagens baseadas em *frames*. O servidor Ontolingua[9] possui exemplos desse tipo de ontologia, como o OKBC e a Frame Ontology (BREITMAN, 2005).

- Ontologias gerais e de uso comum: incluem um vocabulário relacionado a coisas, eventos, tempo, espaço casualidade, comportamento, funções, etc. São usadas para representar conhecimentos de senso comum. Um exemplo é a ontologia time.daml[10].
- Ontologias de topo, genéricas ou de nível superior (*upper ontologies*): descrevem conceitos gerais. As ontologias SUMO[11] – *Standard Upper Merged Ontology* (IEEE Standard, 1995), a Ontologia de Guarino (GUARINO e WELTY, 2000) e a CYC Ontology (REED e LENHAT, 2002) são os exemplos mais conhecidos de *upper ontologies*. Tais ontologias são padrões do grupo de trabalho de ontologias genéricas da *IEEE– Institute of Electrical and Electronics Engineers*. Uma ontologia mais genérica (*upper ontology*) limita-se aos conceitos que são genéricos, abstratos e filosóficos. Conceitos específicos de um dado domínio não são incluídos nas ontologias genéricas. Assim, estas ontologias fornecem uma estrutura para a construção de outras ontologias de várias áreas de domínio.
- Ontologias de domínio: podem ter seus conceitos reutilizados dentro de um domínio específico. Termos e propriedades de uma ontologia de domínio são obtidos através da especialização de conceitos de uma ontologia de topo.
- Ontologias de tarefas: fornecem um vocabulário sistemático de termos, especificando tarefas que podem ou não estar no mesmo domínio.
- Ontologias de domínio-tarefa: são ontologias de tarefas que podem ser reutilizadas em um dado domínio, porém não em domínios similares. A OntoResearcher pode ser encaixada neste tipo, pois foi modelada com o intuito de armazenar o perfil dos pesquisadores e pode ser reutilizada no domínio acadêmico.
- Ontologias de métodos: fornecem definições para os conceitos e relacionamentos relevantes para um processo de modo a se atingir um objetivo.
- Ontologias de aplicação: são dependentes de uma determinada aplicação. Esse tipo de ontologia é usado para especializar e estender ontologias de domínio ou tarefa para uma dada aplicação.

3.2.3 Quanto ao grau de formalismo

Segundo Uschold e King (1995) de acordo com o formalismo que as ontologias representam, elas podem se dividir em quatro tipos:

[9] http://www.ksl.stanford.edu/software/ontolingua/

[10] Disponível em http://www.daml.org

[11] http://protege.stanford.edu/ontologies/sumoOntology/sumo_ontology.html

- Ontologias altamente informais: expressas livremente em linguagem natural. A OntoResearcher se encaixa neste tipo, pois não apresenta restrições formais aos conceitos que representa;
- Ontologias semi-informais: expressa em linguagem natural de forma restrita e estruturada;
- Ontologias semi-formais: expressas em uma linguagem artificial definida formalmente;
- Ontologias rigorosamente formais: expressas por meio de termos definidos com semântica, teoremas e provas.

3.3 Componentes de uma ontologia

De acordo com Noy e McGuinness (2001), o desenvolvimento de uma ontologia inclui definir as classes na ontologia, estruturar essas classes em uma hierarquia taxonômica, definir os *slots* (ou propriedades) e descrever os valores permitidos para estes *slots* (isto é, definir as restrições). Utilizando uma ontologia sobre vinhos para exemplificar, Noy e McGuinness (2001), definem os componentes ontológicos, que são:

- Classes ou conceitos: descrevem conceitos referentes ao domínio, sendo, muitas vezes, o foco das ontologias. Na classificação hierárquica as classes podem se subdividir em superclasses e subclasses. Uma subclasse herda as propriedades de sua superclasse. Por exemplo: uma classe de vinhos representa todos os vinhos, e pode ser subdividida em vinhos tintos, brancos e roses. Além disso, pode-se dividir em vinhos frisantes e não frisantes.
- Propriedades, *Slots*, atributos ou papéis: são as várias características e atributos que descrevem cada conceito, são as propriedades das classes. Por exemplo: a classe vinho pode ter as seguintes propriedades: cor; corpo; sabor; nível de açúcar e nível de tanino.
- Restrições ou facetas: são as restrições impostas às propriedades. As propriedades podem ter diferentes restrições descrevendo os tipos de valores, valores permitidos, números de valores (cardinalidade), etc. Por exemplo: na propriedade produtor da classe vinícola os valores serão os vinhos produzidos pela vinícola. A propriedade uvas de um vinho pode assumir o valor “um” ou “mais de um” se o vinho for feito por uma ou por mais de uma variedade de uva.
- Instâncias: representam os elementos de uma ontologia, ou seja, são os indivíduos que populam a ontologia. Uma instância é um conceito que pertence a uma classe e que possui determinadas propriedades. Por exemplo: a instância individual Chateau-Morgon-Beaujolais representa um tipo específico de vinho da classe Beaujolais. Esta instância tem as seguintes propriedades definidas:
    - Corpo: leve
    - Cor: vermelho
    - Sabor: delicado
    - Nível de tanino: baixo

- Uva: gamay (instância da classe de uva-do-vinho)
- Região: Beaujolais (instância da classe de região-dovinho)

3.4 Linguagens para representar Ontologias

Segundo Breitman (2005), na última década várias linguagens para desenvolvimento de ontologias foram propostas. Além disso, linguagens de representação de conhecimento, que nem foram criadas com esse propósito têm sido utilizadas no desenvolvimento de ontologias. Na década de 90, foram criadas algumas linguagens baseadas em princípios de inteligência artificial, sendo a maioria delas baseada em lógica de primeira ordem. Sendo que, a rápida evolução da internet foi o que fez com que surgissem linguagens de ontologias que ao mesmo tempo dão suporte e exploram características da rede. Essas linguagens são conhecidas como "linguagens leves" ou *lightweight* (BREITMAN, 2005).

De acordo com a arquitetura em camadas da Web Semântica (BERNERS-LEE, 2000), é que foram estruturadas as linguagens para ontologias. Essa estrutura em camadas é apresentada na Figura 3.2.

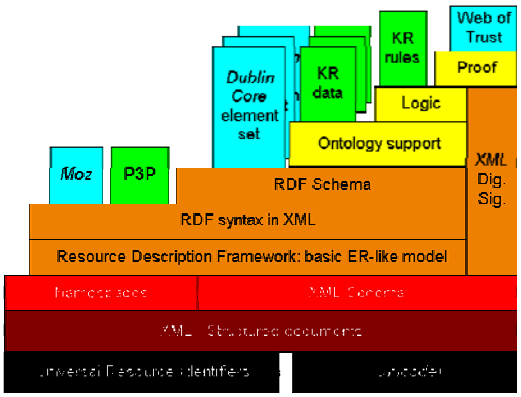

Figura 3.2: Arquitetura da Web Semântica (BERNERS-LEE, 2000).

Neste capítulo, será apresentada uma breve descrição das linguagens para construção de ontologias presentes na arquitetura da Web Semântica. Da arquitetura da Web Semântica, as camadas que possuem as linguagens abordadas são a *RDF syntax in XML*, *RDF Schema* e *Ontology Support*.

3.4.1 RDF, RDF Schema e RDF(S)

O RDF[12] e o RDF Schema[13] são as fundações da Web Semântica. Sendo que outras linguagens foram desenvolvidas com base nelas. O modelo de dados RDF foi proposto como uma recomendação W3C. Foi desenvolvido para a criação de metadados visando a descrição de recursos Web (BREITMAN, 2005).

O modelo RDF consiste de três tipos de objetos, que são: recursos (tudo aquilo que pode ser descrito por uma expressão RDF, por exemplo, uma página específica da Web, que possui um identificador único, URI), propriedades (aspecto, característica, atributo ou relação utilizada pra descrever um recurso) e declarações (acontece quando um valor

[12] http://www.w3.org/RDF/

[13] http://www.w3.org/TR/rdf-schema/

é atribuído a um recurso específico por meio de uma propriedade) (W3C Recommendation, 2004). A utilização de identificadores para os recursos e propriedades faz com que se tenha uma maneira global e única de nomear os itens em RDF (BREITMAN, 2005).

O modelo RDF não possui mecanismos para definição de relacionamentos entre propriedades e recursos. Para cumprir esta finalidade, surgiu o RDF Schema, que pode ser visto como uma extensão baseada em *frames* do RDF. A junção do RDF com o RDF Schema é conhecida como RDF(S), a qual combina redes semânticas com *frames*, porém não provê todas as primitivas que são geralmente encontradas em sistemas de representação do conhecimento baseadas em *frames*.

Os modelos RDF, RDF Schema e RDF(S) não podem ser considerados linguagens ontológicas, mas sim linguagens gerais que servem para descrever metadados na Web. Apesar de não ser uma linguagem e sim um modelo, a especificação de RDF propõe além do modelo uma sintaxe para sua especificação, baseada em XML. O RDF proporciona um modelo para descrição de metadados, já o XML fornece uma sintaxe de forma a permitir armazenar instâncias do modelo em arquivos processáveis e permitir a troca dessas instâncias entre as aplicações.

RDF é baseado na idéia de identificar coisas usando URI's e descrever recursos em termos de propriedades simples e valores de propriedades. Estas características habilitam o RDF a representar simples indicações sobre recursos como um grafo de nós e arcos representando os recursos, suas propriedades e valores. Os arcos são direcionados do recurso para o valor. Esse tipo de grafo é conhecido, na comunidade de inteligência artificial, como rede semântica. Um grafo RDF para a seguinte expressão: "existe uma pessoa identificada por http://www.w3.org/People/EM/contact#me, seu nome é Eric Miller, seu e-mail é em@w3.org" e seu título é Dr." (W3C Recommendation, 2004) pode ser representado como na Figura 3.3.

Figura 3.3: Um grafo RDF descrevendo Eric Miller (W3C Recommendation, 2004).

Grafos servem para transmitir as informações entre seres humanos, entretanto, na Web Semântica é necessária uma representação passível de processamento por máquinas. Assim, teremos uma representação do grafo em termos da sintaxe XML. Nessa representação, um documento RDF é representado utilizando-se um elemento XML com a etiqueta `rdf:RDF`. O conteúdo desse elemento é um conjunto de descrições que utilizam a etiqueta `rdf:Description`. Cada descrição se refere a um recurso, identificado em uma das formas seguintes:

- Atributo do tipo `about`, que faz referência a um recurso existente;
- Atributo do tipo `ID`, que cria um novo recurso;
- Sem nome, criando um atributo anônimo.

Desse modo, o grafo pode ser representado como na Figura 3.4.

```
<?xml version="1.0"?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
   xmlns:contact="http://www.w3.org/2000/10/swap/pim/contact#">
   <contact:Person rdf:about="http://www.w3.org/People/EM/contact#me">
   <contact:fullName>Eric Miller</contact:fullName>
   <contact:mailbox rdf:resource="mailto:em@w3.org"/>
   <contact:personalTitle>Dr.</contact:personalTitle>
  </contact:Person>
</rdf:RDF>
```

Figura 3.4: Descrição de Eric Miller baseado na sintaxe XML (W3C Recommendation, 2004).

3.4.2 SHOE

SHOE[14] (*Simple HTML Ontology Extension*) é um projeto da Universidade de Maryland, é uma extensão da linguagem HTML (*Hypertext Markup Language*) e seve para anotar conteúdo de páginas da Web, sendo que a informação é embebida nas páginas HTML. Esta linguagem oferece *tags* específicas que permitem a descrição de ontologias. E como essas *tags* não fazem parte da especificação HTML, elas não são mostradas através dos *browsers* padrão. O objetivo principal da linguagem SHOE é fornecer uma marcação para disponibilizar informações relevantes sobre o conteúdo das páginas, permitindo que agentes de software possam utilizar essas anotações para realizar buscas semânticas na rede (BREITMAN, 2005).

SHOE tem menos expressividade que RDF e apresenta grandes dificuldades na manutenção das páginas anotadas. O projeto SHOE foi descontinuado, e os pesquisadores que trabalhavam nele migraram para as linguagens DAML+OIL e OWL. Entretanto, a página ainda é mantida pela Universidade de Maryland.

3.4.3 OIL

A linguagem OIL[15] (*Ontology Inference Layer*) inclui-se no projeto On-to-Knowledge e foi patrocinada por um consórcio da Comunidade Européia. Essa linguagem foi criada para suprir a necessidade de uma linguagem que permitisse a modelagem de ontologias na Web, visto que o RDF não provê a semântica necessária nem o formalismo suficiente para permitir suporte a mecanismos de inferência (BREITMAN, 2005).

OIL combina primitivas de linguagens baseadas em *frames* com semântica formal e serviços de raciocínio providos pela lógica de descrição. Ou seja, combina ao mesmo tempo (BREITMAN, 2005):

- Lógica de descrição, fornecendo semântica formal e suporte à inferência;

[14] http://www.cs.umd.edu/projects/plus/SHOE/

[15] http://www.ontoknowledge.org/oil/

- Sistemas baseados em *frames*, portanto fornece primitivas de modelagem epistemológica;
- Linguagens da Web, portanto OIL é baseada nas sintaxes de XML e RDF.

OIL apresenta uma abordagem baseada em camadas, onde cada camada adicionada acrescenta funcionalidade e complexidade à camada anterior. Isto foi feito para que os agentes (humanos ou máquinas) que podem apenas processar as camadas inferiores, consigam compreender as ontologias que são desenvolvidas em um nível mais alto, mesmo que parcialmente. As camadas são[16]:

- *Core* OIL: coincide com RDF Schema (com exceção da característica de retificação). Isto significa que qualquer agente RDF Schema consegue processar ontologias OIL, e melhorar seu significado.
- *Standard* OIL: é uma linguagem para capturar as principais primitivas de modelagem, provê poder de expressividade, bem como permite que a semântica seja precisamente especificada e as inferências sejam viáveis.
- *Instance* OIL: enquanto a camada Standard OIL inclui a modelagem de construtos que permite que um *filler* seja especificado em termos de definição, *Instance* OIL inclui a capacidade completa de uma base de dados.
- *Heavy* OIL: pode incluir capacidades de representação e raciocínio adicionais. Ainda não tem uma sintaxe definida.

A comunidade de pesquisadores que utiliza OIL desenvolveu uma série de ferramentas para edição e verificação (através de mecanismos de inferência) para ontologias. Existem três editores disponíveis para OIL, que são: OntoEdit[17], OILEd[18] e o Protégé-2000[19] (BREITMAN, 2005).

3.4.4 DAML e DAML + OIL

A linguagem DAML[20] (*DARPA Agent Markup Language*) foi desenvolvida pelo DARPA[21] (*Defence Advanced Research Projects Agency*) em conjunto com o W3C. DAML é uma extensão de RDF que acrescenta construtos mais expressivos. O principal objetivo dessa linguagem foi o de facilitar a interação de agentes de software autônomos na Web (BREITMAN, 2005).

A linguagem DAML herdou muitos aspectos presentes em OIL, podendo-se dizer que as duas linguagens têm funcionalidades similares. No entanto, DAML não provê um motor de inferência. A combinação dessas duas linguagens gerou a DAML+OIL.

A semântica formal de DAML+OIL é fornecida através do mapeamento da linguagem para a linguagem KIF (*Knowledge Interchange Format*). DAML+OIL é

[16] http://www.ontoknowledge.org/oil/

[17] http://ontoserver.aifb.uni-karlsruhe.de/ontoedit

[18] http://oiled.man.ac.uk/

[19] http://smi.stanford.edu/projects/protege/

[20] http://www.daml.org/

[21] http://www.darpa.mil/

dividida em duas partes, a saber: domínio dos objetos (consiste nos objetos que são membros de classes definidas na ontologia DAML) e o domínio dos tipos de dados (consiste nos valores importados do modelo XML). Essa separação foi feita para permitir a implementação de mecanismos de inferência, visto que, realizar inferências sobre tipos concretos de dados não seria possível. A composição de DAML é a seguinte (BREITMAN, 2005):

- Elementos de classe: associam uma classe à sua definição. A definição de classe pode ter os seguintes elementos: rdfs:SubClassOf, daml:DisjointWith, daml:DisjointUnionOf, daml:SameClassAs e daml:EquivalentTo. A expressão SubClassOf indica a generalização da classe, importada diretamente da definição presente no RDF(S). Isso se dá porque a linguagem DAML+OIL funciona como uma camada sobre o RDF(S).
- Expressões de classe: é como se pode representar uma classe. Podem ser dos seguintes tipos: nome de classe, enumeração, restrição e combinação booleana. Na linguagem DAML+OIL não se pode atribuir o mesmo nome às classes distintas, pois o nome funciona como um identificador.
- Propriedades: associam uma propriedade a sua respectiva definição. Propriedades podem ser definidas de acordo com os seguintes elementos: rdfs:SubPropertyOf, domínio, rdfs:range, daml:SamePropertyAs, daml:EquivalentTo, daml:InverseOf. Algumas das propriedades são definidas na camada DAML (aquelas que iniciam por daml:) outras são importadas da camada RDF (as que começam por rdf:).

3.4.5 OWL

RDF Schema auxilia a criar classificações, grupos de conceitos e para escrever instruções, mas não é expressivo o suficiente para algumas entidades do mundo real. Os esforços feitos em DAML e OIL conduziram ao OWL[22], que é a linguagem para expressar ontologias (LUGANO, 2005).

As linguagens ontológicas disponíveis dificultavam a tarefa de compartilhar e reusar ontologias entre diferentes aplicações de um mesmo domínio, ou de domínios inter-relacionados. Para resolver o problema da interoperabilidade e definir um paradigma universal para a troca de informação ontológica baseada na Web, o W3C criou a OWL, a qual começou como uma *W3C Recomendation* em fevereiro de 2004 (LUGANO, 2005).

A sintaxe da linguagem OWL é fornecida pelo XML, com o esquema de *tags* (rótulos escondidos com anotações). O *framework* para a representação de informação na Web através da modelagem de seus metadados e das relações entre eles é fornecido pelo RDF. Já o vocabulário para descrição dos conceitos e relações dos recursos, com semântica para as hierarquias de especialização das relações e dos conceitos, é fornecido pelo RDF Schema (FREITAS, 2005). Portanto, OWL estende as funcionalidades do RDF e RDF Schema, mantendo compatibilidade com o *design* arquitetural básico da Web. Isto é, aberto, não proprietário, distribuído para muitos sistemas e permite compartilhamento de dados (através das ontologias) (LUGANO, 2005).

[22] http://www.w3.org/TR/owl-features/

OWL permite qualificar os relacionamentos, por exemplo, explicitar que uma relação é simétrica, transitiva, funcional, funcional inversa, etc. Também é possível definir a cardinalidade entre as relações. Propriedades ou classes podem ser definidas como equivalentes a outras propriedades ou classes. Além disso, através de restrições OWL é possível especificar quantificações existenciais ou universais (VIEIRA et al., 2005).

A linguagem OWL é dividida em três sublinguagens, de acordo com a sua expressividade (SMITH et al., 2003):

- *OWL Lite:* abrange a expressividade de *frames* e lógica de descrições, com algumas restrições. Por exemplo, a cardinalidade máxima ou mínima assume apenas os valores 0 ou 1. Apesar disso, a linguagem é dotada de riqueza semântica, sendo, por isto, ideal para usuários iniciantes e desenvolvedores que preferem *frames* à lógica de descrições. Atributos (aqui chamados de propriedades) podem ter transitividade, simetria, atributos inversos, propriedades funcionais (se `P(x,y) ^ P(y,x) => x=y`), funcionais inversas (se `P(x,y) ^ P(z,x) => x=z`) e papéis.
- *OWL DL:* garante completude, decidibilidade e toda a expressividade da lógica de descrições, almejando satisfazer engenheiros de conhecimento familiarizados com esta tecnologia. A expressividade torna-se ainda maior do que em *OWL Lite*, pois classes podem ser construídas por união, interseção e complemento, pela enumeração de instâncias e podem ter disjunções. Tipos são mantidos cuidadosamente separados (por exemplo, uma classe não pode ser instância e propriedade ao mesmo tempo).
- *OWL Full:* fornece a expressividade de OWL e a liberdade de usar RDF, inclusive permitindo novas metaclasses, já que elas são subclasses definidas em RDFS. Entretanto, fazendo este uso mais complexo, não há garantia de computabilidade. Nesta sublinguagem, não cabem as restrições de separação de tipos da versão anterior, sendo possível manipular e modificar metaclasses.

As linguagens menos expressivas estão contidas dentro das mais expressivas, de maneira que uma ontologia definida numa linguagem menos expressiva é aceita por uma linguagem mais expressiva; a recíproca, naturalmente, não é verdadeira. A OntoResearcher foi descrita em OWL DL, pois esta linguagem apresenta os construtos e formalismos necessários e suficientes para o desenvolvimento da ontologia de perfil.

3.4.5.1 Tipos Básicos do OWL

- *Headers*: para definir uma ontologia em OWL, é preciso em primeiro lugar, dizer onde estão (na Web) as classes primitivas das ontologias, para que seja possível definir novas classes como subclasse destas. Também é necessário determinar um *namespace* para a nova ontologia. Isto é codificado da seguinte forma num trecho da ontologia chamado de *Header*s, como no exemplo da Figura 3.5 (CHEN et al., 2003).

```
<rdf:RDF
xmlns="file:/G:/myclasses#"
xmlns:eyeglass="file:/G:/Glasses#"
xmlns:owl="http://www.w3.org/2003/02/owl#"
xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
xmlns:xsd="http://www.w3.org/2000/10/XMLSchema#">
```

Figura 3.5: Headers (CHEN et al., 2003).

As classes a serem definidas estarão localizadas no *namespace* da primeira definição. A segunda definição serve para que ontologias externas possam referenciar a ontologia sendo definida. As restantes localizam as definições primitivas de OWL, RDF, RDFS e XMLSchema.

- Classes e atributos: classes podem ser construídas de várias formas - por herança, união, interseção, complemento, pela enumeração de instâncias ou por restrições de propriedades. Por exemplo, o trecho de código da Figura 3.6 (COSTELLO et al., 2003) a classe "rio fluvial" é subclasse de "rio", e o atributo "desemboca" (emptiesInto) tem como imagem instâncias da classe "águas" (BodyOfWater).

```
<owl:Class rdf:ID="Flueve">
<rdfs:subClassOf rdf:resource="#River"/>
</owl:Class>
<rdf:Property rdf:ID="emptiesInto">
<rdfs:domain rdf:resource="#River"/>
<rdfs:range rdf:resource="#BodyOfWater"/>
</rdf:Property>
```

Figura 3.6: Classe Flueve e subclasse River (COSTELLO et al., 2003).

Note-se as referências às superclasses primitivas owl:Class e rdf:ID, que conseguem ser localizadas por causa dos *Headers*. Outro ponto a ser analisado, é que, em lógica de descrição, a definição dos atributos não precisa estar junto da classe. A expressividade da lógica de descrição possibilita definir a classe "rio fluvial" como subclasse de "rio" que desemboca (atributo emptiesInto) em mares. Ou seja, a classe é definida com o auxílio de uma restrição, como na Figura 3.7 (CHEN et al., 2003). Em OWL, propriedades são usadas para criar restrições (BREITMAN, 2005).

```
<owl:Class rdf:ID="Flueve">
<rdfs:subClassOf rdf:resource="#River"/>
<rdfs:subClassOf>
<owl:Restriction>
<owl:onProperty rdf:resource="#emptiesInto"/>
<owl:allValuesFrom rdf:resource="#Sea"/>
</owl:Restriction>
</rdfs:subClassOf>
</owl:Class>
```

Figura 3.7: Atributos de River (COSTELLO et al., 2003).

3.4.6 Considerações

A Tabela 3.1 apresenta um breve resumo das linguagens estudadas. A ontologia de perfil de pesquisador, OntoResearcher, foi desenvolvida em OWL, mais especificamente na sublinguagem OWL DL. Isto porque a OWL Lite não permite certas restrições de cardinalidade e a OWL Full não tem as restrições de separação de tipos da OWL DL (uma classe pode ser instância e propriedade ao mesmo tempo) e não tem garantia de computabilidade.

Tabela 3.1: Quadro resumo das linguagens ontológicas.

	Proposto por/ autores	**Objetivos**	**Primitivas de Representação**	**Propriedades lógicas**
OWL	W3C/ Dean and Schereiber	Web Semântica	Classes Indivíduos Propriedades Relações Axiomas	Cardinalidade Equivalência Disjunção Simetria
OIL	European IST project On-To-Knowledge Horrocks et al, 2000	Web Semântica	Classes Indivíduos Propriedades Relações Axiomas	Transitiva Funcional Simétrica
DAML +OIL	DARPA e W3C	Facilitar a integração de agentes de softwares autônomos na Web.	Classes Indivíduos Propriedades Relações Axiomas União Disjunção Equivalência	
RDF	W3C, 1999	Criação de metadados visando a descrição de recursos Web.	Classes Propriedades Indivíduos	Não possui propriedades lógicas.
SHOE	Universidade de Maryland. OBS: projeto descontinuado	Fornecer marcação sobre informações relevantes de páginas, para que agentes de software usem essas anotações nas buscas semânticas.		

3.5 Ferramentas para editoração de ontologias

Algumas ferramentas têm sido utilizadas para auxiliar o desenvolvimento, construção e manipulação de ontologias. Essas ferramentas podem ser classificadas em três categorias: editores de ontologias (Protégé, OntoEdit[23], etc.), metadados e ferramentas de visualização em mecanismos de inferência.

Os mecanismos de inferência são ferramentas de software capazes de derivar novos fatos ou associações a partir de informações existentes, alguns exemplos são JESS[24],

[23] Foi desenvolvido pela Ontoprise, empresa criada na Universidade de Karlsruhe (Alemanha). A primeira versão foi implementada em 1992.

[24] http://herzberg.ca.sandia.gov/

Pellet[25], RACER[26]. Nesse capítulo serão descritas as ferramentas utilizadas para desenvolver a OntoResearcher.

3.5.1 Protégé

É um ambiente interativo para projeto de ontologias, de código aberto, que oferece uma interface gráfica para edição de ontologias e uma arquitetura para a criação de ferramentas baseadas em conhecimento. A arquitetura é modulada e permite a inserção de novos recursos (NOY; McGUINESS, 2001).

O Protégé sempre procurou crescer em número de usuários, passando por várias reengenharias e reimplementações, provendo ferramentas simples e configuráveis (FREITAS, 2006). Assim, surgiu uma arquitetura integrável a diversas aplicações, via componentes que podem ser conectados ao sistema. Como conseqüência desta decisão e de sua difusão, componentes elaborados por grupos de pesquisa de usuários, foram adicionados ao sistema, sem necessitar o redesenvolvimento. Foram aproveitados, por exemplo, o Jambalaya[27], um utilitário com animação e vários outros recursos em visualização de dados, e o OntoViz[28], um componente que faz com que o gerador de gráficos Graphviz[29] da AT&T produza gráficos com instâncias, heranças e outros tipos de relacionamento.

O projeto do Protégé tem seu modelo de conhecimento extensível, ou seja, é possível redefinir declarativamente as classes primitivas (ou metaclasses) de um sistema de representação. O conjunto de metaclasses usados por default pelo sistema implementa características comuns a frames, tornando-o fácil de usar mesmo para usuários leigos. Todavia, se forem utilizadas metaclasses complexas e distintas das originais - como as para definir classes em RDF, por exemplo - as instâncias alcançarão a expressividade e complexidade desejada. Por isso, o Protégé pode ser adaptado a diversos usos. Algumas características do Protégé são (HORRIDGE et al., 2004):

- A linguagem axiomática PAL (*Protégé Axiomatic Language*), permite a inserção de restrições e axiomas que incidem sobre as classes e instâncias de uma ou mais ontologias.
- A geração de arquivos de saída alteráveis, permitindo que sejam implementados componentes para traduzir o conhecimento para outros formalismos através das metaclasses.
- Possui uma interface agradável para entrada de conhecimento, incluindo um gerador automático de formulários para as classes definidas, admitindo ainda a reposição da interface original por componentes mais adequados às aplicações específicas. Esta interface facilita o gerenciamento de conhecimento de uma ou mais ontologias.

[25] http://pellet.owldl.com/

[26] http://www.racer-systems.com/

[27] http://protegewiki.stanford.edu/index.php/Jambalaya

[28] http://protegewiki.stanford.edu/index.php/OntoViz

[29] http://www.graphviz.org/

Na Figura 3.8 apresenta-se a tela principal da ferramenta, onde foram sinalizadas algumas áreas da interface, tais como: na seta 1 têm-se as guias que permitem a alternância na edição das classes, propriedades, formulários, instâncias e metadados. Em 2 é apresentada a visualização da hierarquia de classes, bem como a classe base "owl:Thing" da qual, todas as outras são especializações. A área 3 é para preenchimento do nome e outras informações relevantes de cada classe. A área 4 serve para exibir a descrição lógica de cada classe. Na área 5 são mostradas as propriedades relacionadas com a classe que está em edição. Por fim, em 6 são apresentadas as classes disjuntas (VIEIRA et al., 2005).

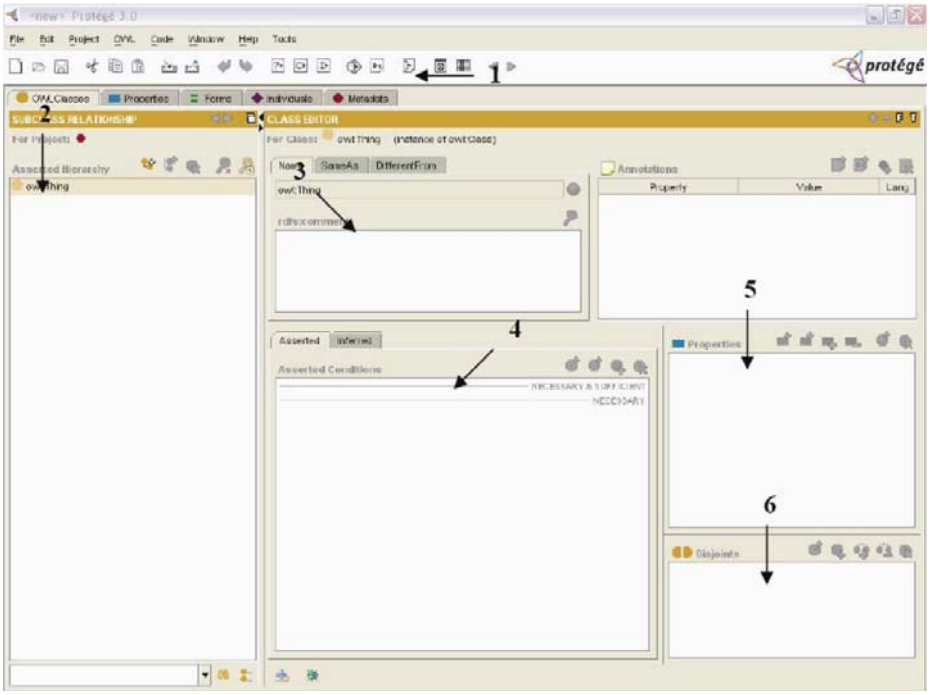

Figura 3.8: Tela principal do Protégé (Modificado de: VIEIRA et al., 2005).

O Protégé foi a ferramenta escolhida para desenvolvimento da OntoResearcher, por algumas razões como: é grátis e tem disponibilidade para *download*; existem comunidades de pesquisa confiáveis que utilizam a ferramenta; a ferramenta passa por freqüentes atualizações; a interface é amigável; é fácil de usar e possui diversos *plugins*[30] que aumentam as funcionalidades do Protégé. Neste trabalho foram utilizados alguns *plugins*, tais como: OWL *plugin*[31] que permite exportar e importar ontologias no formato OWL (HORRIDGE et al., 2004); o OntoViz[32] e o OWLViz[33] que permitem ao usuário do Protégé trabalhar no desenvolvimento de ontologias na forma de diagramas.

3.5.2 Jena

A partir de uma ontologia é possível recuperar conhecimento, de acordo com sua semântica, através do uso de um motor de inferência e também descobrir novas

[30] http://protege.cim3.net/cgi-bin/wiki.pl?ProtegePluginsLibraryByTopic

[31] http://protege.stanford.edu/overview/protege-owl.html

[32] http://protege.cim3.net/cgi-bin/wiki.pl?OntoViz

[33] http://www.co-ode.org/downloads/owlviz/

informações através de linguagens de consulta. Por meio das regras de inferência, e de linguagens de consultas podem-se derivar novos fatos baseados em fatos existentes. A ferramenta utilizada, nesta dissertação, é o Jena (JENA, 2006). O Jena é um projeto *open-source* desenvolvido pelo *HP Labs Semantic Web Programme*. É uma API (*Application Programming Interface*) para construção de aplicações voltadas à Web Semântica que fornece um ambiente de programação para RDF (e também OWL, RDFS, DAML) e inclui um motor de inferência baseado em regras. Por meio do Jena é possível não só manipular, consultar e persistir arquivos OWL, mas também criar novos motores de inferência ou mesmo estender os motores já existentes.

A API para ontologias *Jena Ontology API* é independente de linguagem, então para representar as diferenças entre as várias representações, cada linguagem ontológica tem um *profile*, o qual lista os construtos permitidos e as URI's das classes e propriedades. Por exemplo, o OWL *profile* é *owl:ObjectProperty*. O *profile* é limitado a um *ontology model* que é uma versão estendida do *Model class* do Jena. O *OntModel* extende o *Model* geral adicionando suporte para os tipos de objetos que se espera ter uma ontologia: classes, propriedades e indivíduos. Na Figura 3.9 é apresentada, de forma resumida, a arquitetura base da API Jena. O *Ontology Model* é usado para representar modelos ontológicos em memória; o *Reasoner* permite inferir informações acerca dos modelos e *Base RDF Graph* usa a API de RDF para representar os modelos. Os *reasoners* de OWL do Jena funcionam aplicando regras tipo *if-then-else* sobre instâncias OWL. Como o OWL está definido sobre RDF(S) o Jena usa suas APIs de RDF para poder manipular as ontologias, e por isso a arquitetura possui o terceiro módulo.

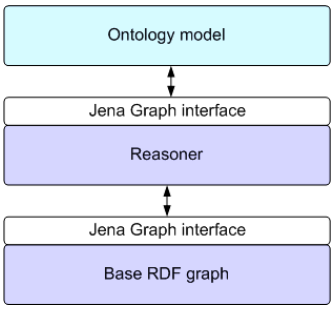

Figura 3.9: Arquitetura base da API Jena.

O Jena possui classes para executar consultas (*query*) sobre modelos ontológicos, essas consultas são feitas em RDQL[34] (A *Query Language for RDF*). O RDQL executa consultas sobre modelos OWL, tratando estes como um conjunto de triplos (Sujeito Propriedade Valor). Assim as consultas RDQL são padrões desse triplo e seguem a mesma sintaxe básica de SQL (SELECT *variáveis* WHERE *condições*). As varáveis são representadas com o ponto de interrogação seguido pelo nome da variável, por exemplo, ?a, ?b. Os recursos são enclausurados entre "<>".

É possível realizar consultas "inteligentes", utilizando primeiramente o motor de inferência baseado em regras (que contêm boa parte das regras semânticas da linguagem OWL) para gerar as triplas que serão relevantes (ou não) para a consulta desejada, depois é realizada uma consulta direta sobre os dados originais junto com os dados inferidos. O Jena foi utilizado para manipulação da OntoResearcher e para realização de

[34] http://www.w3.org/Submission/2004/SUBM-RDQL-20040109/

consultas, pois possui uma API Java que permitiu fácil integração com o código da aplicação desenvolvida nesta dissertação.

3.6 Perfis descritos como ontologias

No contexto da Web semântica têm-se recursos que representam diversos objetos do mundo real, tais como: pessoas, dispositivos, serviços, empresas, agentes de software, dentre outros. Os recursos possuem relacionamentos entre si. Então, através de inferências computacionais sobre as propriedades dos recursos e dos relacionamentos é possível descrever características lógicas úteis no uso dessas informações (VIEIRA et al., 2005). A possibilidade de utilização de inferências computacionais é um dos benefícios do uso de ontologias para descrição de perfis (VIEIRA et al., 2005). Outros benefícios são (GUIZZARDI, 2000):

- Comunicação e interoperabilidade: ontologias são úteis para ajudar as pessoas a tratarem sobre um determinado conhecimento. Atuam na definição de um consenso acerca do vocabulário técnico comum a ser usado nas suas iterações;
- Formalização: a notação formal utilizada elimina as contradições e inconsistências resultando em uma especificação não ambígua do domínio representado. Por utilizar uma notação formal pode ser automaticamente verificada e validada, sendo possível também, realizar inferências de forma automática;
- Representação do conhecimento de forma organizada e reuso: ontologias são construídas para representar o conhecimento do domínio de forma explícita, não ambígua, possuindo um potencial enorme de reuso.

3.7 Considerações

Nesta dissertação, a definição do perfil é justificada pela necessidade de conhecer a atuação acadêmica de um pesquisador. O perfil dos pesquisadores foi modelado em uma ontologia. Como a abordagem ontológica adotada é a base do protótipo desenvolvido, este capítulo apresentou a fundamentação teórica sobre o tema ontologia. Foram apresentadas as justificativas para o uso de ontologias, bem como uma breve descrição dos conceitos, das linguagens, das ferramentas para o desenvolvimento da ontologia de perfil de pesquisador descrita nesta dissertação.

4 QUALIFICAÇÃO DE PESQUISADORES

Este capítulo descreve a abordagem proposta para o problema de descobrir a qualificação dos pesquisadores nas áreas da Ciência da Computação. Primeiramente é apresentado o processo de desenvolvimento da ontologia de perfil, desde a definição dos requisitos até a descrição das classes e propriedades definidas. Para qualificar os pesquisadores foi desenvolvido o protótipo de um sistema Web, o qual é centrado na OntoResearcher e considera o reuso de outras ontologias para extração de informações baseadas em regras. Além de prover o reuso de outras ontologias, o protótipo também utiliza extração de informação de arquivos XML e da Web tradicional (através de consultas ao site Google Scholar).

A OntoResearcher é uma ontologia modelada com termos, definições e indicadores de qualidade científica, os quais foram definidos com base em conceitos da Ciência da Computação. Esses termos, definições e indicadores foram utilizados para o processo de definição da OntoResearcher. Entretanto, alguns pesquisadores podem considerar outros indicadores que não apenas os utilizados nesta dissertação, ou mesmo considerar apenas alguns desses indicadores. Existem muitas questões a serem discutidas neste sentido, sendo esta abordagem o início da busca por uma proposta mais abrangente para a qualificação dos pesquisadores.

A seção 4.1 apresenta o processo de definição do perfil dos pesquisadores. A seção 4.2 mostra a ontologia OntoResearcher desenvolvida nesta dissertação. A seção 4.3 apresenta os indicadores utilizados e o cálculo para qualificar os pesquisadores nas áreas da Ciência da Computação. A seção 4.4 apresenta uma descrição das funcionalidades do sistema. A seção 4.5 discute as implementações para obter as informações sobre os pesquisadores e então qualificá-los nas áreas da Ciência da Computação.

4.1 Definição do perfil de pesquisador

A necessidade de medir a qualidade do trabalho científico torna imprescindível descobrir o quão competente um pesquisador é em determinada área. Para quantificar e qualificar a atuação e competência dos pesquisadores nas áreas da Ciência da Computação é que a OntoResearcher foi desenvolvida. A descoberta das competências dos pesquisadores está relacionada com o seu perfil acadêmico, por isso, existe a necessidade de identificar esses perfis.

Para modelar o perfil do pesquisador foram identificadas as características relevantes (indicadores) através da análise de duas fontes de informações, que são: o

currículo Lattes do CNPq e os critérios utilizados pelo CNPq para conceder a bolsa de produtividade científica.

A escolha pelo currículo Lattes se deu por alguns motivos, como: (i) no Lattes encontram-se grande parte dos dados necessários para qualificar um pesquisador; (ii) o Lattes é um padrão de currículo brasileiro; (iii) o Lattes é disponibilizado pelo CNPq no formato XML, o que facilita o processo de obtenção dos dados e população automática da ontologia. Entretanto, o currículo Lattes no formato XML só pode ser obtido pelo próprio pesquisador, por esta razão é necessário solicitar que ele submeta tal arquivo ao sistema.

A Bolsa de Produtividade em Pesquisa do CNPq é concedida como prêmio para os pesquisadores que mais produzem no Brasil. As áreas do conhecimento têm comitês de assessoramento específicos que se baseiam nos critérios definidos pelo CNPq e acrescentam outros critérios, aumentando o rigor e detalhamento do perfil do pesquisador e a qualidade intelectual (NIEDERAUER, 2002). Para esta dissertação, foram analisadas as informações do comitê de assessoramento da área da Ciência da Computação[35].

Do currículo Lattes foram selecionadas as informações:

- Nome do pesquisador, e-mail, homepage, endereço profissional ou residencial (depende do que o pesquisador selecionou como preferencial), país, instituição (pode ser mais de uma instituição) e país da instituição;
- Formação acadêmica (nível: pós-doutorado, doutorado, mestrado, especialização e graduação), o título do trabalho de diplomação, o orientador, a instituição e o ano de conclusão da formação acadêmica;
- Disciplinas ministradas (se é para doutorado e mestrado, especialização ou graduação), e o nome da disciplina;
- Orientações (pós-doutorado, doutorado, mestrado, especialização ou graduação), o nome do orientando, o título do trabalho, o ano de conclusão e a instituição;
- Idiomas (se o pesquisador compreende, escreve, fala e lê), nome do idioma;
- Produção bibliográfica:
  - Para artigos em conferências: título do artigo, DOI, idioma da publicação, título do evento, país e ano do evento, título dos *proceedings*;
  - Para capítulos de livro e livros: título do capítulo e do livro, idioma da publicação, ISBN e nome dos autores;
  - Para artigos em Journals: título do artigo, DOI, idioma da publicação, título do *Journal* e ISSN.
- Projeto de pesquisa: o papel (coordenador ou colaborador) do pesquisador em um projeto de pesquisa, o título do projeto de pesquisa e o ano de conclusão;

[35] http://portal.cnpq.br/cas/ca-cc.htm#criterios

- Participação em comitê de programa de conferências científicas.

Da análise dos critérios do CNPq para conceder bolsa de produtividade científica aliadas ao documento Qualis-Capes para avaliação dos veículos de publicação foram selecionadas as seguintes informações:

- Número de citações de cada publicação do pesquisador;
- A área: das disciplinas ministradas, dos trabalhos orientados, das publicações, dos projetos de pesquisa e da formação acadêmica;
- O Qualis das publicações, o qual serve como um indício da qualidade das publicações de um pesquisador.

Algumas informações como: participação em bancas, produção técnica, outras produções, prêmios e títulos, e dados complementares não foram consideradas neste trabalho. A principal razão para desconsiderar essas informações foi encontrada em uma análise feita por Cazella (2006) que demonstrou que tais informações não têm muita influência na competência dos pesquisadores, estas informações estão no Anexo A.

4.2 OntoResearcher

O desenvolvimento da OntoResearcher foi baseado nos indicadores de qualidade descritos na seção 4.1. Ela foi desenvolvida utilizando a linguagem OWL-DL e o software Protégé. Como citado na seção 3.3, o desenvolvimento de uma ontologia inclui definir as classes na ontologia, estruturar essas classes em uma hierarquia taxonômica, definir os *slots* (ou propriedades) e descrever os valores permitidos para estes *slots*. Esta seção apresenta e descreve os elementos da ontologia de perfil.

4.2.1 As classes

Noy e McGuinness (2001) afirmam que as classes descrevem conceitos referentes ao domínio em questão, elas podem se subdividir em superclasses e subclasses, sendo que cada subclasse herda as propriedades de sua superclasse. Em OWL, as classes são interpretadas como conjuntos que contém indivíduos, por exemplo, a classe *Instituition* contém indivíduos que são instituições de ensino ou instituições que trabalham com pesquisa. A Figura 4.1 mostra a estrutura de classes da OntoResearcher. As classes *RA:ResearchArea*, *C:Country* e *L:Langua*ge são referentes as ontologias importadas. A classe *owl:Thing* é criada por *default* no Protégé.

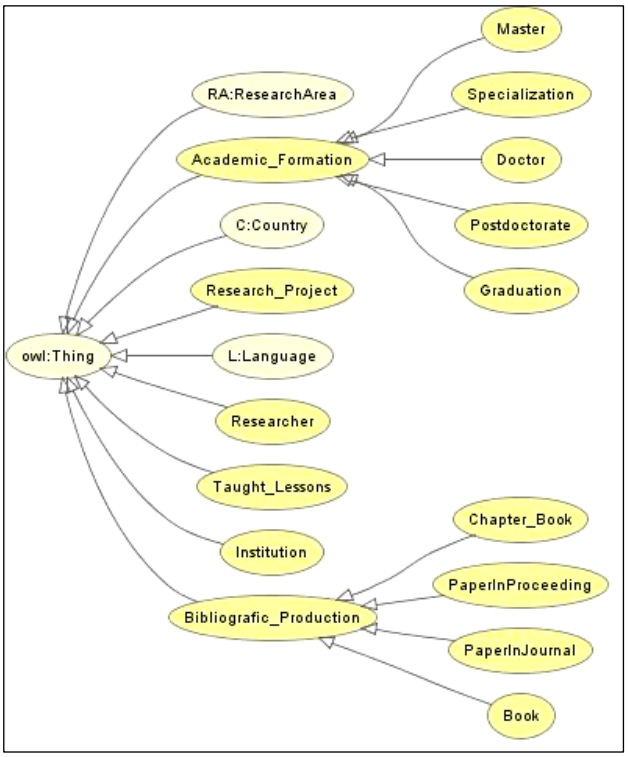

Figura 4.1: Estrutura das classes da OntoResearcher

As ontologias (*Language*, *Country* e *ResearchArea*) foram importadas por uma questão conceitual, que é o reuso de ontologias. Como a OntoQualis e a OntoDoc também necessitam dos conceitos, que as ontologias importadas representam, optou-se por desenvolver ontologias separadas e que pudessem ser reusadas (evitando repetições dos conceitos presentes nas 3 ontologias importadas) no contexto do projeto DIGITEX.

A *ResearchArea* representa as áreas em que um pesquisador pode atuar e foi desenvolvida baseando-se nas áreas da Ciência da Computação descritas pela ACM[36]. A ontologia *Country* representa o nome dos países em inglês, ela foi baseada na norma ISO 3166-1 (ISO, 2007). A *Language* representa o nome das linguagens em inglês e foi baseada na norma ISO 639-2 alpha-3 (ISO, 1998). A Figura 4.2 apresenta o trecho de código OWL da OntoResearcher com o processo de importação das três ontologias.

```
<owl:Ontology rdf:about="">
<owl:imports rdf:resource="http://www.inf.ufrgs.br/~khannel/Ontology/ResearchArea.owl"/>
<owl:imports rdf:resource="http://www.inf.ufrgs.br/~khannel/Ontology/Language.owl"/>
<owl:imports rdf:resource="http://www.inf.ufrgs.br/~khannel/Ontology/Country.owl"/>
```

Figura 4.2: Código da importação das ontologias.

As 15 classes e subclasses da OntoResearcher são:

- *Academic_Formation*: representa o quanto um pesquisador é graduado. Esta classe possui cinco subclasses, que são:
  - *Graduation*: o pesquisador possui graduação;
  - *Specialization*: o pesquisador possui especialização;
  - *Master*: o pesquisador possui mestrado;

[36]http://portal.acm.org/ccs.cfm?part=author&coll=GUIDE&dl=GUIDE&CFID=23845712&CFTOKEN=62061851

- *Doctor*: o pesquisador possui doutorado;
- *Pos-Doc*: o pesquisador possui pós-doutorado;

- *Research_Project*: projetos de pesquisa em que um pesquisador atua.
- *Researcher*: representa o conceito pesquisador;
- *Taught_Lessons*: representa as disciplinas ministradas por um pesquisador;
- *Institution*: representa as instituições, que podem ser sociedades que trabalham com pesquisa ou podem ser de ensino e pesquisa;
- *Bibliografic_Production*: representa as publicações do pesquisador. Tem 4 subclasses, que são:
  - *Chapter_Book: representa os capítulos de livros publicados;*
  - *PaperInProceeding: artigos publicados em conferências;*
  - *PaperInJournal: representa os artigos publicados em Journal;*
  - *Book: representa os livros publicados.*

4.2.2 As propriedades

A OntoResearcher possui 40 propriedades, destas, 23 são propriedades do tipo "*object*" (relaciona indivíduo(s) de uma classe a outro(s) indivíduo(s)) e as outras 17 são propriedades do tipo "*datatype*" (relacionam indivíduo(s) a um tipo de dado RDF literal ou a um valor XML *Schema Datatype*). Em OWL, as propriedades representam relações entre indivíduos, por exemplo, a propriedade *hasAuthor* relaciona indivíduo(s) da classe *Bibliografic_Production* com indivíduo(s) da classe *Researcher*.

Segundo Horridge et al. (2004) propriedades têm um domínio (*domain*) e um escopo (*range*) especificados. Assim, as propriedades ligam indivíduos do domínio a indivíduos do escopo. Por exemplo, para a propriedade *hasAuthor* o domínio é *Bibliografic_Production* e o escopo é *Researcher*. A Tabela 4.1 lista as propriedades *object* da OntoResearcher e a Tabela 4.2 lista as propriedades *datatype* da OntoResearcher. Ambas as tabelas apresentam o domínio, o escopo e uma breve descrição.

As propriedades das ontologias importadas não fazem parte das 40 propriedades, e são analisadas separadamente. Na ontologia *Country*, têm-se as seguintes propriedades: C:countryCodeISO3166Alpha2 e C:countryNameISO3166Short que são referentes aos códigos da norma ISO (ISO- INTERNATIONAL ORGANIZATION FOR STANDARIZATION, 2007) para os nomes dos países e são ambas do tipo *datatype* com o escopo *String*. Na ontologia *Language* as propriedades são: L:NameISO639 e L:CodeISO639 que se referem aos códigos da norma ISO (ISO- INTERNATIONAL ORGANIZATION FOR STANDARIZATION, 1998) para os nomes dos idiomas, são do tipo *datatype* com escopo *String*. Já na ontologia *ResearchArea* temos as propriedades: RA:hasSubArea que é inversa de RA:isSubAreaOf e representam a hierarquia das áreas da Ciência da Computação definidas pela ACM. Estas propriedades são do tipo *object* e o escopo é a RA:ResearchArea.

Tabela 4.1: Propriedades *Object da* OntoResearcher

Propriedades *Object*	Domínio/Escopo	Descrição
hasArea	*Academic_Formation, Bibliografic_Production, Taugh_Lessons* e *ResearchProject / RA:ResearchArea*.	Indivíduos das 4 classes do domínio possuem uma área que é um indivíduo da ontologia *ResearchArea*.
Advisor	*Researcher / Academic_Formation*	Pesquisador orienta uma formação acadêmica. É inversa de *hasAdvisor*.
hasAdvisor	*Academic_Formation / Researcher*	A formação acadêmica do pesquisador tem orientador.
WasFinishedIn	*Academic_Formation / Institution*	A formação acadêmica é realizada em uma instituição.
hasCountry	*Instituition, Researcher / C:Country*	Toda instituição e pesquisador têm um país.
hasPublication	*Researcher / Bibliografic_Production*	Pesquisadores possuem publicações (livro, capítulo de livro, *paper* em proceedind e em *journal*). É inversa de *hasAuthor*.
hasAuthor	*Bibliografic_Production / Researcher*	Toda produção bibliográfica tem pelo menos um autor.
ResearchProjectCoordinator	*Researcher / Research_Project*	Os pesquisadores são coordenadores de projeto de pesquisa.
Lesson_Tought_By_Researher	*Taught_Lesson / Researcher*	As disciplinas são ministradas por um pesquisador.
ReserchProjectColaborator	*Researcher / Research_Project*	Pesquisador atua como colaborador em projetos de pesquisa. Inversa de *hasColaborator*.
hasColaborator	*Research_Project / Researcher*	Os projetos de pesquisa têm colaboradores.
hasAcademicFormation	*Researcher / AcademicFormation*	Pesquisador tem formação acadêmica (graduado, especialista, mestre, doutor ou pós-doutor).
hasLevel	*Taught_Lessons / Academic_Formation*	Disciplinas são ministradas na graduação, especialização ou mestrado e doutorado.
hasTeachingActivity	*Research / Taught_Lessons*	Um pesquisador ministra disciplinas.
hasResearchProject	*Researcher / Research_Project*	Pesquisador faz parte de um projeto de pesquisa.
hasInstituition	*Researcher / Instituition*	Pesquisador faz parte de pelo menos uma instituição.
BibligraficProductionL	*Bibligrafic_Production /*	Uma produção bibliográfica

anguage	*L:Language*	tem um idioma.
Read	*Researcher / L:Language*	Pesquisador lê em determinado idioma.
Write	*Researcher / L:Language*	Pesquisador escreve em determinado idioma.
Speak	*Researcher / L:Language*	Pesquisador fala um idioma.
Understand	*Researcher / L:Language*	Pesquisador entende idioma.
EventCountry	*PaperInProceeding / C:Country*	País do evento onde o artigo foi publicado.

Tabela 4.2: Propriedades Datatype da OntoResearcher

Propriedades *Datatype*	Domínio/Escopo	Descrição
EventYear	*Paper / String*	Ano do evento onde o artigo foi publicado.
Email	*Researcher / String*	Pesquisador possui e-mail(s).
hasQualis	*PaperInJournal e PaperInProceeding / String*	Publicações possuem Qualis. Valores permitidos: A, B ou C.
Homepage	*Reseracher / String*	Pesquisador tem *homepage*.
finalPage	*PaperInJournal e PaperInProceeding / String*	Página Final de um *paper*.
initialPage	*PaperInJournal e PaperInProceeding / String*	Página inicial de um *paper*.
Name	*Researcher, Intitution e Taught_Lessons / String*	Nome de pesquisador, instituição e disciplinas ministradas.
hasBookTitle	*Chapter_Book / String*	Capítulo de livro tem o título do livro ao qual pertence.
hasJournalTitle	*PaperInJournal / String*	*Paper* em *Journal* tem o título do *Journal* onde foi publicado.
DOI	*PaperInProceeding e PaperInJournal / String*	DOI (sistema de identificação de objetos digitais).
ConclusionYear	*Academic_Formation e Research_Project / String*	Formação acadêmica e projeto de pesquisa têm ano de conclusão.
Citation	*Bibliografic_Production / int*	Produções bibliográficas possuem citações.
hasProceedingsTitle	*PaperInProceeding / String*	*Paper* em evento possui o título do *Proceding*.
hasTitle	*Bibliografic_Production, Academic_Formation e Research_Project / String*	Título da publicação, do trabalho de diplomação e dos projetos de pesquisa.
hasISBN	*Book, Chapter_Book / String*	O número identificador ISBN.
hasISSN	*Book, Chapter_Book e PaperInJournal/ String*	O número identificador ISSN.
hasEvent	*PaperInProceeding / String*	*Paper* é publicado em evento.

4.3 Definição dos indicadores de qualidade e do cálculo das qualificações

Após a definição das informações que fazem parte da OntoResearcher, foram definidos os critérios que servem de base para o cálculo da qualificação do pesquisador. Tais critérios foram ponderados para representar sua importância em relação aos demais critérios considerados para a qualificação. Os critérios e seus respectivos pesos (ou impactos) foram definidos de acordo com os trabalhos de Cazella (2006) e Rech (2007). Os valores obtidos nos trabalhos de Cazella e Rech são apresentados no Anexo A, desta dissertação.

A abordagem adotada para a definição dos pesos foi a MAUT (*Multi-Attribute Utility Theory* ou, em português, Teoria de Utilidade Multiatributo). Esta abordagem requer a representação das preferências de quem julga para cada critério. O julgador deve, a partir de seu próprio julgamento, confrontar os diferentes critérios, definindo limites de perdas para os demais ao optar por um critério específico. O resultado desta avaliação é a ordenação de todos os critérios, de acordo com sua importância. Para efetuar tal avaliação existem técnicas, que servem para que quem julga possa traduzir seus próprios julgamentos de valor em uma informação objetiva. Algumas técnicas de atribuição de pesos aos critérios são: SMART (*Simple Multi-Attribute Rating Technique*); Métodos Ordinais; AHP (*Analytic Hierarchy Process*); Atribuição Direta de Peso ou Pontuação Direta (*Direct Rating*); *Swing Weighting*; e *Trade-off Weighting* (BORCHERDING et al., 1991). O procedimento adotado nesta dissertação foi o *Swing Weighting* e para a tomada de decisão de importância de cada um dos indicadores foram utilizadas as definições apresentadas no Anexo A, ou seja, os valores dos trabalhos de Cazella e Rech foram os julgadores na abordagem adotada.

A técnica *Swing Weighting* funciona da seguinte maneira: primeiramente defini-se uma situação hipotética, caracterizada como sendo a pior hipótese possível, onde todos os critérios/sub-critérios tenham a pior avaliação possível. Depois da definição do pior cenário, o julgador decide qual dos sub-critérios é mais importante e assim sucessivamente para todos os critérios. Usando esta técnica, quem está julgando expressa suas preferências com relação aos critérios, partindo da pior hipótese possível, na qual todos os critérios obtêm a pior classificação (BORCHERDING et al., 1991). Os cálculos realizados para obter os pesos são apresentados no Anexo B. Os indicadores considerados para o cálculo bem como a respectiva ponderação de cada um são apresentados na Tabela 4.3.

Tabela 4.3: Indicadores considerados e respectiva ponderação

Categoria /Importância	Indicador	Impacto
Formação acadêmica (14,63%)	Pós-Doutorado	4,64%
	Doutor	3,96%
	Mestre	2,78%
	Especialista	1,86%
	Graduado	1,39%
Publicações (24,43%)	Livro	6,26%
	Capítulo de Livro	4,18%
	Paper em *Journal*	7,95%
	Paper em *Proceeding*	6,04%
Citações das Publicações (12,19%)	Número de Citações	12,19%
Qualis (*Paper* em *Journal* e *Proceedings* e das Conferências que o pesquisador é membro) (12,19%)	Qualis A	6,25%
	Qualis B	3,75%
	Qualis C	2,19%
Disciplinas Ministradas (10,97%)	Para Doutorado ou Mestrado	5,49%
	Especialização	3,29%
	Graduação	2,19%
Orientações Concluídas (9,75%)	Pós- Doutorado ou doutorado	4,48%
	Mestrado	2,93%
	Especialização	1,37%
	Graduação	0,97%
Participação em Projeto de Pesquisa (7,31 %)	Coordenador	4,09%
	Colaborador	3,22%
Membro de Comitê de Programa (8,53%)	É membro de Comitê de Programa de Conferências Científicas	8,53%

No critério "Formação Acadêmica" é considerado se o pesquisador possui "Graduação", "Especialização", "Mestrado", "Doutorado" e "Pós-Doutorado", cada um com seu respectivo impacto. Nos critérios "Publicações" ("Livro", "Capítulo de Livro", "*Paper* em *Journal*", "*Paper* em *Proceeding*") e "Membro de Comitê de Programa", são considerados, além de seus pesos, o peso do critério "Qualis". Por exemplo, se o pesquisador é membro de comitê de programa de uma conferência Qualis B, será considerado para o cálculo da qualificação o peso do critério "Membro de Comitê de Programa", 8,53%, e o peso do critério "Qualis B", 3,75%.

O critério "Citações das Publicações" considera o seu peso, 12,19%, multiplicado pela quantidade de citações de cada publicação para o cálculo das qualificações. O critério "Qualis" ("Qualis A", "Qualis B", "Qualis C") é utilizado para dar mais valor às publicações e à atuação do pesquisador como membro de comitê de programa. O critério "Disciplinas Ministradas" leva em conta o nível para o qual a disciplina é ministrada: "Doutorado", "Mestrado", "Especialização" e "Graduação".

O critério "Orientações Concluídas" considera o nível da orientação: "Pós-Doutorado ou Doutorado", "Mestrado", "Especialização" e "Graduação", e o critério "Projeto de Pesquisa" considera o papel do pesquisador no projeto, se ele é "Coordenador" ou "Colaborador".

Após a definição dos critérios e seus respectivos pesos, foi definido o cálculo das qualificações (CQ) que consiste de uma média aritmética ponderada representada pela Equação 2.

$$CQ = \frac{\sum_{i=1}^{n} ind_i * p_i}{\sum_{i=1}^{n} p_i} \qquad (2)$$

Na Equação 2, $\sum_{i=1}^{n} ind_i * p_i$ indica o somatório de todos os indicadores multiplicados pelos seus respectivos pesos e $\sum_{i=1}^{n} p_i$ indica o somatório do peso de todos os indicadores.

Este cálculo é efetuado para cada uma das áreas em que o pesquisador atua. Por exemplo, se um pesquisador possui na área de "I.3_COMPUTER_GRAPHICS", com: graduação (peso 1,39%), mestrado (peso 2,78%), uma publicação de *paper* em *proceeding* (peso 6,04%) classificado como Qualis A (peso 6,25%) e com 10 citações (peso 12,19%), o seu CQ será calculado como segue:

$$CQ = \frac{(1*1{,}39 + 1*2{,}78 + 1*6{,}04 * 6{,}25 * 10 * 12{,}19)}{1{,}39 + 2{,}78 + 6{,}04 + 6.25 + 12{,}19} = 160{,}76$$

O valor encontrado no cálculo do exemplo anterior (CQ= 160,76) representa o quanto o pesquisador é qualificado na área, caso o pesquisador não possua qualificação em outra área, este valor corresponderá à 100% de sua qualificação. Caso o pesquisador possua qualificação em outras áreas, os valores do CQ de cada área serão somados e o total será 100%, então para encontrar a porcentagem de atuação em cada área é aplicada uma regra de três simples.

4.4 Descrição das funcionalidades do sistema

O sistema web para qualificação de pesquisadores foi projetado de acordo com a arquitetura apresentada na Figura 4.3. Este sistema foi desenvolvido para obter as informações de diferentes fontes, popular essas informações na ontologia OntoResearcher e então calcular as qualificações dos pesquisadores nas áreas da Ciência da Computação.

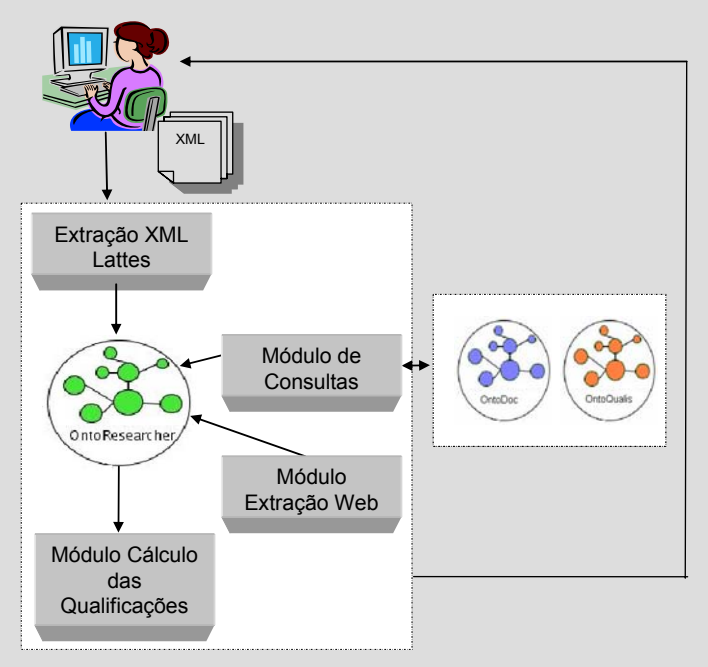

Figura 4.3: Arquitetura do sistema (Modificado de HANNEL; LIMA, 2007).

De acordo com a Figura 4.3, a entrada do sistema consiste no envio do XML do currículo Lattes do pesquisador através de um navegador Web. Após o envio do currículo Lattes o "Módulo Extração XML Lattes" é executado. Este módulo é responsável pela extração das informações do currículo (formação acadêmica, disciplinas ministradas, idiomas, instituição, país, projetos de pesquisa e produção bibliográfica) e população destas informações na OntoResearcher.

Após as informações do currículo serem populadas na OntoResearcher, o "Módulo de Consultas" efetua as consultas necessárias às ontologias OntoDoc (dissertação de mestrado em andamento) e OntoQualis (SOUTO et al., 2007) e popula tais informações na OntoResearcher. São realizadas consultas à ontologia OntoQualis para saber o Qualis de todas as conferências em que o pesquisador publicou e também das quais ele foi membro do comitê de programa. E à ontologia OntoDoc são feitas consultas sobre a área das disciplinas ministradas, orientações, formações acadêmicas, projetos de pesquisa e publicações.

O "Módulo Extração Web" é responsável por extrair o número de citações para cada uma das publicações do pesquisador, através de consultas realizadas ao site Google Scholar. Neste módulo também é feita uma análise de similaridade para saber se as citações retornadas do Google Scholar são mesmo do pesquisador. Apenas as informações similares são populadas na ontologia OntoResearcher. O "Módulo Cálculo das Qualificações" consiste da implementação de técnicas para pontuar as qualificações

dos pesquisadores em cada uma das áreas em que ele atua. Todos os módulos descritos nesta seção serão detalhados na seção 4.5.

4.5 Implementações

Arquitetura do sistema apresentada na seção 4.4 norteou a implementação do protótipo do sistema desenvolvido. O protótipo automatiza aspectos relacionados à obtenção das informações necessárias para qualificar a atuação dos pesquisadores através da integração das diferentes bases de informações. Permitindo que os pesquisadores verifiquem a distribuição de sua atuação nas áreas da Ciência da Computação.

O protótipo do sistema foi desenvolvido para execução em ambiente Web. A entrada do sistema consiste de uma página Web com uma breve descrição das funcionalidades, que é apresentado na Figura 4.4.

Figura 4.4: Página inicial do sistema.

Clicando no link cadastro, o pesquisador realiza um cadastro simples, apenas seu e-mail, uma senha e envia o seu currículo Lattes no formato XML, como apresentado na Figura 4.5. Este cadastro serve para garantir que cada pesquisador só tenha acesso as informações do seu próprio perfil é apresentado na Figura. Além disso, como trabalho futuro, a qualificação dos pesquisadores por área de atuação será enviada para o e-mail cadastrado na entrada no sistema.

Figura 4.5: Página de cadastro.

As próximas seções detalham a implementação dos módulos do sistema. Tais módulos foram previamente descritos na seção 4.3.

4.5.1 Extração XML do Lattes

Para extrair os dados do XML do Lattes e popular a ontologia foi realizado um mapeamento entre os elementos do XML e os conceitos da ontologia. O mapeamento é a análise para identificação das tags do XML do Lattes e a correspondência dessas com as classes e propriedades da OntoResearcher. Para ilustrar o mapeamento apresenta-se o exemplo dos dados referentes à formação acadêmica de doutorado de um pesquisador. A Figura 4.6 mostra um trecho do código XML do Lattes com os elementos referentes à formação acadêmica de doutorado em destaque.

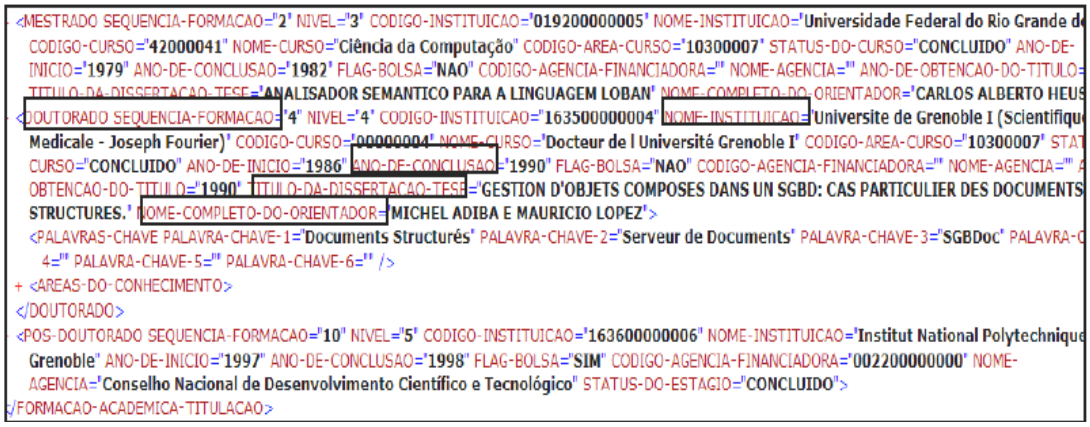

Figura 4.6: Trecho do currículo Lattes referente à formação acadêmica de doutorado.

Os dados referentes à formação acadêmica (nível doutorado) de um pesquisador, destacados na Figura 4.6, foram mapeados para a OntoResearcher. A Tabela 4.4 mostra esse mapeamento dos elementos do XML do Lattes para a classe *Doctor* (subclasse de *Academic_Formation*) e suas propriedades.

Tabela 4.4: Exemplo de mapeamento das informações do XML do Lattes para a OntoResearcher

***Tag* do XML**	**OntoResearcher**	**Observações**
DOUTORADO-SEQUENCIA-DE-FORMAÇÃO	Classe *Academic_Formation* subclasse *Doctor*	*Tag* que identifica que o pesquisador possui doutorado.
NOME-INSTITUIÇÃO	Propriedade *WasFinishedIn*	Instituição onde realizou doutorado.
ANO-CONCLUSÃO	Propriedade *ConclusionYear*	Ano de conclusão do doutorado.
TITULO-DA-DISSERTAÇÃO-TESE	Propriedade *hasTitle*	*Tag* referente ao título da tese.
NOME-COMPLETO-DO-ORIENTADOR	Propriedade *hasAdvisor*	Orientador da tese.

A extração automática dos dados contidos no currículo Lattes dos pesquisadores foi implementada em Java utilizando a API DOM (Document Object Model)[37]. O DOM cria uma estrutura de representação em árvore, onde todos os elementos do XML são nodos, pelos quais é possível navegar. Os nodos têm um relacionamento hierárquico entre eles, sendo que os termos *parent* e *child* (pai e filho, respectivamente) são usados para descrever esses relacionamentos. Alguns nodos podem ter nodos *child* ou serem nodos *leaf* (folha). Como os dados XML são estruturados em forma de árvore, é possível percorrer esta árvore sem conhecer a estrutura exata, e sem saber que tipo de dados ela contém. O SAX (*Simple API for XML*)[38] não foi utilizado pois é uma API baseada em eventos e trata o documento XML como um fluxo contínuo de dados, não permitindo navegar pelos dados XML no sentido contrário. As informações extraídas do currículo Lattes são populadas na OntoResearcher através do *framework* Jena.

4.5.2 Módulo Extração Web

Este módulo foi desenvolvido para obter o número de citações de cada publicação de um pesquisador. As citações são obtidas do site Google Scholar[39] o qual é um sistema público que recolhe na Web as publicações e computa as citações para cada publicação recolhida. Segundo Rech (2007), as respostas do Google Scholar possuem sempre a mesma estrutura, as publicações do autor são retornadas em uma lista contendo várias "referências", cada referência é composta pelo título da publicação, abaixo o nome dos autores, local de publicação e ano, uma descrição da publicação e o número de citações, que está circulado na Figura 4.7.

[37] http://www.w3.org/DOM/

[38] http://www.saxproject.org/

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

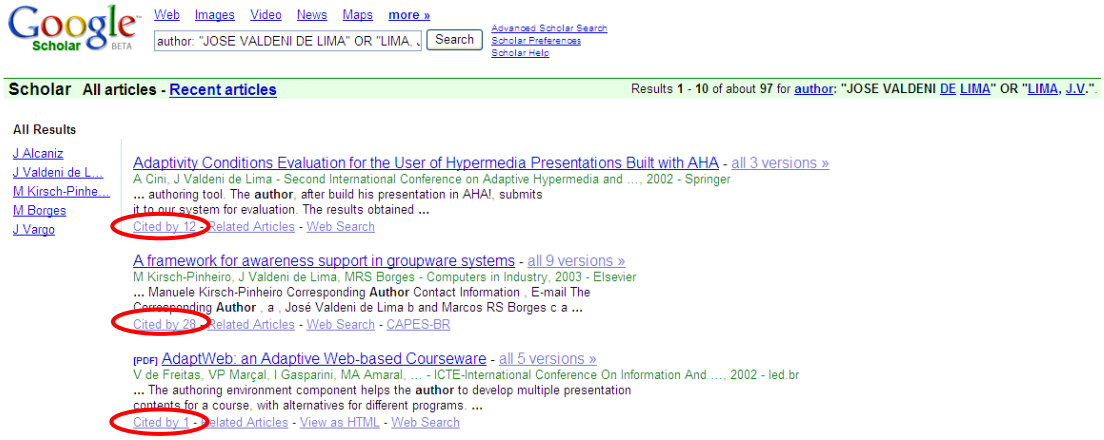

Figura 4.7: Consulta ao Google Scholar.

Para obter o número de citações das publicações dos pesquisadores o módulo de extração Web consulta o site Google Scholar e retorna para cada publicação o número de citações. Este módulo foi reusado do trabalho de Rech (2007), o qual utilizou a ferramenta Web-Harvest[40]. Esta ferramenta fornece uma API que permite consultar servidores Web, obter uma página HTML de resposta, transformá-la para XHTML (*eXtensible Hypertext Markup Language*) e aplicar tecnologias para manipulação de texto e de XML como XSLT (*eXtensible Stylesheet Language Transformations*), XQuery[41] (*XML Query Language*) e XPath (*XML Path Language*)[42]. O Web-Harvest foi configurado para receber a consulta (*query*) por parâmetro (extraída da *tag* do XML do Lattes NOME-EMCITAÇÕES-BIBLIOGRÁFICAS seguida da palavra "OR" e da informação extraída da *tag* do Lattes NOME-COMPLETO), obter as 10 primeiras páginas HTML retornadas pelo Google Scholar, e, para cada resultado retornado, extrair o título, nome dos autores e número de citações. Tanto a consulta como o número de páginas retornadas são valores configuráveis.

É possível que os títulos dos trabalhos do pesquisador não sejam exatamente os mesmos no currículo e no que foi retornado do Google Scholar. Por esta razão, após a extração das informações é utilizada uma função de similaridade. O objetivo da análise de similaridade é encontrar duas instâncias de dados (cadeias de caracteres: *strings*, árvores, etc.) que representam o mesmo objeto do mundo real (SILVA et al., 2006). Com esta análise de similaridade evita-se que citações que não são do pesquisador e foram retornadas pelo Google Scholar sejam atribuídas a ele. Partindo do estudo apresentado em Rech (2007), a função Smith-Waterman foi utilizada como função de similaridade, e o *threshold* (limiar) adotado foi 0,814 (RECH, 2007, p. 64-66).

No algoritmo Smith-Waterman (SMITH; WATERMAN, 1981) qualquer escore negativo é substituído por zero e o escore do alinhamento é o melhor escore dentre todos. Isto possibilita que nem o começo nem o fim das duas *strings* precisem estar alinhados. Considerando duas *strings* $A=a_1a_2..a_n$ e $B=b_1b_2...b_m$ a similaridade $s(a,b)$ é dada entre a seqüência de elementos a e b. Para encontrar os elementos com maior grau de similaridade é criada a matriz H de tamanho $(n + 1) \times (m + 1)$ que é calculada de

[40] http://web-harvest.sourceforge.net

[41] http://www.w3.org/XML/Query/

[42] http://www.w3.org/TR/xpath

acordo com a Equação 3. Onde $p(s_i, t_j)$ representa uma função de custo e g a penalidade de alinhamento com um *gap*. Os valores de $H[i,0]$ e $H[0, j]$ são inicializados com zero.

$$H[i, j] = \max \begin{cases} 0, \\ H[i-1, j-1] + p(s_i, t_j) \\ H[i-1, j] - g \\ H[i, j-1] - g \end{cases} \quad (3)$$

A implementação da função de similaridade adotada foi a da biblioteca SimMetrics[43]. As publicações similares são populadas na OntoResearcher usando o *framework* Jena.

4.5.3 Módulo de Consultas

Este módulo realiza consultas nas ontologias reusadas nesta dissertação, a *OntoQualis* (SOUTO et al., 2006; SOUTO et al., 2007) e a *OntoDoc* (dissertação de mestrado de Luis Henrique G. Oliveira, em andamento no contexto do projeto DIGITEX na UFRGS).

4.5.3.1 OntoQualis

Foram identificadas duas formas para obter o Qualis da conferências científicas, que são: usar a base de dados Qualis-Capes ou usar a OntoQualis. Uma limitação da base de dados Qualis-Capes foi apresentada por Rech:

> 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 (2007, p. 62).

Para contornar este problema apresentado por Rech, as conferências científicas foram classificadas pela OntoQualis. A OntoQualis visa (semi) automatizar o processo de classificação de conferências científicas de acordo com os critérios definidos pelo Comitê de Computação da Capes. O Qualis tem por objetivo classificar todos os veículos de publicação relatados pelos Cursos de Pós-Graduação, tais como *Journal*, Periódicos, Conferências Nacionais, Conferências Internacionais, etc. Cada veículo tem suas regras específicas e podem ser classificados em Tipo A, B, C ou D (não classificado). A cada ano/período, o conjunto de veículos poderá ser ajustado, assim como os critérios de avaliação poderão ser revistos para contemplar a evolução e as particularidades das subáreas da Ciência da Computação. O Qualis de Conferências Científicas subdivide-se em internacional e nacional, entretanto a OntoQualis foi modelada apenas as características de conferências internacionais (SOUTO et al., 2006).

Então, para obter os dados sobre o Qualis das publicações em conferências e para saber se um pesquisador faz parte de um comitê de programa de uma conferência científica foi desenvolvido um sistema de consultas à OntoQualis. Esse sistema de consultas foi desenvolvido em Java utilizando o *framework* Jena. Por exemplo, para

[43] http://www.dcs.shef.ac.uk/~sam/simmetrics.html

saber se o pesquisador "José Valdeni de Lima" é membro de algum comitê de programa, a consulta realizada é apresentada na Figura 4.8.

```
SELECT ?y
      WHERE (?x <www.inf.ufrgs.br/~khannel/Ontology/OntoResearcher#hasName> "José
                        Valdeni de Lima"^^xsd:string),
         (?x < www.inf.ufrgs.br/~khannel/Ontology/OntoQualis#isMemberOf> ?y)
```

Figura 4.8: Consulta para membro de comitê de programa.

4.5.3.2 OntoDoc

A *OntoDoc* representa o domínio dos artigos científicos utilizando para isso um conjunto de metadados do padrão *Dublin Core*[44]. A *Ontodoc* não é contribuição desta dissertação, ela faz parte da dissertação de mestrado de Luis H. G. Oliveira (em desenvolvimento no mestrado em computação da UFRGS). Através dos metadados modelados, é possível classificar documentos digitais nas áreas da Ciência da Computação de acordo com a classificação da ACM. A classe principal da *OntoDoc* é a *Document*, a qual possui propriedades equivalentes aos elementos do Dublin Core como: *title* (referente ao título de um documento), *date* (referente a data do documento), *description* (uma descrição do documento), *type* (referente ao tipo do documento), *format* (referente ao formato do documento), *creator* (referente ao(s) autor(es) do documento), *subject* (referente a área em que o documento se encaixa) e *language* (referente ao idioma do documento).

A obtenção das informações para popular a ontologia *OntoDoc* é efetuada, resumidamente, de acordo com a Figura 4.9.

Figura 4.9: Arquitetura do sistema de classificação de documentos digitais.

Na Figura 4.9 é possível observar que a entrada do sistema consiste em um documento no formato PDF. Esse documento passa por um processo de extração de texto e devolve um arquivo XML. A etapa de pré-processamento consiste em gerar um vetor de características para cada linha do artigo. Este vetor é composto por um conjunto de palavras mais um conjunto de características específicas de linhas, como quantidade de palavras na linha, posição da linha no texto. Esses vetores de características passam pelos classificadores SVM, que já foram previamente treinados, que identificam a área do documento de acordo com as áreas da ACM. E essa informação da área é populada na *OntoDoc*.

A informação sobre a área (das publicações, disciplinas ministradas, orientações, formação acadêmica e projetos de pesquisa) será obtida com consultas à ontologia *OntoDoc*. As consultas à *OntoDoc* também foram realizadas através do *framework* Jena. A Figura 4.10 mostra a consulta sobre a área da publicação "Increasing XML interoperability in Visual rewriting Systems".

[44] http://dublincore.org/

```
SELECT ?y
        WHERE (?x <www.inf.ufrgs.br/~khannel/Ontology/OntoResearcher#hasArea> "
 Increasing XML interoperability in Visual rewriting Systems"^^xsd:RA:ResearchArea),
        (?x < www.inf.ufrgs.br/~khannel/Ontology/OntoDoc#Area> ?y)
```

Figura 4.10: Consulta área de uma publicação.

As consultas são realizadas pelo título, como no exemplo da Figura 4.10 onde foi feita a consulta pelo título de uma publicação. Supõe-se que a *OntoDoc* tenha os mecanismos para obter as informações sobre tal publicação e assim possa classificá-la.

4.5.4 Módulo Cálculo das Qualificações

Este módulo realiza o cálculo das qualificações dos pesquisadores para cada uma das áreas em que ele atua. O cálculo consiste na aplicação da fórmula do CQ, $CQ = \frac{\sum_{i=1}^{n} ind_i * p_i}{\sum_{i=1}^{n} p_i}$, a qual foi apresentada na Equação 2 da seção 4.3.

Para o cálculo são considerados os indicadores de qualidade (apresentados na Tabela 4.3 da seção 4.3).

4.5.5 Tecnologias Utilizadas

Para a implementação do sistema foram utilizadas as tecnologias apresentadas na Tabela 4.5.

Tabela 4.5: Tecnologias utilizadas na implementação do sistema.

Categoria	Tecnologia	URL (acesso em: dez. 2007)
Framework Web Semântica	Jena- 2.4	http://jena.sourceforge.net/
Linguagem para Descrição de Ontologias	OWL-DL	http://www.w3.org/TR/owl-guide/
Ambiente para Editoração de Ontologia	Protégé-3.2.1	http://protege.stanford.edu/
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/
Framework para Extração de Dados do Google Scholar	Web-Harvest 0.3	http://web-harvest.sourceforge.net
Manipulação de Arquivos XML	DOM	http://www.w3.org/DOM/
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

4.5.6 Considerações

Durante a implementação do protótipo foram identificadas algumas dificuldades e limitações, como:

- No momento de popular a OntoResearcher não é feita uma análise de similaridade para verificar se a instância que está sendo populada já existe. Isto é, se a informação a ser populada já é uma instância da ontologia, porém tiver sido escrita de forma diferente, será populada novamente. Por exemplo, o pesquisador "José Valdeni de Lima" é uma instância da OntoResearcher, caso outro pesquisador tenha uma publicação com ele e tenha descrito em seu currículo "J.V. de Lima" serão criadas duas instâncias para o mesmo pesquisador. Como trabalho futuro é sugerida a implementação de uma técnica de similaridade para evitar esta repetição de informações.
- Como já havia sido identificado por Rech (2007, p. 61-62), o processo de extração de dados da Web utilizando a ferramenta Web-Harvest é dependente da lógica da página HTML, de modo que se a lógica do site Google Scholar for alterada será necessário alterar as configurações da ferramenta.

5 APLICAÇÃO E RESULTADOS

Este capítulo trata da aplicação do protótipo desenvolvido e dos resultados obtidos. A principal aplicação é o cálculo das qualificações por área de atuação dos pesquisadores da amostra utilizada. Além disso, apresenta-se as consultas realizadas na OntoResearcher a fim de descobrir novas informações sobre os pesquisadores e a criação de conglomerados (*clusters*) de pesquisadores.

5.1 Conjunto de Dados

Foram utilizados 12 currículos Lattes, no formato XML, de pesquisadores doutores da área da Ciência da Computação da UFRGS (Universidade Federal do Rio Grande do Sul). Os pesquisadores foram identificados pelo conjunto {P1, P2,...,P12}. Estes currículos foram os mesmos utilizados em Rech (2007). Destes 12 currículos foi obtido um total de 791 publicações que variam do ano 1974 a 2007. Para cada publicação foram realizadas consultas ao Google Scholar com "NOME-EMCITAÇÕES-BIBLIOGRÁFICAS" seguida da palavra "OR" e "NOME-COMPLETO" (ambos extraídos de da *tags* do XML do Lattes), como descrito previamente na seção 4.5.2.

Além das publicações, dos 12 currículos Lattes foram obtidas informações sobre as disciplinas ministradas, os projetos de pesquisa, as orientações concluídas, formação acadêmica e participação em comitê de programa. Para cada uma destas informações foram realizadas consultas à OntoDoc para obter a área. Para saber o Qualis das conferências em que o pesquisador foi membro do comitê de programa foram realizadas consultas a *OntoQualis*. Porém foram realizadas apenas algumas consultas a *OntoQualis* e a *OntoDoc* para validar o modelo proposto. Entretanto, como o volume de informações necessários para popular a *OntoQualis* e *OntoDo*c é grande, por uma questão de tempo, os experimentos foram realizados com informações (sobre o Qualis e as áreas) descobertas manualmente. O processo de descoberta manual das informações sobre o Qualis e as áreas ocorreu da seguinte forma:

- Para obter o Qualis: como o sistema Qualis, para conferências científicas, possui poucas conferências classificadas procedeu-se uma análise baseando-se nas regras Qualis-Capes. Para isso, é necessário obter as informações (na Web) sobre as conferências. Entretanto não foi possível encontrar informações sobre grande número de publicações anteriores a 2002. Assim, o escopo de obtenção do Qualis foi reduzido para os anos de 2002 a 2007.

- Para obter as áreas: foi efetuada uma análise manual das informações com base nas áreas, nas palavras-chave[45] e na biblioteca digital[46] da ACM. A partir do título das publicações, das disciplinas ministradas, dos trabalhos orientados, dos trabalhos da formação acadêmica e dos projetos de pesquisa descobrir as áreas dos mesmos. Caso o título das publicações (também dos trabalhos da formação acadêmica e dos trabalhos orientados) seja muito geral, impossibilitando a identificação da área, é analisado o *abstract* para identificar a área (isso se for possível encontrar o trabalho na Web). Para as disciplinas ministradas, quando o título não é suficiente para encontrar a área, analisa-se a súmula da disciplina. Se mesmo assim não for possível identificar a área, considera-se como NÃO_CLASSIFICADO;

Resumidamente, dos currículos Lattes foram extraídas as seguintes informações: "Formação Acadêmica" (3 pós-doutorados, 12 doutorados, 11 mestrados, 1 especialização e 14 graduações); "Publicações" (17 livros, 26 capítulos de livros, 79 *journal* e 701 *proceeding*); "Disciplinas Ministradas" (67 para mestrado e doutorado, 5 para especialização e 132 para graduação); "Orientações Concluídas" (21 pós-doutorado e doutorado, 10 especialização e 127 graduação) e "Projeto de Pesquisa" (22 coordenador e 47 colaborador). Do Google Scholar foram obtidas 4344 citações. Para obter o Qualis das publicações e das conferências em que o pesquisador é membro do comitê de programa, reduziu-se o escopo para os anos de 2002 a 2007. Obteve-se então 47 Qualis A, 16 Qualis B e 20 Qualis C. A classificação de áreas da ACM foi utilizada até o segundo nível de classificação o que resultou em 62 diferentes áreas.

Com os dados populados na OntoResearcher, foi aplicada realizar o cálculo das qualificações para todas as áreas em que o pesquisador atua, que será descrita na próxima seção.

5.2 Cálculo das Qualificações

O cálculo das qualificações consiste da aplicação da Equação 2 descrita na seção 4.3 para cada um das áreas de atuação identificada para os pesquisadores. A abordagem adotada nesta dissertação é a de apresentar a porcentagem de atuação do pesquisador em cada área. Salienta-se que as áreas utilizadas são as da ACM, então para verificar quais são as áreas e subáreas é necessário acessar o *Computing Classification System*[47] da ACM.

Para facilitar a visualização, as áreas que têm menos de 1% de atuação foram agrupadas. As Figuras 5.1 a 5.12 apresentam os gráficos das áreas em que atuam, respectivamente, para cada um dos 12 pesquisadores. O Anexo C desta dissertação apresenta os valores obtidos para todas as áreas em que os pesquisadores atuam.

[45] http://portal.acm.org/subjects.cfm?part=author&coll=GUIDE&dl=GUIDE&CFID=51237847&CFTOKEN=12010035

[46] http://portal.acm.org/dl.cfm

[47] http://portal.acm.org/ccs.cfm?part=author&coll=GUIDE&dl=GUIDE&CFID=50872872&CFTOKEN=65714136

O cálculo das qualificações do pesquisador 1 é apresentado na Figura 5.1. Pode-se verificar que a área em que o pesquisador mais atua é H.2_DATABASE_MANAGEMENT, na qual obteve 71,70% de atuação.

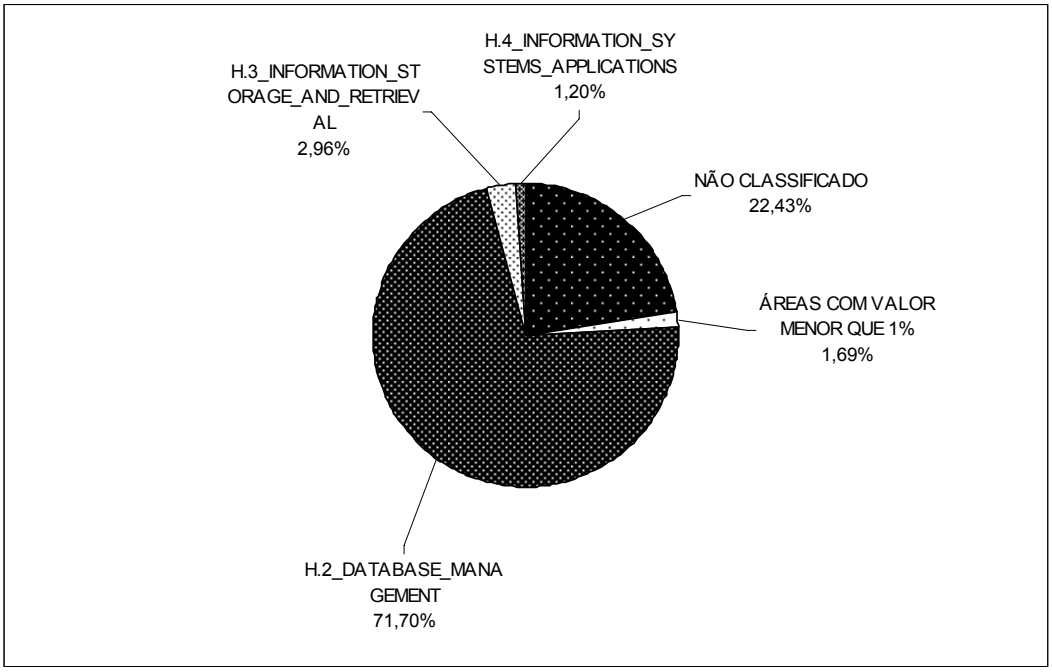

Figura 5.1: Gráfico para o pesquisador 1.

A Figura 5.2 apresenta o cálculo das qualificações para o pesquisador P2, o qual possui maior atuação na área K.3_COMPUTERS_AND_EDUCATION.

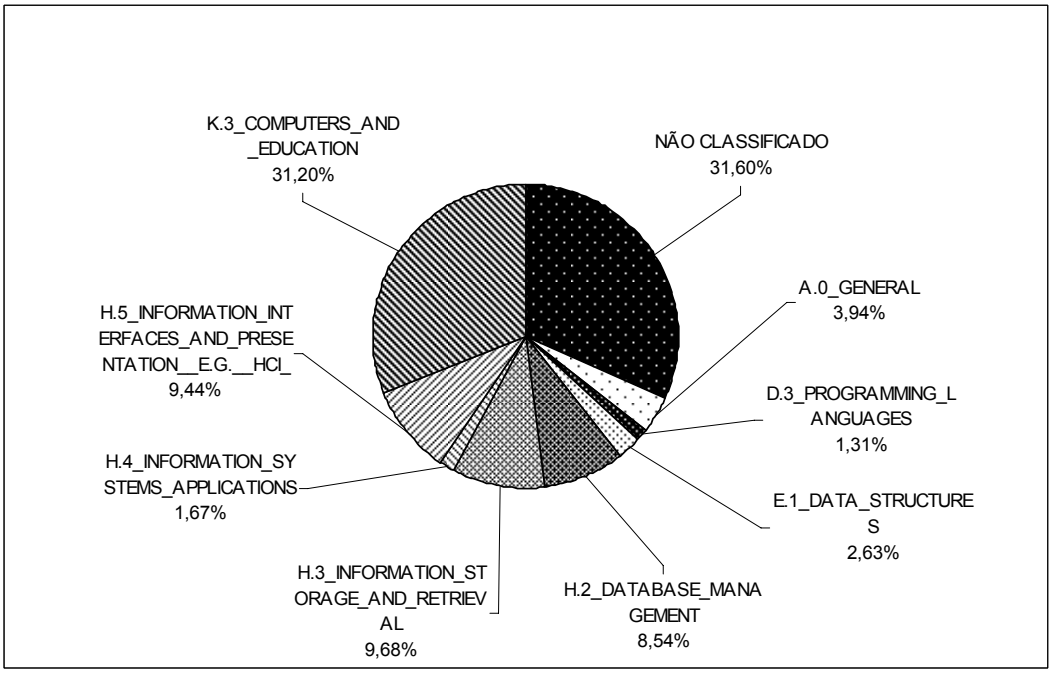

Figura 5.2: Gráfico para o pesquisador 2.

O pesquisador P3 possui 61,31% de sua atuação na área I.3_COMPUTER_GRAPHICS, como pode ser visto na Figura 5.3.

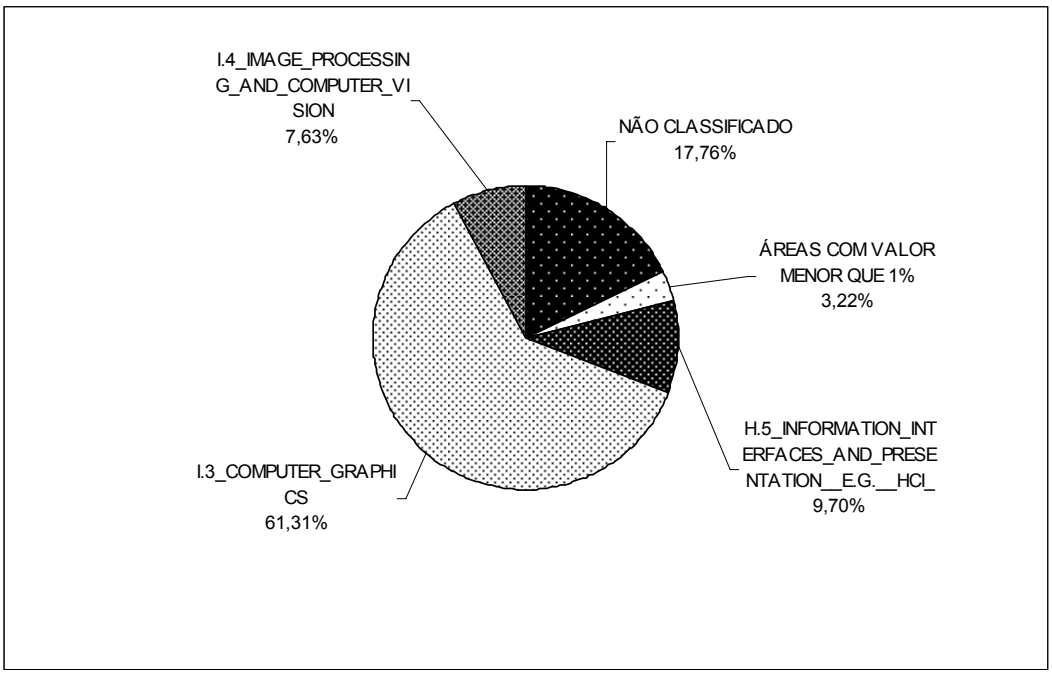

Figura 5.3: Gráfico para o pesquisador 3.

O pesquisador P4 se caracteriza pela atuação distribuída em diversas áreas como apresentado na Figura 5.4. A área em que o P4 mais atua é H.2_DATABASE_MANAGEMENT com 9,49%.

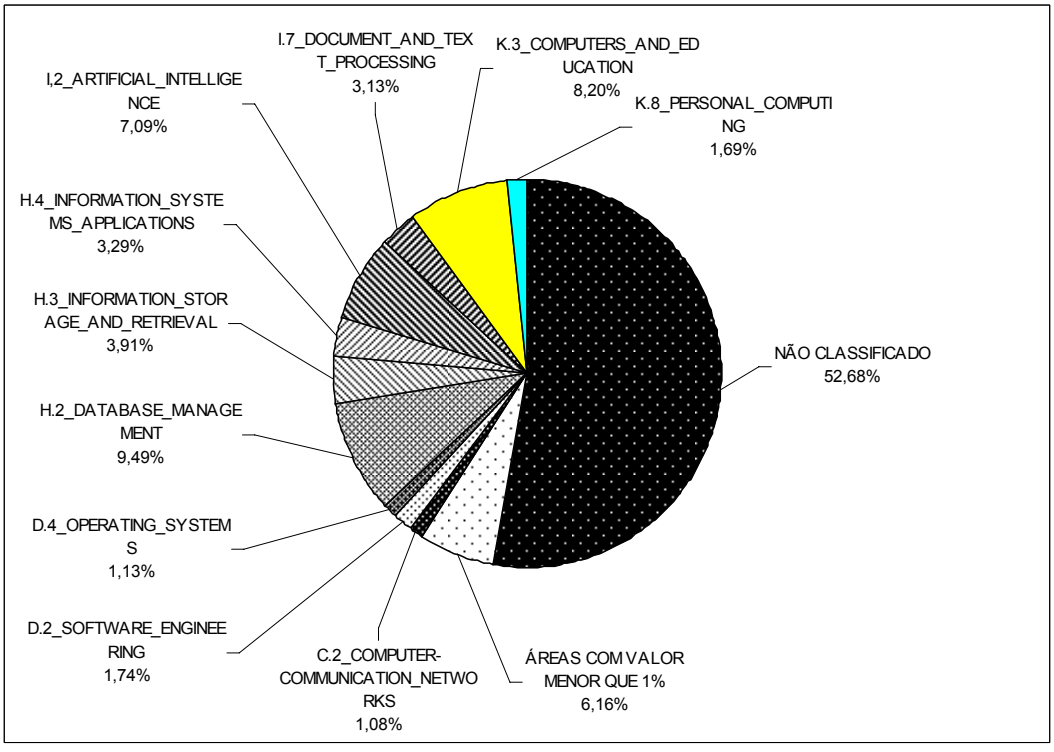

Figura 5.4: Gráfico para o pesquisador 4.

O cálculo das qualificações do pesquisador P5 é apresentado na Figura 5.5. O P5 atua com 40,48% na área de H.5_INFORMATION_INTERFACES_AND_PRESENATTION_E.G._HCI_.

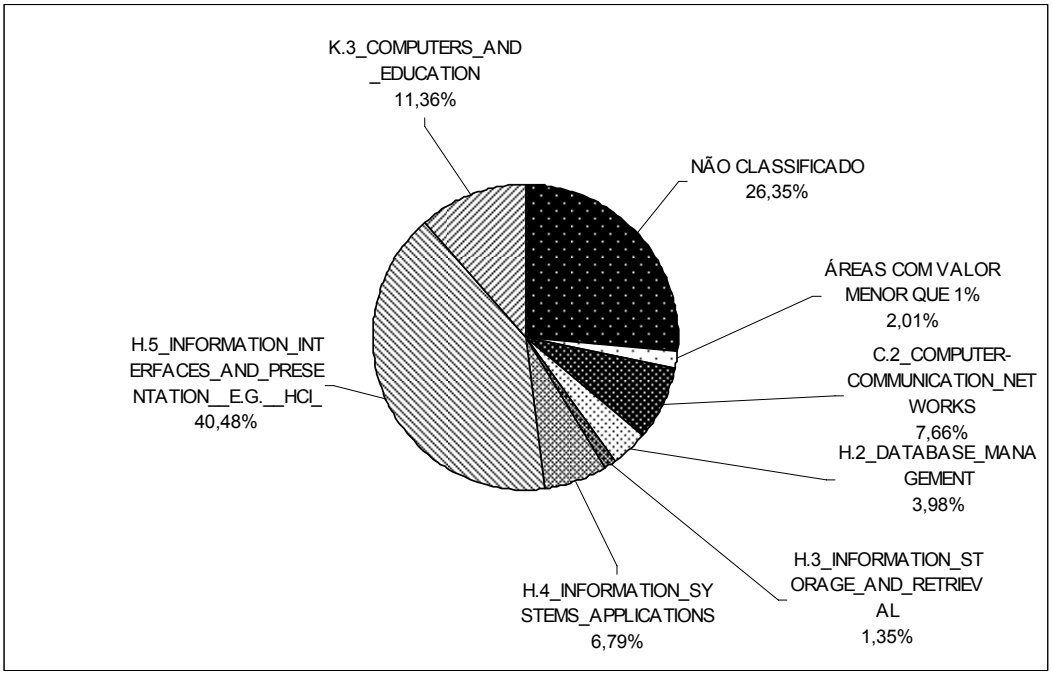

Figura 5.5: Gráfico para o pesquisador 5.

A Figura 5.6 apresenta a distribuição da atuação do pesquisador P6. O P6 tem 20,49% de sua atuação na área H.3_INFORMATION_STORAGE_AND_RETRIEVAL.

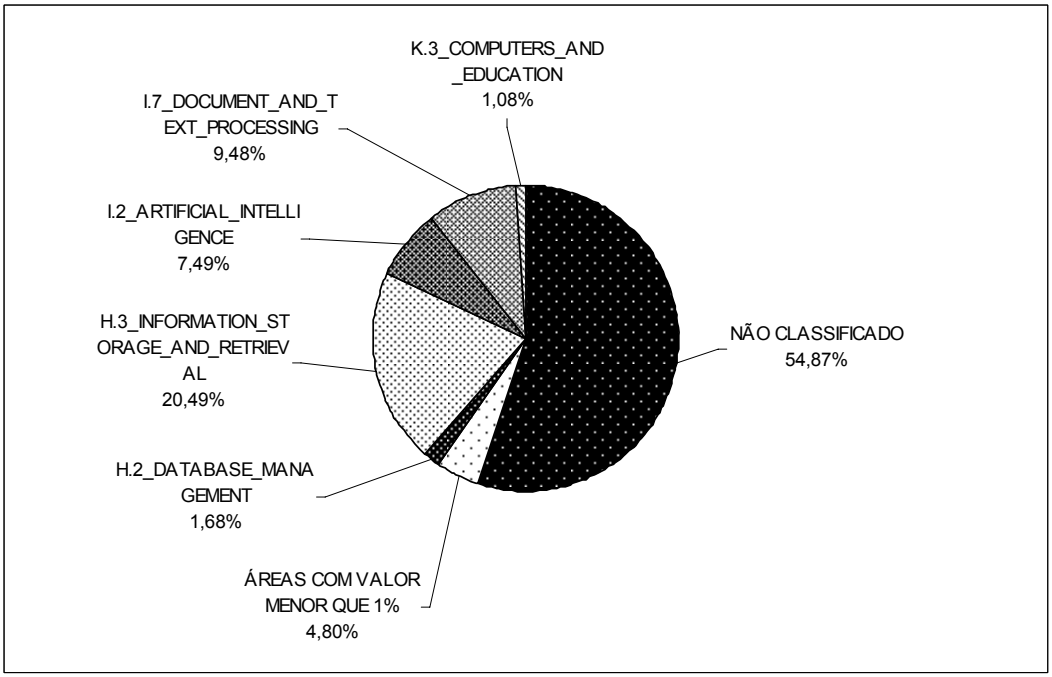

Figura 5.6: Gráfico para o pesquisador 6.

A Figura 5.7 apresenta a distribuição das áreas de atuação do pesquisador P7. Como pode ser observado, P7 possui 57,10% de sua atuação na área H.2_DATABASE_MANAGEMENT.

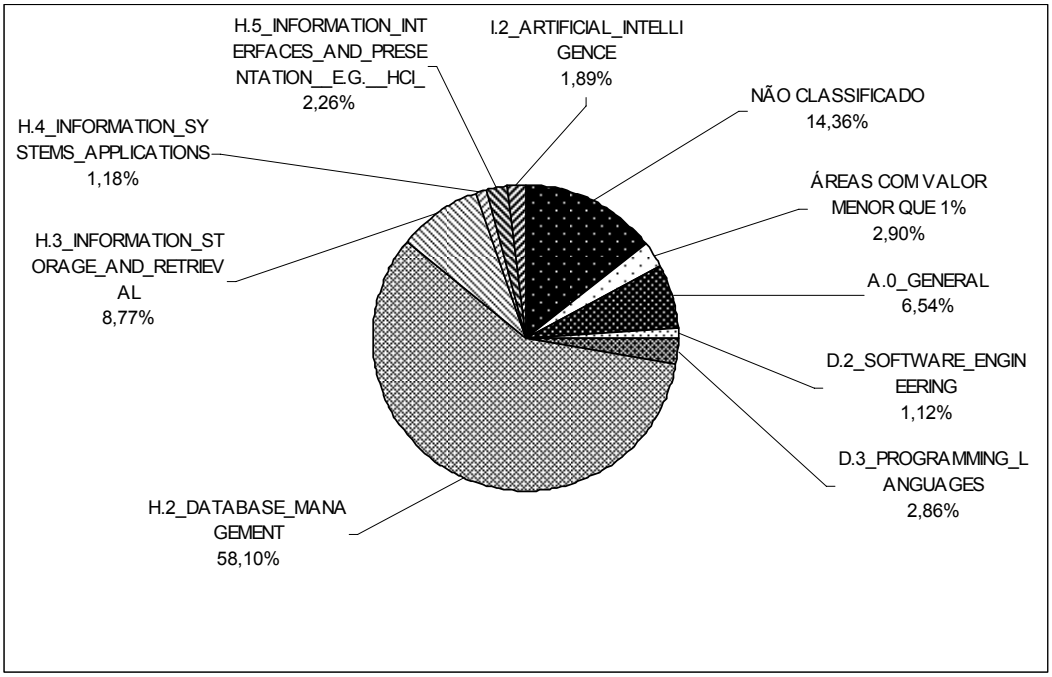

Figura 5.7: Gráfico para o pesquisador 7.

A Figura 5.8 mostra a distribuição da atuação do pesquisador P8. O pesquisador P8 possui 20,32% de sua atuação na área H.2_DATABASE_MANAGEMENT.

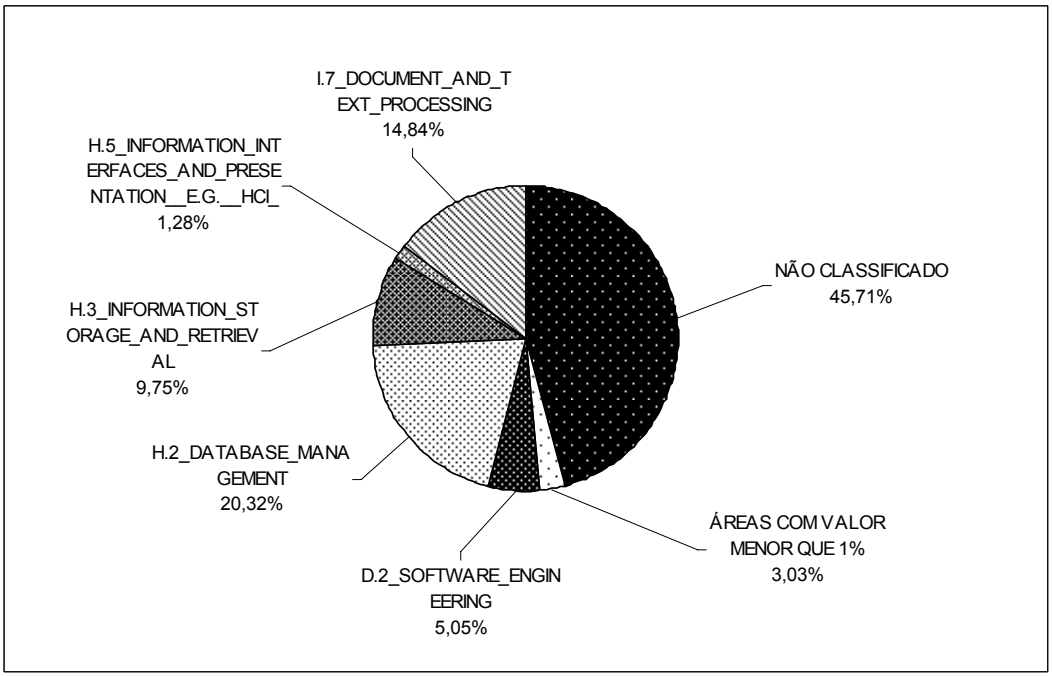

Figura 5.8: Gráfico para o pesquisador 8.

A Figura 5.9 apresenta a distribuição da atuação do pesquisador P9. Como pode ser observado, o P9 possui 60,92% de sua atuação na área C.2_COMPUTER-COMMUNICATION_NETWORKS.

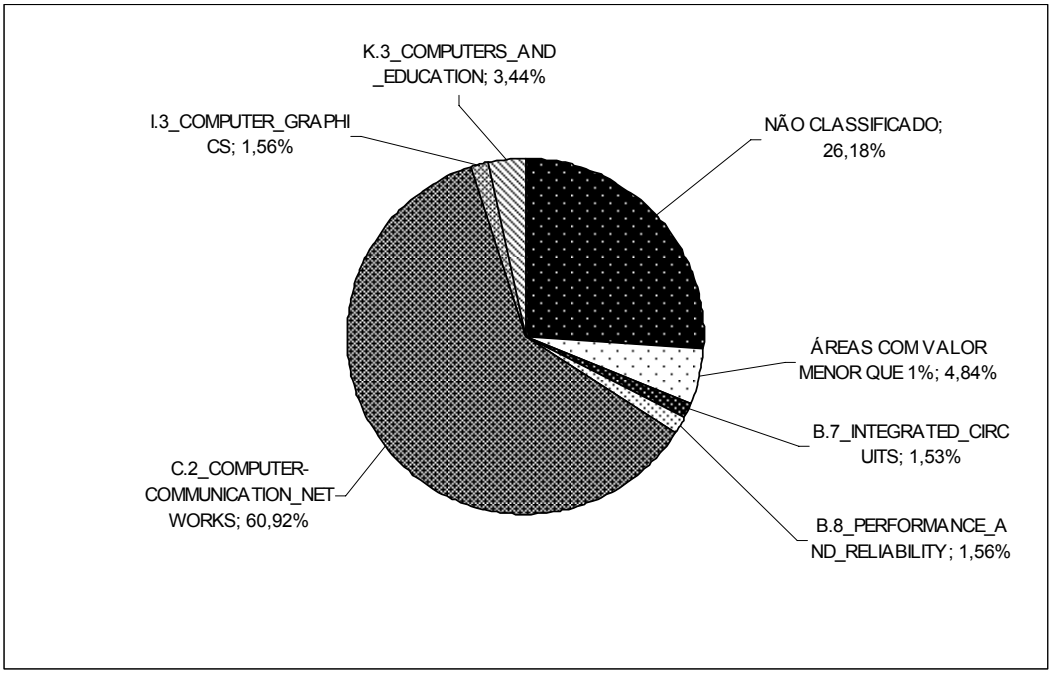

Figura 5.9: Gráfico para o pesquisador 9.

A Figura 5.10 mostra a distribuição da atuação do pesquisador P10. Como pode ser observado, o P10 possui 42,67% de sua atuação na área I.3_COMPUTER_GRAPHICS.

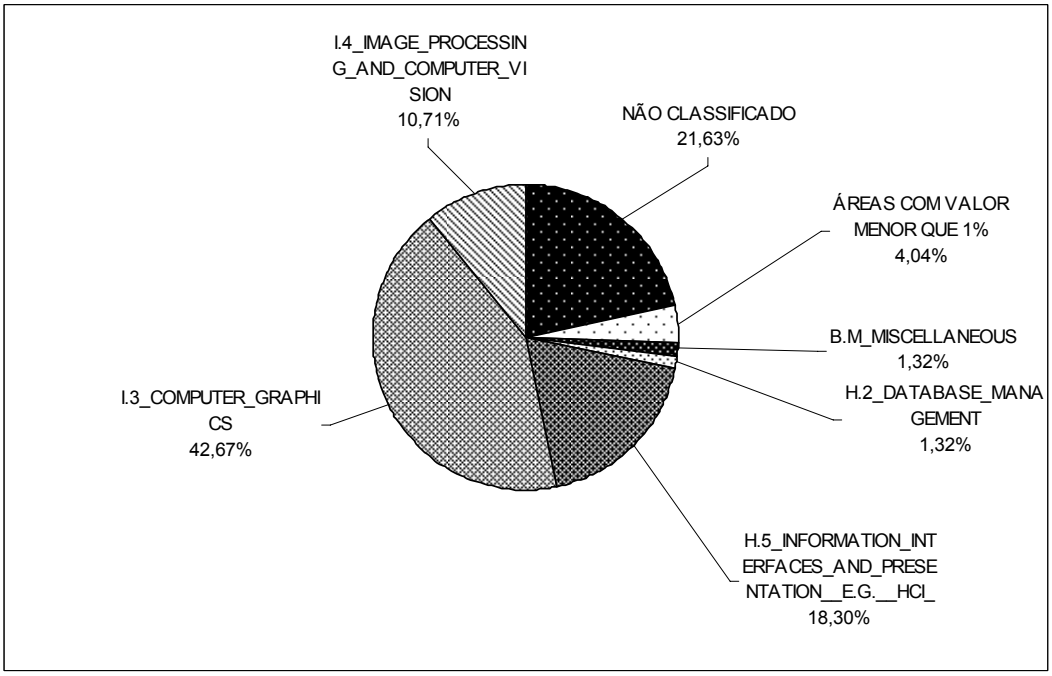

Figura 5.10: Gráfico para o pesquisador 10.

A Figura 5.11 mostra a distribuição da atuação do pesquisador P11. Como pode ser observado, o P11 possui 89,25% de sua atuação na área B.8_PERFORMANCE_AND_RELIABILITY.

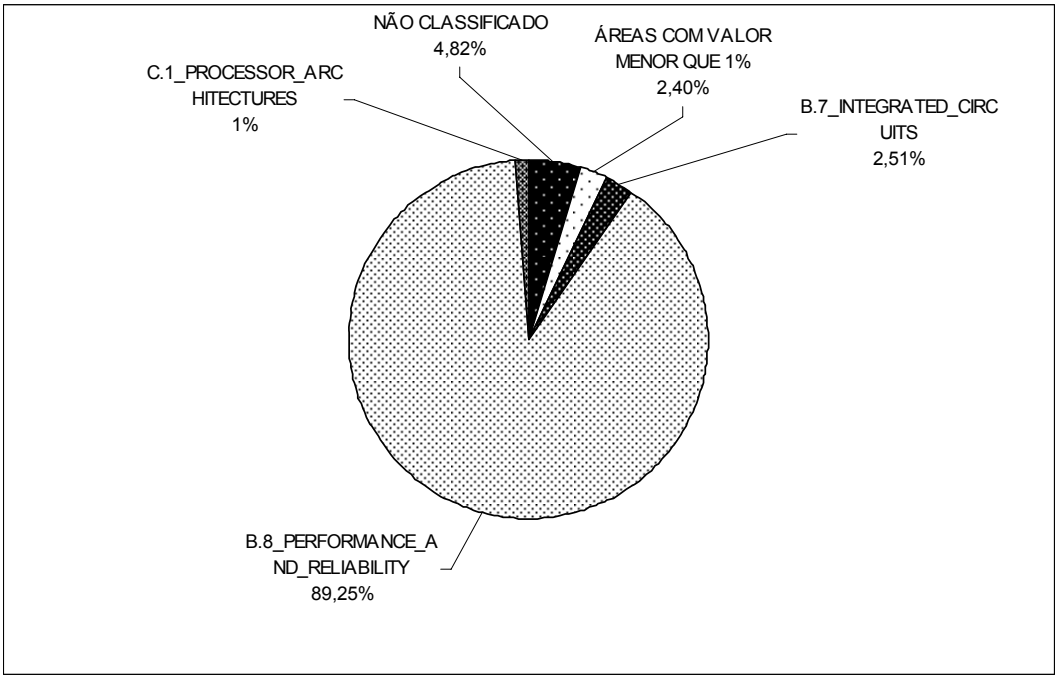

Figura 5.11: Gráfico para o pesquisador 11.

A Figura 5.12 mostra a distribuição da atuação do pesquisador P12. Como pode ser observado, o P12 possui 20,91% de sua atuação na área I.2_ARTIFICIAL_INTELLIGENCE.

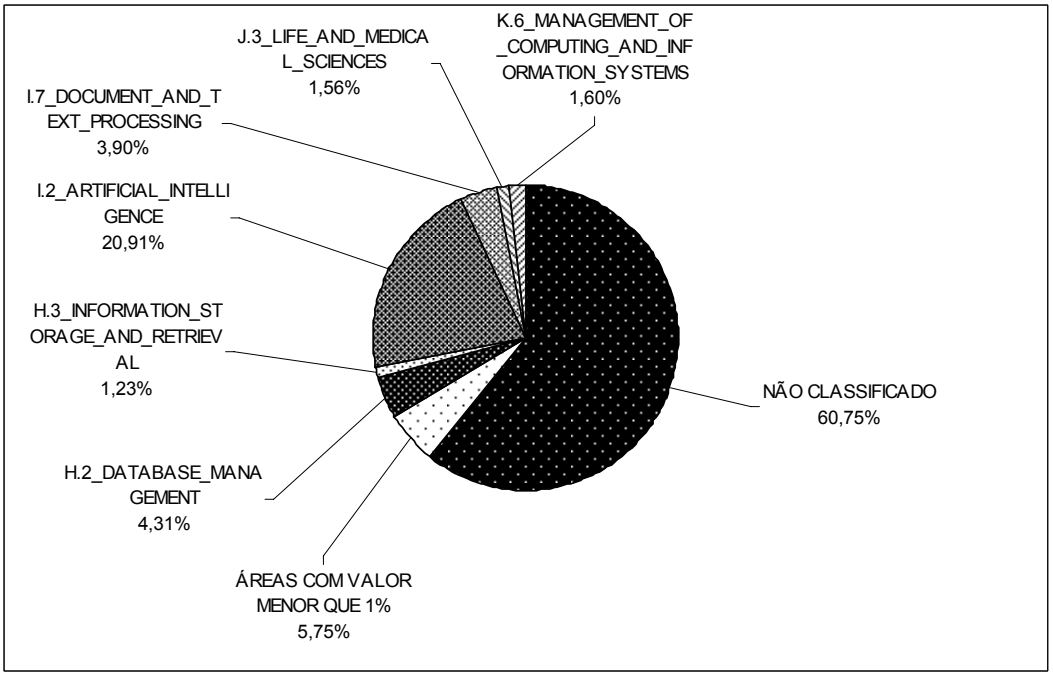

Figura 5.12: Gráfico para o pesquisador 12.

Todos os pesquisadores obtiveram tiveram informações que não puderam ser enquadradas em nenhuma das áreas, e por isso foram consideradas como "NÃO_CLASSIFICADA". Foi efetuada uma análise manual para verificar o motivo de não encontrar a área de algumas informações. Na maioria das vezes, isto ocorre por razões, como:

- O título da publicação, disciplina ministrada, etc. estão descritos no Lattes de modo abreviado ou sem informações adicionais que permitam obter a área;
- Alguns ministram disciplinas como "Tópicos Especiais em Computação", que não podem ser classificadas, pois o tema abordado varia e não há uma descrição de tais temas no currículo do pesquisador;
- Alguns trabalhos não podem ser encontrados na Web e por isso não se tem acesso ao *abstract* para tentar inferir a área;
- Alguns pesquisadores têm publicações fora do escopo da Computação.

O objetivo desta dissertação, de qualificar os pesquisadores nas áreas da Ciência da Computação, foi alcançado com a realização desta aplicação. Demonstrando assim, a viabilidade do modelo proposto.

5.2.1 Comparação com outros trabalhos

A qualificação dos pesquisadores descrita nesta dissertação foi comparada com o trabalho de Rech (2007). Em Rech, o cálculo das competências foi aplicado sem a especificação das áreas. Rech calcula dois coeficientes, o *CCc* (competências referentes ao currículo dos pesquisadores) e o *CCb* (indicadores quantitativos das relacionados à produção bibliográfica). Após, aplica a junção dos indicadores e respectiva atribuição dos pesos gerando o coeficiente *CC*. Como nesta dissertação os indicadores qualitativos e quantitativos são agrupados no Cálculo das Qualificações, o *CQ* (descrito na Equação 2 da seção 4.3) o resultado obtido será comparado com o *CC* de Rech.

A Tabela 5.1 apresenta o *ranking* dos pesquisadores gerado a partir do cálculo do *CQ* e a Tabela 5.2 apresenta o *ranking* dos pesquisadores obtidos no cálculo do *CC* por Rech, ambos foram criados apenas ordenando de forma decrescente os valores obtidos no cálculo do *CQ* e do *CC*, respectivamente.

Tabela 5.1: *Ranking* dos pesquisadores usando cálculo das qualificações *CQ*.

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

Tabela 5.2: *Ranking* dos pesquisadores pelo *CC*.

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

Fonte: RECH, 2007.

Como os indicadores utilizados são, em sua grande maioria, os mesmos já era esperado que não ocorressem muitas diferenças entre os dois *rankings*. De fato, ambas as abordagens não divergem em relação à competência dos pesquisadores. Sendo que, a identificação das áreas dos pesquisadores é um diferencial desta dissertação.

5.3 Descoberta de Conhecimento sobre os Perfis

Uma das grandes vantagens do uso de ontologias é a possibilidade de descobrir novas informações sobre o domínio em questão, no caso o perfil dos pesquisadores. Para realizar as consultas utilizou-se a linguagem RDQL que trabalha sobre os modelos do *framework* Jena. As consultas realizadas não servirão para o cálculo das qualificações dos pesquisadores, e sim para aprimorar o perfil dos pesquisadores.

5.3.1 Co-autoria

Esta consulta é justificada pela necessidade de conhecer quais pesquisadores publicaram juntos. Por exemplo, para o pesquisador "José Valdeni de Lima", a consulta realizada é apresentada na Figura 5.14.

```
SELECT ?w
WHERE (?z prop:Name "Jose Valdeni de Lima"),
       (?y prop:hasAuthor ?z),
       (?y prop:hasAuthor ?x)
AND (?x ne ?z)
USING prop FOR <http://www.inf.ufrgs.br/~khannel/Ontology#>
```

Figura 5.14: Exemplo de consulta sobre os co-autores.

A consulta da Figura 5.14 gera uma lista de pesquisadores que tiveram alguma publicação em conjunto com "José Valdeni de Lima". Para fins de visualização, a lista de co-autores obtida foi organizada de forma gráfica e apresentada na Figura 5.15.

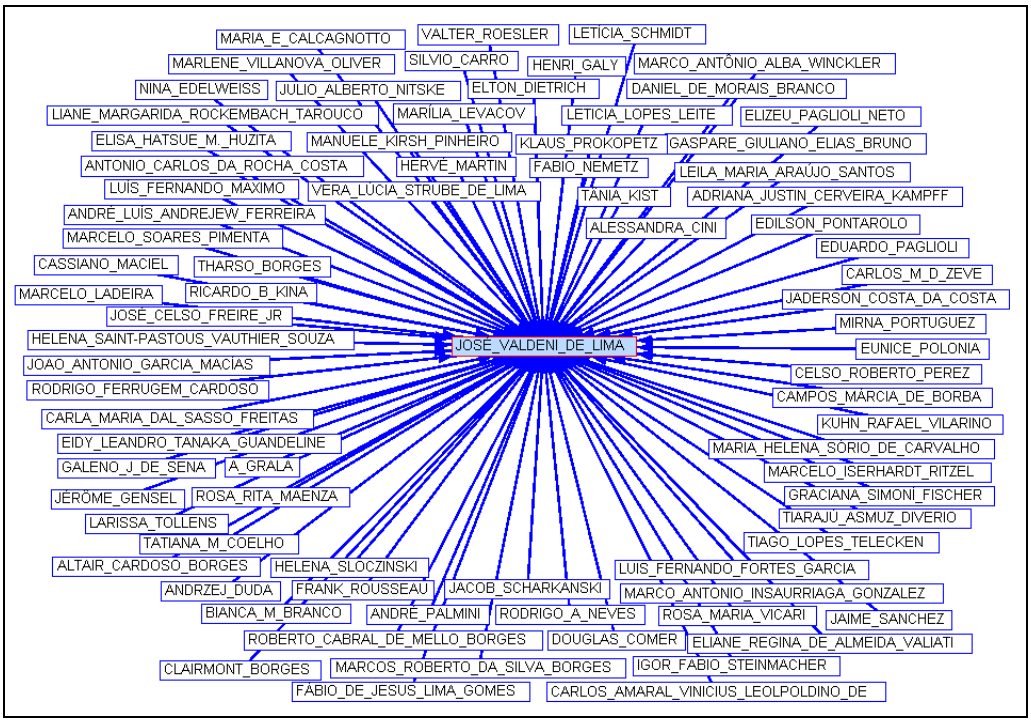

Figura 5.15: Grafo de co-autoria.

5.3.2 Fator de Impacto (FI)

O FI demonstra o quanto as publicações de um pesquisador têm repercussão na comunidade. Para o cálculo do FI são utilizadas todas as publicações do pesquisador e número de citações das publicações (obtido do Google Scholar em dezembro de 2007). Ou seja, é uma análise das publicações que tiveram citações (são considerados os trabalhos que têm repercussão) em relação ao total de publicações do pesquisador. Um exemplo das consultas realizadas para obter o FI é apresentado na Figura 5.16.

```
SELECT ?y
WHERE (?x prop:Name "Jose Valdeni de Lima"),
      (?x prop:hasPublication ?y),
      (?y prop:Citation ?z)
AND ?z >= 1
SELECT ?y
WHERE (?x prop:Name "Jose Valdeni de Lima"),
      (?x prop:hasPublication ?y)
```

Figura 5.16: Exemplo de consultas para obter o FI.

Na Figura 5.16 a primeira consulta (*select*) retorna as publicações do pesquisador "José Valdeni de Lima" com citações maiores ou iguais a 1. A segunda consulta retorna todas as publicações (com ou sem citações). Para o cálculo do FI divide-se o resultado da primeira consulta pelo resultado da segunda. Quanto mais próximo de 1, maior o FI. A Tabela 5.4 apresenta os dados obtidos para o cálculo do FI considerando apenas a área em que o pesquisador mais atuou. Por exemplo, o P6 possui 48 publicações e 15 delas possuem citações, enquanto P4 possui 160 publicações, mas somente 17 possuem citações. Nesse caso, o FI de P6 é 0.31 e o do P4 é 0.10, ou seja, a repercussão dos trabalhos de P6 é maior que de P4, mesmo que P4 possua bem mais trabalhos que P6.

Tabela 5.4: Cálculo do fator de impacto.

Pesquisador	Total de Publicações	Publicações com Citações	FI
P1	22	1	0.05
P2	19	0	0.00
P3	66	16	0.16
P4	160	17	0.10
P5	113	20	0.18
P6	48	15	0.31
P7	22	6	0.27
P8	93	15	0.16
P9	36	5	0.14
P10	134	18	0.13
P11	187	46	0.25
P12	35	2	0.06

A partir do cálculo do FI foi criado um *ranking* onde um FI maior significa que o pesquisador tem trabalhos de maior repercussão na comunidade. Este *ranking* é apresentado na Tabela 5.5. Não foram efetuadas análises complementares, por exemplo, para saber quem citou e quem foi citado, sendo que este tipo de análise é sugerido como trabalho futuro.

Tabela 5.5: *Ranking* dos pesquisadores pelo fator de impacto.

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

É importante salientar que o número de citações de uma publicação deve ser considerado como um indicador parcial de qualidade. Isto porque o número de citações depende obviamente da qualidade de um artigo, mas também depende de outras variáveis como o prestígio do autor (ou autores), do reconhecimento da instituição do autor, da atualidade do tema da publicação, do idioma da publicação e também do reconhecimento do meio de publicação. Então, o FI por medir a repercussão das publicações funciona apenas como um indicador indireto da qualidade das mesmas.

O FI aliado à co-autoria pode ser uma importante informação a ser analisada. Por exemplo, é possível verificar quem citou quem, para avaliar qual o nível de profundidade da citação, ou seja, se quem citou trabalha junto com o autor ou não (na mesma instituição, mesmo grupo de pesquisa ou costumam ser co-autores) ou até mesmo analisar auto-citações. Entretanto, tais análises não são contempladas nesta dissertação, ficando como sugestão de trabalho futuro.

5.4 Criação de Conglomerados (*Clusters*) de Pesquisadores

O processo de descoberta de conglomerados, *clustering* do inglês, é um método de descoberta de conhecimento que identifica associações ou correlações entre objetos, os quais são classificados por semelhança em agrupamentos (do inglês, *clusters*) relativamente homogêneos (WIVES, 2004). Wives afirma:

> Como o objetivo do agrupamento é organizar os objetos em conglomerados de objetos similares, ele está baseado na identificação da similaridade entre os objetos. Uma vez identificada a similaridade, eles são atribuídos a um conglomerado de objetos que possuem alguma relação de similaridade. Por conseqüência, objetos pertencentes a um mesmo conglomerado tendem a ser mais similares entre si do que em relação a outros objetos pertencentes a outros conglomerados. (2004, p. 28)

A descoberta de conglomerados, nesta dissertação, é importante para identificar se existem algumas relações entre a vida acadêmica dos pesquisadores e suas áreas de atuação. Por exemplo, verificar relações como: os pesquisadores que publicam mais em *journal* são de uma área X, já os que publicam mais em conferências são de uma área Y. O experimento realizado consiste na descoberta de conglomerados sem a definição das áreas de atuação dos pesquisadores, ou seja, são utilizados os indicadores de qualidade para criar os clusters. Os resultados desse experimento são comparados à qualificação dos pesquisadores por área a fim de verificar a existência de relações.

Segundo Wives (2004) existem diversos métodos para a descoberta de conglomerados, entre esses estão a aglomeração hierárquica e o K-means. A utilização do K-means é justificada devido a sua eficiência, enquanto a de aglomeração hierárquica é justificada por sua qualidade. Como o objetivo é ter o máximo de precisão na criação dos aglomerados, para poder identificar a existência ou não de padrões, optou-se por utilizar o método de aglomeração hierárquica. A técnica utilizada foi a CAH, Análise Aglomerativa Hierárquica (do inglês *Agglomerative Hierarchical Clustering*) também realizado através de um suplemento estatístico para o Microsoft Excel chamado XLSTAT[48]. Os dados utilizados para a descoberta de conglomerados foram os da Tabela 5.6 (não foram atribuídos pesos para os critérios, e não foram identificadas as áreas de atuação). Na Tabela 5.6 as colunas representam os pesquisadores (P1 à P12) e as linhas representam os valores dos 19 indicadores utilizados.

[48] http://www.xlstat.com/en/support/tutorials/cluster.htm

Tabela 5.6: Valores dos indicadores para cada pesquisador.

Indicador	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12
1- Formação acadêmica Pós-doutorado	0	0	1	0	1	0	0	0	0	0	1	0
2- Formação acadêmica Doutorado	1	1	1	1	1	1	1	1	1	1	1	1
3- Formação acadêmica Mestrado	0	1	2	1	1	1	1	1	1	0	1	1
4- Formação acadêmica Especialização	0	1	0	0	0	0	0	0	0	0	0	0
5- Formação acadêmica Graduação	1	1	1	1	3	1	1	1	1	1	1	1
6- Publicação Livro	0	0	3	4	2	0	0	6	0	0	2	0
7- Publicação Capítulo de Livro	0	2	3	6	2	4	0	0	0	0	7	2
8- Publicação *Paper* em *journal*	3	3	7	15	9	3	1	6	2	3	24	3
9- Publicação *Paper* em *proceeding*	19	14	53	135	100	41	21	81	34	19	154	30
10- Número de Citações	8	0	86	112	117	115	16	206	16	8	3628	32
11- Disciplina Ministrada para doutorado ou mestrado	2	0	6	26	0	0	2	4	15	2	6	4
12- Disciplina Ministrada para especialização	0	0	0	0	0	1	0	0	0	0	0	4
13- Disciplina Ministrada para graduação	2	8	22	14	1	14	38	3	17	2	5	6
14- Orientações concluídas para pós-doutorado e doutorado	0	0	0	9	4	0	0	5	0	0	3	0
15- Orientações concluídas para mestrado	0	0	0	0	0	0	0	0	0	0	0	0
16- Orientações concluídas para especialização	0	0	0	5	0	0	0	0	5	0	0	9
17- Orientações concluídas para graduação	4	0	9	20	1	11	20	0	34	4	13	11
18- Coordenador de projeto de pesquisa	0	1	2	0	2	1	1	3	6	0	1	5
19- Colaborador de projeto de pesquisa	6	3	3	0	0	6	11	4	3	6	0	5

Na técnica utilizada os dados são inicialmente distribuídos de modo que cada exemplo represente um conglomerado, então esses conglomerados são recursivamente agrupados considerando uma medida de similaridade (a utilizada foi a Distância Euclidiana) criando uma matriz de proximidade (ou dissimilaridade), até que todos os exemplos pertençam a apenas um conglomerado como representado no eixo Y do dendograma[49] da Figura 5.17.

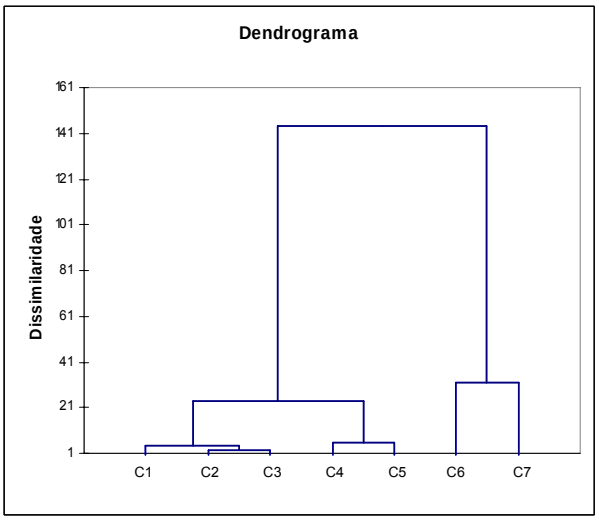

Figura 5.17: Dendograma.

Como pode ser observado no dendograma da Figura 5.17, foram criados 7 conglomerados, representados no eixo X. Posteriormente a técnica utilizada calcula os centróides e distância entre eles para obter um resultado por conglomerado. O resultado obtido é apresentado na Tabela 5.7 (os indicadores observados são os da Tabela 5.6).

Tabela 5.7: Conglomerados criados para os indicadores.

Indicador observado	Conglomerado
1, 2, 3, 4, 5, 6, 7,12, 14, 15, 16 e 18	1
8 e 19	2
9	3
10	4
11	5
13	6
17	7

Os conglomerados 1 e 2 são caracterizados por serem mais genéricos, pois englobam diferentes critérios. Já os conglomerados 3 ao 7 são específicos de um critério. O conglomerado 3 é caracterizado por *paper* em *proceeding*, o 4 pelo número de citações, o 5 são as disciplinas ministradas para doutorado e mestrado, o 6 são as disciplinas ministradas para graduação e o 7 são as orientações concluídas para graduação.

[49] O dendograma é um tipo especial de árvore na qual os nós pais agrupam os exemplos representados pelos nós filhos.

A verificação do enquadramento dos pesquisadores nos conglomerados criados é interessante para verificar a atuação acadêmica de cada um. Então, para encontrar a distribuição dos 12 pesquisadores por conglomerado criado foi utilizada a técnica MDPREF (*Muldimensional Analysis of Preference Data*) também realizado através do XLSTAT[50]. Com base nos dados da Tabela 5.6 é calculada a preferência de cada pesquisador pelos conglomerados. A ordem decrescente de preferência dos pesquisadores por conglomerado é apresentada na Tabela 5.8.

Tabela 5.8: Preferência dos pesquisadores por conglomerado.

1	2	3	4	5	6	7
P4	P11	P11	P11	P4	P9	P9
P11	P4	P4	P4	P9	P12	P12
P9	P5	P5	P5	P12	P7	P4
P12	P3	P3	P8	P11	P4	P7
P3	P8	P8	P3	P3	P6	P3
P8	P6	P9	P2	P8	P3	P6
P5	P2	P6	P6	P6	P10	P8
P6	P9	P12	P10	P7	P1	P11
P7	P10	P2	P1	P5	P8	P10
P2	P1	P10	P9	P2	P2	P1
P10	P12	P1	P12	P10	P11	P2
P1	P7	P7	P7	P1	P5	P5

Analisando a Tabela 5.8 é possível fazer algumas considerações como:

- Os pesquisadores P11, P4 e P5 preferem os conglomerados 2, 3 e 4, mesmo que P4 e P5 apareçam na segunda e terceira posição, respectivamente de tais conglomerados. Mesmo sendo de áreas diferentes (P11 é da área B.8_PERFORMANCE_AND_RELIABILITY, P4 é da área H.2_DATABASE_MANAGEMENT e P5 é da área H.5_INFORMATION_INTERFACES_AND_PRESENTATION__E.G.__HCI_), é interessante notar que P4 e P11 possuem atualmente bolsa de produtividade em pesquisa do CNPq e P5 já possuiu tal bolsa, sendo que entre os pesquisadores analisados só P10 e P3 possuem esta bolsa atualmente.
- P12 prefere os conglomerados 6 e 7 na segunda posição e P7 prefere o conglomerado 6 na terceira posição e o 7 na quarta posição. Pode-se dizer que P12 e P7, mesmo sendo de áreas diferentes (respectivamente, I.2_ARTIFICAL_INTELLIGENCE e H.2_DATABASE_MANAGEMENT) têm proximidade de perfis por preferirem os conglomerados relacionados a ministrar disciplinas para graduação e ter orientações concluídas para graduação.
- Os pesquisadores P1, P2, P3, P6, P8 e P10 têm preferência por algum conglomerado, entretanto tal preferência é considerada fraca, pois é encontrada a partir da quinta posição no *ranking*.

[50] http://www.xlstat.com/en/support/tutorials/prefmap.htm

5.5 Considerações

Diversas análises poderão ser efetuadas a partir do perfil dos pesquisadores. Algumas análises poderiam ser: quais pesquisadores ministraram ou ministram uma disciplina X; quais pesquisadores são sênior e quais são júnior; dos 12 pesquisadores analisados quais possuem bolsa de produtividade científica; dentre outras análises. Entretanto, esta dissertação focou apenas em algumas análises, consideradas mais relacionadas ao contexto em que este trabalho está inserido, sendo que outras análises são sugeridas como trabalhos futuros.

De modo geral, o protótipo permitiu a qualificação de todos os pesquisadores. Alguns pesquisadores como o P11, por exemplo, têm sua atuação mais concentrada em uma área. Já outros, como o P4, por exemplo, possuem sua atuação distribuída em diversas áreas. Tal fato não significa que o P4 seja melhor ou pior que o P11 ou vice-versa, significa apenas que os pesquisadores têm atuações acadêmicas diferentes.

Os pesquisadores possuem algumas competências que não puderam ser classificadas em nenhuma área. Para isso foi criada uma categoria chamada de NÃO_CLASSIFICADO. Tal fato necessita ser mais analisado, em trabalhos futuros, para que o sistema classifique tudo que está nos currículos. Durante os experimentos foram identificadas algumas limitações e dificuldades, como:

- O currículo Lattes permite muita liberdade no seu preenchimento, isso possibilita que os pesquisadores digitem importantes informações sem padronização. Este fato causa discrepâncias nos dados, por exemplo: um pesquisador colocou em seu Lattes a participação na seguinte conferência “Eurographics/IEEE VGTC Symposium on Visualization”, entretanto o nome da conferência é “EuroVis/Joint Eurographics - IEEE TCVG Symposium on Visualization”. Esse ruído se propaga e desta forma o próprio pesquisador fica prejudicado na sua qualificação;
- Alguns pesquisadores não preenchem todas as informações necessárias para a sua qualificação. Por exemplo, um pesquisador pode colocar no currículo que possui doutorado, mas não preencher o título do trabalho impossibilitando a descoberta da área do doutorado. Então, para uma qualificação correta é necessário que os pesquisadores preencham corretamente e completamente seus currículos, pois todo o processo de qualificação é baseado na descoberta de informações a partir das informações descritas no currículo Lattes;
- A utilização de uma amostra não-probabilística caracteriza os resultados apenas como indícios de que as análises efetuadas estejam corretas, entretanto tais análises não podem ser generalizadas;
- É necessário inserir informações sobre as conferências na ontologia OntoQualis, para que esta possa calcular o Qualis das publicações dos pesquisadores. E como os sites das conferências (que possuem as informações necessárias para qualificar as conferências) não seguem um padrão, não é possível instanciar tais informações de forma automática. Isso é um problema num primeiro momento, pois uma vez que várias conferências forem instanciadas a descoberta do Qualis é simples. Além disso, é necessário que a OntoDoc possua as informações sobre as publicações, sobre as áreas das disciplinas ministradas, sobre os projetos de pesquisa para que possa inferir as áreas.

6 CONCLUSÕES E TRABALHOS FUTUROS

Esta dissertação apresenta um sistema Web que busca a descoberta das qualificações dos pesquisadores por área de atuação na Ciência da Computação, baseado em uma ontologia de perfil. As principais contribuições desta pesquisa são a definição dos indicadores de qualidade de pesquisadores, o desenvolvimento da ontologia de perfil OntoResearcher e a qualificação dos pesquisadores por área de atuação. Outras contribuições são: utilização de diferentes fontes de informação, o reuso de ontologias e a implementação de um protótipo acessível via web.

Primeiramente foram identificadas as informações necessárias para modelar o perfil dos pesquisadores e desenvolver a OntoResearcher. Posteriormente foram identificadas as informações modeladas no perfil que permitem qualificar os pesquisadores. A próxima fase consistiu na obtenção das informações necessárias para popular a ontologia (currículo Lattes, Google Scholar, *OntoQualis* e *OntoDoc*). Posteriormente foi efetuada uma análise para atribuição de pesos aos indicadores e realizado o cálculo das qualificações por área.

O escopo da qualificação de pesquisadores da área da Ciência da Computação, já restringe consideravelmente a ambigüidade da qualificação de pesquisadores no âmbito de diversas áreas do conhecimento. Mesmo assim, o processo de qualificação pode não ser considerado justo por muitos pesquisadores. De fato, não há um consenso quanto à eficiência do uso dos indicadores para medir qualitativamente e quantitativamente a produção científica. Ainda que existam tais ambigüidades e discussões a cerca de qual seria a melhor maneira de qualificar os pesquisadores, buscou-se nesta dissertação, descrever o perfil dos pesquisadores mais próximo de uma abordagem completa e fiel de sua atuação acadêmica.

Algumas informações modeladas na ontologia, como: endereço, e-mail, homepage, idiomas, não servem para o processo de qualificação. Estas informações servirão, por exemplo, em um sistema de recomendação de artigos científicos. Pois é necessário saber que idiomas o pesquisador tem conhecimento para poder recomendar um artigo para ele, por exemplo, se um pesquisador não compreende alemão não adianta recomendar um artigo neste idioma para ele. Assim, os perfis definidos podem servir como base em um sistema de recomendação de artigos científicos. Além disso, a descoberta da qualificação do pesquisador pode ser utilizada para a criação de comunidades virtuais, baseadas nas áreas da Ciência da Computação; bem como em processos de seleção que necessitem saber quais pesquisadores são especialistas em determinada área ou mesmo em disputas por recursos pode-se criar um ranking de pesquisadores.

O sistema Web desenvolvido nesta dissertação é uma ferramenta (semi) automatizada para a descoberta da qualificação dos pesquisadores. Isto porque é necessária a intervenção em alguns momentos como: popular a *OntoQualis* com as informações sobre as conferências para que ela possa inferir o Qualis das mesmas e popular a *OntoDoc* com os documentos para que ela infira as áreas.

É importante salientar que o objetivo da qualificação dos pesquisadores não é o de substituir a avaliação pelos pares, e sim complementar tal processo, apresentando uma análise da atuação dos pesquisadores. Os resultados obtidos no processo de qualificação dos pesquisadores demonstraram que:

- Uma análise completa do currículo de um pesquisador efetuada manualmente é praticamente inviável tomando muito tempo, pois é necessário pesquisar o número de citações para cada publicação do currículo, encontrar a área de cada publicação, das disciplinas ministradas, dos projetos de pesquisa, encontrar o Qualis das publicações, etc. Assim sendo, um processo (semi) automatizado que encontre essas informações e calcule a qualificação é útil para diversos processos que não dispõem de tempo para análises manuais.
- Uma qualificação precisa depende quase exclusivamente das informações que o pesquisador disponibiliza em seu currículo. Quanto mais informações forem colocadas no currículo mais precisa será a qualificação obtida.

Apesar das vantagens, o uso de ontologias apresenta alguns problemas. Desde problemas como processo de escolha de uma ontologia, isto porque nenhuma ontologia pode ser totalmente adequada a todos os indivíduos ou grupos, passando por desenvolvimento e manutenção, até problemas do tratamento da evolução de ontologias (GUIZZARDI, 2000). Nesta dissertação, o principal problema encontrado foram as diversas modificações feitas na OntoResearcher, pois a cada modificação é necessário revisar todo o sistema para que a ontologia seja populada corretamente e as informações possam ser extraídas como esperado.

6.1 Trabalhos Futuros

No decorrer desta dissertação, foram identificados alguns trabalhos futuros, como:

- Disponibilizar o sistema desenvolvido para a comunidade da Ciência da Computação para obter um número mais expressivo de currículos e então poder fazer análises com dados estatísticos.
- Tornar o módulo de atribuição de pesos configurável via Web, permitindo que o próprio pesquisador altere os pesos no momento do cadastro. Por exemplo, dar a liberdade ao pesquisador para que este possa dar um peso maior para o Qualis das publicações ou dar o mesmo peso para todos os indicadores, por exemplo.
- Implementar uma técnica de similaridade para evitar a população de instâncias repetidas na OntoResearcher.
- Enviar o resultado das qualificações para os pesquisadores para verificar se eles concordam com os resultados obtidos.
- Testar o perfil do pesquisador no contexto de um sistema de recomendação de artigos científicos.

- Realizar mais análises sobre o perfil dos pesquisadores, tais como: verificar quem citou quem (análise da profundidade das citações); inferir áreas de interesse dos pesquisadores, considerando períodos da vida acadêmica; verificar quais pesquisadores possuem bolsa de produtividade científica, dentre outras análises descritas ao longo desta dissertação ou que forem importantes dependo do contexto em que os perfis forem aplicados.
- A evolução do perfil dos pesquisadores, por exemplo, quando um pesquisador que já está cadastrado no sistema quiser enviar um novo currículo como armazenar informações sobre as diferenças entre os currículos. No momento quando um pesquisador envia um novo currículo, são apenas armazenadas as novas informações acrescentadas no currículo, entretanto não são armazenadas informações sobre as versões.
- Revisar os indicadores de qualidade utilizados para qualificar os pesquisadores, inclusive incluindo novos indicadores. Por exemplo, analisar a viabilidade de incluir o *h-Index* como indicador.
- Testar outras técnicas de atribuição de pesos aos indicadores para comparar com os resultados e então decidir qual seria a técnica mais adequada.
- Criação de comunidades virtuais onde os pesquisadores possam trocar experiências em suas áreas de atuação, após obter um número mais expressivo de currículos.
- Melhorar a OntoResearcher, acrescentando maior número de informações semânticas para que os pesquisadores construam suas páginas pessoais com tais informações semânticas. E assim possamos utilizar mecanismos de busca por informações dos pesquisadores de forma automática.
- Propor que o CNPq adote a classificação da ACM para as áreas da Ciência da Computação. Isto porque as áreas nas quais o pesquisador pode classificar, tanto suas publicações quanto sua própria atuação são muito genéricas. Então, se os próprios pesquisadores puderem definir as áreas, seria mais simples para o sistema identificar a atuação dos pesquisadores.
- Integrar o sistema desenvolvido nesta dissertação ao sistema de editoração aberta do projeto DIGITEX.

REFERÊNCIAS

ALMEIDA, M. B. **Um modelo baseado em ontologias para representação da memória organizacional**. 2006. 345f. Tese (Doutorado em Ciência da Informação) - Programa de Pós Graduação da Escola de Ciência da Informação da Universidade Federal de Minas Gerais, Belo Horizonte.

BERNERS-LEE, T. B. **Building the future**. Disponível em: <http://www.w3.org/2000/Talks/0906-xmlweb-tbl/slide9-6.html>. Apresentação titulada: XML and the Web, realizada no ano de 2000. Acesso em: out. 2006.

BERNERS-LEE, T.; HENDLER, J.; LASSILA, O. The Semantic Web. **Scientific American**, [S.l.], v.284, n.5, p. 34-43, May 2001.

BORCHERDING, K.; EPPEL, T.; WINTERFELDT, D. von. Comparison of Weighting Judgements in Multiattribute Utility Measurement. **Management Science**, [S.l.], v. 37, n. 12, p. 1603-1619, 1991.

BREITMAN, K. **Web Semântica:** a internet do futuro. Rio de Janeiro: LTC, 2005.

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.

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

CHEN, W.; MIZOGUCHI, R. Communication Content Ontology for Learner Model Agent in Multi-Agent Architecture. In: ADVANCED RESEARCH IN COMPUTERS AND COMMUNICATIONS IN EDUCATION, AIED, 1999. **Proceedings…** [S.l.:s.n.], 1999. p. 95-102.

CHEN, C. et al. **Web Ontology Language-OWL**. 2003. Disponível em: <ww.cs.concordia.ca/~faculty/haarslev/teaching/semweb/OWL.ppt>. Acesso em: set. 2006.

COSTELLO, R.; JACOBS, D. B. **OWL Web Ontology Language**. 2003. Disponível em: <http://scholar.google.com/scholar?hl=pt-BR&lr=&q=cache:c_4ExYzP46IJ:cerebra.com/downloads/MITRE-OWL-Tutorial.pdf+COSTELLO,+R.,+JACOBS,+.+2003.+OWL+Web+Ontology+Language.+>. Acesso em: nov. 2006.

DOLOG, P. et al. Personalization in Elena: How to cope with personalization in distributed eLearning Networks. In: CONFERENCE ON WORLDWIDE COHERENT

WORKFORCE, SATISFED USERS – NEW SERVICES FOR SCIENTIFIC INFORMATION, 2003. **Proceedings…** [S.l.:s.n.], 2003.

FELICÍSSIMO, C. H. **Interoperabilidade Semântica na Web:** uma Estratégia para o Alinhamento Taxonômico de Ontologias. 2004. 180 f. Dissertação (Mestrado) – Departamento de Informática, PUC-Rio, Rio de Janeiro.

FREITAS, F. L. G. **Ontologias e a Web Semântica**. 2005. Disponível em: <http://www.inf.unisinos.br/~renata/cursos/topicosv/ontologias-ws.pdf>. Acesso em: set. 2006.

GÓMEZ-PÉREZ, A.; FERNÁNDEZ-PÉREZ, M.; CORCHO, O. **Ontological Engineering whit Examples from the Áreas of Knowledge Management, E-Commerce and Semantic Web**. London: Springer, 2004.

GRUBER, T. Toward principles for the Design of Ontologies Used for Knowledge Sharing. **International Journal of Human and Computer Studies**, [S.l.], v. 43, n.5/6, p.907-928, 1995.

GUARINO, N.; GIARETTA, P. Ontologies and knowledge bases: Towards a terminological clarification. In: MARS, N.J.I. **Towards Very Large Knowledge Bases:** Knowledge Building and Knowledge Sharing. Amsterdã: IOS Press, 1995. p. 25–32.

GUARINO, N.; WELTY, C. **Ontological Analysis of Taxonomic Relationships**. 2000. Disponível em: <http://citeseer.ist.psu.edu/guarino00ontological.html>. Acesso em: abr. 2006.

GUIMARÃES, M. C. S. **Avaliação em ciência e tecnologia:** um estudo prospectivo em química. 1992. 289f. Dissertação (Mestrado em Ciência da Informação)- UFRJ-ECO/CNPq-IBICT, Rio de Janeiro- RJ.

GUIZZARDI, G. **Desenvolvimento para e com Reuso**: um Estudo de Caso no Domínio de Vídeo Sob Demanda. 2000. 202f. Dissertação (Mestrado em Informática) – Centro Tecnológico da Universidade Federal do Espírito Santo, Vitória, Espírito Santo.

HANNEL, K.; LIMA, J. V. de. Qualificação de Pesquisadores por Área da Ciência da Computação com Base em uma Ontologia de Perfil. In: WORKSHOP DE TESES E DISSERTAÇÕES, WTDWeb, 2007, Gramado. **Anais**... Porto Alegre, RS: SBC, 2007.

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.

HORRIDGE, M. et al. **A Practical Guide To Building OWL Ontologies Using The Protégé-OWL Plugin and CO-ODE Tools Edition 1.0**. 2004. Disponível em: <http://www.co-ode.org/resources/tutorials/ProtegeOWLTutorial.pdf>. Acesso em: out. 2006.

INSTITUTE OF ELECTRICAL AND ELECTRONICS ENGINEERS. **IEEE Standard 1074:** Standard for developing software life cycle processes. [S.l.], 1995. Disponível em:

<http://ieeexplore.ieee.org/xpls/abs_all.jsp?isnumber=10452&arnumber=490501&count=2&index=0>. Acesso em: dez. 2006.

ISO - INTERNATIONAL ORGANIZATION FOR STANDARIZATION. **ISO 3166-1:** English country names and code elements. [S.l.], 2007. Disponível em: <http://www.iso.org/iso/country_codes/iso_3166_code_lists/english_country_names_and_code_elements.htm>. Acesso em: jun. 2007.

ISO - INTERNATIONAL ORGANIZATION FOR STANDARIZATION. **ISO 639-2 alpha 3:** English language names and code elements. 1998. Disponível em: <http://www.loc.gov/standards/iso639-2/php/code_list.php>. Acesso em: jun. 2007.

JENA - A Semantic Web Framework for Java. 2006. Disponível em: <http://jena.sourceforge.net/>. Acesso em: dez 2007.

KOIVUNEM, M.; MILLER, E. **W3C Web Semanic Activity**. 2001. Disponível em: <http://www.w3.org/2001/12/semweb-fin/w3csw>. Acesso em: set. 2006.

KORHONEN, P.; SILJAMÄKI, A.; SOISMAA, M. On the use of value efficiency analysis and some further developments. **Journal of Productivity Analysis,** Boston, v.17, n. 1-2, p. 49–64, Jan. 2002.

LUGANO, G. **Semantic Web Technologies and the FOAF Project**. 2005. Disponível em: < http://www.cs.helsinki.fi/u/chande/courses/cs/MWS/seminar/Articles/12.pdf>. Acesso em: nov. 2006.

MACHADO, R. P. **Um Serviço de Matching de Interesses Dependentes de Localização**. 2005. Dissertação (Mestrado em Informática) - Programa de Pós-graduação em Informática da PUC-Rio, Rio de Janeiro.

MAEDCHE, A. **Ontology Learning for The Semantic Web**. Boston: Kluwer Academic, 2002. 244p.

MIDDLETON, S. E.; SHADBOLT, N. R.; DE ROURE, D. C. Ontological User Profiling in Recommender Systems. **ACM Transactions on Information Systems**. New York, v. 22, n. 1, p. 54-88, Jan. 2004.

MUSA, D. L. **Compartilhamento de Modelos de Alunos via Ontologia e Web Services**. 2006. 113f. Tese (Doutorado em Ciência da Computação) – Instituto de Informática, UFRGS, Porto Alegre - RS.

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

NOY, N. F.; MCGUINNESS, D. L. **Ontology Development 101:** A Guide to Creating Your First Ontology. Stanford: Stanford University, 2001. Disponível em: <http://www.ksl.stanford.edu/people/dlm/papers/ontology101/ontology101-noy-mcguinness.html>. Acesso em: out. 2006.

OLIVEIRA, J. P. M. de.; GALANTE, R. M.; MUSA, D. L.: EDELWEISS, N. Uma proposta de Editoração, Indexação e Busca de Documentos Científicos em um Processo de Avaliação Aberta. In: WORKSHOP DE BIBLIOTECAS DIGITAIS, WDL, 1., 2005.

Uberlândia. **Anais...** [S.l.: s.n.], 2005. Disponível em: <http://www.lbd.dcc.ufmg.br/wdl2005/JPalazzoWDL05.pdf>. Acesso em: maio 2006.

PARNAS, D. L. Stop the Numbers Game. **Commun. of the ACM**, New York, v. 50, n.11, p. 19-21, Nov. 2007.

REED, S.L.; LENAT, D.B. **Mapping Ontologies into Cyc**. 2002. Disponível em: <http://www.cyc.com/doc/white_papers/mapping-ontologies-into-cyc_v31.pdf>. Acesso em: nov. 2006.

RECH, R. **Um Modelo de Pontuação na Busca de Competências Acadêmicas de Pesquisadores.** 2007. 92f. Dissertação (Mestrado em Ciência da Computação) – Instituto de Informática, UFRGS, Porto Alegre - RS.

REN, J.; TAYLOR, R. N. Automatic and versatile publications ranking for research institutions and scholars. **Commun. of the ACM**, New York, NY, USA, v.50, n.6, p.81–85, June 2007.

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, ERBD, 1., 2005. **Anais...** Porto Alegre: SBC, 2005. Disponível em: <http://www.inf.ufrgs.br/~erbd2005/Artigos/7866.pdf>. Acesso em: mar. 2007.

RODRIGUES, S.; OLIVEIRA, J. Competence mining for vitual scientific community creation. **Int. J. Web Based Comunities**, [S.l.], v.1, n.1, p. 90-102, July. 2004.

RODRIGUEZ, M. A.; BOLLEN, J.; VAN de SOMPEL, H. A Pratical Ontology for the Large-Scale Modeling of Scholarly Artifacts and their Usage.In: CONFERENCE ON DIGITAL LIBRARIES, JCDL, 2007, Vancouver, Canada. ***Proceedings…*** New York: ACM, 2007. p. 278-287.

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, M.; WELTY, C.; McGUINNESS, D. **Web Ontology Language (OWL) Guide Version 1.0**. 2003. Disponível em: <http://www.w3.org/TR/2003/WD-owl-guide-20030210/>. Acesso em: nov. 2006.

SMITH, T. F.; WATERMAN, M. S. Identification of common molecular subsequences. **J. Mol. Biol.** [S.l.], v.147, p. 195-197, 1981. Disponível em: <http://gel.ym.edu.tw/~chc/AB_papers/03.pdf>. Acesso em: dez. 2007.

SOUTO, M. A. M.; WARPECHOWSKI, M.; OLIVEIRA, J. P. M. de. An Ontological Approach for the Quality Assessment of Computer Science Conferences. In: INTERNATIONAL WORKSHOP ON QUALITY OF INFORMATION SYSTEMS QoIS, 3., 2007. **Proceedings...** [S.l.: s.n.], 2007.

SOUTO, M. A. M.; WARPECHOWSKI, M.; OLIVEIRA, J. P. M. de. Modelo de Avaliação da Qualidade de Conferências Científicas na Área da Ciência da Computação: uma Abordagem Ontológica. In: WORKSHOP ON ONTOLOGIES AND METAMODELING SOFTWARE AND DATA ENGINEERING, WOMSDE, 1., 2006. **Anais…** Florianópolis: SBC, 2006. p. 93-102.

SOWA, J. F. **Ontology, Metadata, and Semiotcs**. Berlin: Springer – Verlag, 2000. p. 55-81. (Lecture Notes in Computer Science, v. 1867).

USCHOLD, M.; KING, M. Towards a Methodology for Building Ontologies. In: WORKSHOP ON BASIC ONOTOLOGICAL ISSUES IN KNOWLEDGE SHARING, 1995, Edinburgh. **Proceedings…** [S.l.:s.n.], 1995.

VELHO, L. A ciência e seu público. **Transinformação**, Campinas, v. 9, n. 3, p. 15-32, set./dez. 1997.

VIEIRA, R. et al. **Web Semântica:** ontologias, lógica de descrição e inferência. 2005. Disponível em: <http://www.inf.unisinos.br/~renata/laboratorio/publicacoes/webmedia-webs.pdf>. Acesso em: out. 2006.

WIVES, L. K. **Utilizando *conceitos* como descritores de textos para o processo de identificação de conglomerados (*clustering*) de documentos**. 2004. 136f. Tese (Doutorado em Ciência da Computação) – Instituto de Informática, UFRGS, Porto Alegre - RS.

W3C Recommendation. **RDF Prime**. 2004. Disponível em: <http://www.w3.org/TR/rdf-primer/>. Acesso em: nov. de 2006.

ANEXO A INDICADORES UTILIZADOS POR CAZELLA (2006) E RECH (2007)

A tabela seguinte mostra o impacto dos indicadores quantitativos calculados por Cazzela (2006) e posteriormente utilizados no trabalho de Rech (2007). O impacto destes indicadores foi calculado em um experimento onde 25 pesquisadores doutores, da Ciência da Computação da UFRGS, responderam um questionário ponderando tais indicadores (CAZELLA, 2006).

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

Fonte: CAZELLA, 2006.

Impacto dos indicadores qualitativos utilizados no trabalho de Rech (2007). O impacto destes indicadores foi calculado empiricamente pelo autor.

Categoria: Classificação dos Veículos de Publicação – Periódicos			**% Imp. da Cat.**
Indicador	*% Imp. do Indicador*	*Peso Calculado*	
PQARI	40	20	50
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	

Fonte: RECH, 2007, p. 73-74.

ANEXO B DEFINIÇÃO DOS PESOS

Os cálculos efetuados, para cada um dos critérios e sub-critétios são apresentados nas tabelas que seguem:

Formação acadêmica	*Ranking*	*Nota*	*Peso*
Pós Doutorado	1	100	0,317
Doutorado	2	85	0,270
Mestrado	3	60	0,190
Especialização	4	40	0,127
Graduação	5	30	0,095
Pior hipótese	6	0	0

Publicações	*Ranking*	*Nota*	*Peso*
Livro	2	30	0,256
Capítulo do Livro	4	20	0,171
Journal	1	38	0,325
Proceeding	3	29	0,247
Pior hipótese	5	0	0

Qualis	*Ranking*	*Nota*	*Peso*
A	1	100	0,513
B	2	60	0,308
C	3	35	0,179
Pior hipótese	4	0	0

Citações	*Ranking*	*Nota*	*Peso*
Número de citações	1	100	1
Pior hipótese	2	0	0

Disciplinas Ministradas	*Ranking*	*Nota*	*Peso*
Mestre e Doutorado	1	100	0,5
Especialização	2	60	0,3
Graduação	3	40	0,2
Pior hipótese	4	0	0

Comitê de Programa	*Ranking*	*Nota*	*Peso*
Membro	1	100	1
Pior hipótese	2	0	0

Orientações Concluídas	*Ranking*	*Nota*	*Peso*
Pós Doutorado ou Doutorado	1	46	0,46
Mestrado	2	30	0,3
Especialização	3	14	0,14
Graduação	4	10	0,1
Pior hipótese	5	0	0

Participação Projeto de pesquisa	*Ranking*	*Nota*	*Peso*
Coordenador	1	100	0,56
Colaborador	2	80	0,44
Pior hipótese	3	0	0

A tabela final com os pesos dos critérios gerais ficou como segue:

Critério	*Ranking*	*Nota*	*Peso*
Formação Acadêmica	2	60	14,63%
Publicações	1	100	24,43%
Qualis	3	50	12,19%
Citações	3	50	12,19%
Disciplinas Ministradas	4	45	10,97%
Comitê de programa	6	35	8,53%
Orientações concluídas	5	40	9,75%
Projeto de pesquisa	7	30	7,31%
Pior hipótese	8	0	0
Total			100%

Após a definição dos pesos finais para cada critério, os valores dos sub-critérios foram calculados utilizando uma regra de três simples e obtendo-se assim os valores apresentados na Tabela 4.4 da seção 4.4.4.

ANEXO C RESULTADOS DO CÁLCULO CQ

A seguir são apresentados os valores encontrados para as áreas de atuação de cada um dos 12 pesquisadores:

P1	NÃO CLASSIFICADO	0,766	22,43%
	A.0_GENERAL	0,022	0,64%
	H._INFORMATION_SYSTEMS	0,022	0,64%
	H.2_DATABASE_MANAGEMENT	2,448	71,7%
	H.3_INFORMATION_STORAGE_AND_RETRIEVAL	0,101	2,96%
	H.4_INFORMATION_SYSTEMS_APPLICATIONS	0,041	1,20%
	J.3_LIFE_AND_MEDICAL_SCIENCES	0,014	0,41%

P2	NÃO CLASSIFICADO	0,529	31,60%
	A.0_GENERAL	0,066	3,94%
	D.3_PROGRAMMING_LANGUAGES	0,022	1,31%
	E.1_DATA_STRUCTURES	0,044	2,63%
	H.2_DATABASE_MANAGEMENT	0,143	8,54%
	H.3_INFORMATION_STORAGE_AND_RETRIEVAL	0,162	9,68%
	H.4_INFORMATION_SYSTEMS_APPLICATIONS	0,028	1,67%
	H.5_INFORMATION_INTERFACES_AND_PRESENTATION__E.G.__HCI_	0,158	9,44%
	K.3_COMPUTERS_AND_EDUCATION	0,522	31,20%

P3	NÃO CLASSIFICADO	3,221	17,76%
	A.0_GENERAL	0,066	0,36%
	C.1_PROCESSOR_ARCHITECTURES	0,022	0,12%
	C.4_PERFORMANCE_OF_SYSTEMS	0,022	0,12%
	D.3_PROGRAMMING_LANGUAGES	0,153	0,84%
	E._DATA	0,022	0,12%
	E.1_DATA_STRUCTURES	0,044	0,24%
	G.4_MATHEMATICAL_SOFTWARE	0,060	0,33%
	H.5_INFORMATION_INTERFACES_AND_PRESENTATION__E.G.__HCI_	1,759	9,70%
	I.2_ARTIFICIAL_INTELLIGENCE	0,121	0,67%
	I.3_COMPUTER_GRAPHICS	11,118	61,31%
	I.4_IMAGE_PROCESSING_AND_COMPUTER_VISION	1,383	7,63%
	J.3_LIFE_AND_MEDICAL_SCIENCES	0,060	0,33%
	K.6_MANAGEMENT_OF_COMPUTING_AND_INFORMATION_SYSTEMS	0,082	0,45%

P4	NÃO CLASSIFICADO	15,083	52,68%
	A.0_GENERAL	0,077	0,27%
	C.2_COMPUTER-COMMUNICATION_NETWORKS	0,310	1,08%
	C.3_SPECIAL-PURPOSE_AND_APPLICATION-BASED_SYSTEMS	0,060	0,01%
	C.5_COMPUTER_SYSTEM_IMPLEMENTATION	0,121	0,42%
	D.1_PROGRAMMING_TECHNIQUES	0,207	0,72%
	D.2_SOFTWARE_ENGINEERING	0,497	1,74%
	D.3_PROGRAMMING_LANGUAGES	0,187	0,65%

	D.4_OPERATING_SYSTEMS	0,324	1,13%
	E._DATA	0,055	0,19%
	E.5_FILES	0,132	0,46%
	H._INFORMATION_SYSTEMS	0,274	0,96%
	H.0_GENERAL	0,148	0,52%
	H.1_MODELS_AND_PRINCIPLES	0,060	0,01%
	H.2_DATABASE_MANAGEMENT	2,718	9,49%
	H.3_INFORMATION_STORAGE_AND_RETRIEVAL	1,119	3,91%
	H.4_INFORMATION_SYSTEMS_APPLICATIONS	0,943	3,29%
	H,5_INFORMATION_INTERFACES_AND_PRESENTATION__E.G.__HCI_	0,089	0,31%
	I,2_ARTIFICIAL_INTELLIGENCE	2,031	7,09%
	I.5_PATTERN_RECOGNITION	0,055	0,19%
	I.6_SIMULATION_AND_MODELING	0,099	0,35%
	I.7_DOCUMENT_AND_TEXT_PROCESSING	0,896	3,13%
	J.1_ADMINISTRATIVE_DATA_PROCESSING	0,014	0,05%
	J.3_LIFE_AND_MEDICAL_SCIENCES	0,121	0,42%
	K.3_COMPUTERS_AND_EDUCATION	2,348	8,2%
	K.4_COMPUTERS_AND_SOCIETY	0,067	0,23%
	K.6_MANAGEMENT_OF_COMPUTING_AND_INFORMATION_SYSTEMS	0,115	0,4%
	K.8_PERSONAL_COMPUTING	0,483	1,69%

P5	NÃO CLASSIFICADO	5,898	26,35%
	C.2_COMPUTER-COMMUNICATION_NETWORKS	1,714	7,66%
	D.2_SOFTWARE_ENGINEERING	0,121	0,54%
	D.3_PROGRAMMING_LANGUAGES	0,088	0,39%
	F.3_LOGICS_AND_MEANINGS_OF_PROGRAMS	0,060	0,27%
	H.2_DATABASE_MANAGEMENT	0,891	3,98%
	H.3_INFORMATION_STORAGE_AND_RETRIEVAL	0,303	1,35%
	H.4_INFORMATION_SYSTEMS_APPLICATIONS	1,519	6,79%
	H.5_INFORMATION_INTERFACES_AND_PRESENTATION__E.G.__HCI_	9,060	40,48%
	I.2_ARTIFICIAL_INTELLIGENCE	0,182	0,81%
	K.3_COMPUTERS_AND_EDUCATION	2,543	11,36%

P6	NÃO CLASSIFICADO	9,875	54,87%
	A.0_GENERAL	0,044	0,24%
	D.1_PROGRAMMING_TECHNIQUES	0,121	0,67%
	D.2_SOFTWARE_ENGINEERING	0,124	0,69%
	D.3_PROGRAMMING_LANGUAGES	0,110	0,61%
	D.4_OPERATING_SYSTEMS	0,066	0,37%
	E.4_CODING_AND_INFORMATION_THEORY	0,041	0,23%
	E.5_FILES	0,022	0,12%
	H._INFORMATION_SYSTEMS	0,055	0,31%
	H.2_DATABASE_MANAGEMENT	0,301	1,68%
	H.3_INFORMATION_STORAGE_AND_RETRIEVAL	3,680	20,49%
	H.5_INFORMATION_INTERFACES_AND_PRESENTATION__E,G.__HCI_	0,080	0,45%
	I.2_ARTIFICIAL_INTELLIGENCE	1,345	7,49%
	I.3_COMPUTER_GRAPHICS	0,060	0,33%
	I.5_PATTERN_RECOGNITION	0,067	0,37%
	I.7_DOCUMENT_AND_TEXT_PROCESSING	1,703	9,48%
	J.3_LIFE_AND_MEDICAL_SCIENCES	0,014	0,08%
	K.3_COMPUTERS_AND_EDUCATION	0,194	1,08%
	K.5_LEGAL_ASPECTS_OF_COMPUTING	0,060	0,33%

P7	NÃO CLASSIFICADO	0,768	14,36%
	A.0_GENERAL	0,350	6,54%
	D.2_SOFTWARE_ENGINEERING	0,060	1,12%
	D.3_PROGRAMMING_LANGUAGES	0,153	2,86%

	E.1_DATA_STRUCTURES	0,022	0,41%
	E.5_FILES	0,022	0,41%
	F.3_LOGICS_AND_MEANINGS_OF_PROGRAMS	0,022	0,41%
	H.2_DATABASE_MANAGEMENT	3,108	58,10%
	H.3_INFORMATION_STORAGE_AND_RETRIEVAL	0,469	8,77%
	H.4_INFORMATION_SYSTEMS_APPLICATIONS	0,063	1,18%
	H.5_INFORMATION_INTERFACES_AND_PRESENTATION__E.G.__HCI_	0,121	2,26%
	I.2_ARTIFICIAL_INTELLIGENCE	0,101	1,89%
	I.7_DOCUMENT_AND_TEXT_PROCESSING	0,038	0,71%
	K.3_COMPUTERS_AND_EDUCATION	0,041	0,77%
	K.4_COMPUTERS_AND_SOCIETY	0,010	0,19%

P8	NÃO CLASSIFICADO	15,94	45,71%
	C.2_COMPUTER-COMMUNICATION_NETWORKS	0,193	0,55%
	C.3_SPECIAL-PURPOSE_AND_APPLICATION-BASED_SYSTEMS	0,060	0,17%
	D.1_PROGRAMMING_TECHNIQUES	0,242	0,69%
	D.2_SOFTWARE_ENGINEERING	1,76	5,05%
	D.3_PROGRAMMING_LANGUAGES	0,060	0,17%
	E._DATA	0,121	0,35%
	E.2_DATA_STORAGE_REPRESENTATIONS	0,060	0,17%
	F.3_LOGICS_AND_MEANINGS_OF_PROGRAMS	0,060	0,17%
	H.1_MODELS_AND_PRINCIPLES	0,105	0,30%
	H.2_DATABASE_MANAGEMENT	7,086	20,32%
	H.3_INFORMATION_STORAGE_AND_RETRIEVAL	3,401	9,75%
	H.4_INFORMATION_SYSTEMS_APPLICATIONS	0,060	0,17%
	H.5_INFORMATION_INTERFACES_AND_PRESENTATION__E.G.__HCI_	0,447	1,28%
	I.2_ARTIFICIAL_INTELLIGENCE	0,101	0,29%
	I.7_DOCUMENT_AND_TEXT_PROCESSING	5,174	14,84%

P9	NÃO CLASSIFICADO	2,065	26,18%
	A.0_GENERAL	0,066	0,83%
	B._HARDWARE	0,032	0,41%
	B.7_INTEGRATED_CIRCUITS	0,121	1,53%
	B.8_PERFORMANCE_AND_RELIABILITY	0,123	1,56%
	C.1_PROCESSOR_ARCHITECTURES	0,032	0,41%
	C.2_COMPUTER-COMMUNICATION_NETWORKS	4,805	60,92%
	C.3_SPECIAL-PURPOSE_AND_APPLICATION-BASED_SYSTEMS	0,073	0,93%
	D.3_PROGRAMMING_LANGUAGES	0,022	0,28%
	E,_DATA	0,022	0,28%
	E.3_DATA_ENCRYPTION	0,01	0,13%
	H.2_DATABASE_MANAGEMENT	0,019	0,24%
	H.3_INFORMATION_STORAGE_AND_RETRIEVAL	0,01	0,13%
	H.4_INFORMATION_SYSTEMS_APPLICATIONS	0,054	0,68%
	H.5_INFORMATION_INTERFACES_AND_PRESENTATION__E,G.__HCI_	0,01	0,13%
	I.3_COMPUTER_GRAPHICS	0,123	1,56%
	I.5_PATTERN_RECOGNITION	0,01	0,13%
	J.3_LIFE_AND_MEDICAL_SCIENCES	0,01	0,13%
	K.3_COMPUTERS_AND_EDUCATION	0,271	3,44%
	K.4_COMPUTERS_AND_SOCIETY	0,01	0,13%

P10	NÃO CLASSIFICADO	4,944	21,63%
	A.0_GENERAL	0,044	0,19%
	B._HARDWARE	0,055	0,24%
	B.6_LOGIC_DESIGN	0,121	0,53%
	B.M_MISCELLANEOUS	0,302	1,32%
	D.1_PROGRAMMING_TECHNIQUES	0,165	0,72%
	D.2_SOFTWARE_ENGINEERING	0,060	0,26%

	D.3_PROGRAMMING_LANGUAGES	0,092	0,40%
	D.4_OPERATING_SYSTEMS	0,044	0,19%
	E.1_DATA_STRUCTURES	0,022	0,1%
	H._INFORMATION_SYSTEMS	0,022	0,1%
	H.2_DATABASE_MANAGEMENT	0,302	1,32%
	H.3_INFORMATION_STORAGE_AND_RETRIEVAL	0,060	0,26%
	H.5_INFORMATION_INTERFACES_AND_PRESENTATION__E.G._HCI_	4,183	18,30%
	I.3_COMPUTER_GRAPHICS	9,753	42,67%
	I.4_IMAGE_PROCESSING_AND_COMPUTER_VISION	2,448	10,71%
	J.3_LIFE_AND_MEDICAL_SCIENCES	0,121	0,53%
	K.3_COMPUTERS_AND_EDUCATION	0,060	0,26%
	K.6_MANAGEMENT_OF_COMPUTING_AND_INFORMATION_SYSTEMS	0,060	0,26%

P11	NÃO CLASSIFICADO	22.056	4,82%
	A.0_GENERAL	0,022	0,005%
	B._HARDWARE	0,123	0,03%
	B.1_CONTROL_STRUCTURES_AND_MICROPROGRAMMING	2,014	0,44%
	B.4_INPUT_OUTPUT_AND_DATA_COMMUNICATIONS	1,465	0,32%
	B.6_LOGIC_DESIGN	0,181	0,04%
	B.7_INTEGRATED_CIRCUITS	11,47	2,51%
	B.8_PERFORMANCE_AND_RELIABILITY	407,842	89,25%
	B.M_MISCELLANEOUS	0,077	0,02%
	C._COMPUTER_SYSTEMS_ORGANIZATION	1,704	0,37%
	C.0_GENERAL	0,060	0,01%
	C.1_PROCESSOR_ARCHITECTURES	4,608	1%
	C,3_SPECIAL-PURPOSE_AND_APPLICATION-BASED_SYSTEMS	1,725	0,38%
	D,1_PROGRAMMING_TECHNIQUES	0,183	0,04%
	D,2_SOFTWARE_ENGINEERING	1,032	0,23%
	D.4_OPERATING_SYSTEMS	2,133	0,47%
	H.4_INFORMATION_SYSTEMS_APPLICATIONS	0,152	0,03%
	I.5_PATTERN_RECOGNITION	0,055	0,01%
	K.3_COMPUTERS_AND_EDUCATION	0,060	0,01%

P12	NÃO CLASSIFICADO	4,725	60,75%
	A.0_GENERAL	0,033	0,42%
	D.3_PROGRAMMING_LANGUAGES	0,066	0,85%
	E.1_DATA_STRUCTURES	0,022	0,28%
	F.4_MATHEMATICAL_LOGIC_AND_FORMAL_LANGUAGES	0,022	0,28%
	H._INFORMATION_SYSTEMS	0,0549	0,71%
	H.1_MODELS_AND_PRINCIPLES	0,0409	0,53%
	H.2_DATABASE_MANAGEMENT	0,335	4,31%
	H.3_INFORMATION_STORAGE_AND_RETRIEVAL	0,0955	1,23%
	H.4_INFORMATION_SYSTEMS_APPLICATIONS	0,0644	0,83%
	H.5_INFORMATION_INTERFACES_AND_PRESENTATION__E.G._HCI_	0,0701	0,90%
	I.2_ARTIFICIAL_INTELLIGENCE	1,626	20,91%
	I.3_COMPUTER_GRAPHICS	0,033	0,42%
	I.7_DOCUMENT_AND_TEXT_PROCESSING	0,303	3,9%
	J.1_ADMINISTRATIVE_DATA_PROCESSING	0,041	0,53%
	J.3_LIFE_AND_MEDICAL_SCIENCES	0,121	1,56%
	K.6_MANAGEMENT_OF_COMPUTING_AND_INFORMATION_SYSTEMS	0,125	1,60%