Pós-Graduação em Ciência da Computação

“UMA ABORDAGEM PARA ROTEAMENTO DE CONSULTAS EM PDMS BASEADA EM ASPECTOS SEMÂNTICOS E DE QUALIDADE”

Por

Crishane Azevedo Freire

Tese de Doutorado

Universidade Federal de Pernambuco
posgraduacao@cin.ufpe.br
www.cin.ufpe.br/~posgraduacao

RECIFE
2014

UNIVERSIDADE FEDERAL DE PERNAMBUCO
CENTRO DE INFORMÁTICA
PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO

CRISHANE AZEVEDO FREIRE

"UMA ABORDAGEM PARA ROTEAMENTO DE CONSULTAS EM PDMS BASEADA EM ASPECTOS SEMÂNTICOS E DE QUALIDADE"

ESTE TRABALHO FOI APRESENTADO À PÓS-GRADUAÇÃO EM CIÊNCIA DA COMPUTAÇÃO DO CENTRO DE INFORMÁTICA DA UNIVERSIDADE FEDERAL DE PERNAMBUCO COMO REQUISITO PARCIAL PARA OBTENÇÃO DO GRAU DE DOUTOR EM CIÊNCIA DA COMPUTAÇÃO.

ORIENTADORA: DRA. ANA CAROLINA SALGADO

CO-ORIENTADORA: DRA. DAMIRES YLUSKA DE SOUZA FERNANDES

RECIFE
2014

Catalogação na fonte
Bibliotecária Jane Souto Maior, CRB4-571

Freire, Crishane Azevedo
Uma abordagem para roteamento de consultas em PDMS baseada em aspectos semânticos e de qualidade / Crishane Azevedo Freire. - Recife: O Autor, 2014.
148 f., fig., tab., quadro

Orientador: Ana Carolina Salgado.
Tese (doutorado) - Universidade Federal de Pernambuco. CIn, Ciência da Computação, 2014.

Inclui referências e apêndice.

1. Banco de dados. 2. Roteamento. 3. Semântica. I. Salgado, Ana Carolina (orientadora). I. Título.

025.04 CDD (23. ed.) MEI2014 – 129

Tese de Doutorado apresentada por Crishane Azevedo Freire à Pós Graduação em Ciência da Computação do Centro de Informática da Universidade Federal de Pernambuco, sob o título "**Uma Abordagem para Roteamento de Consultas em PDMS baseada em Aspectos Semânticos e de Qualidade**" orientada pela **Profa. Ana Carolina Brandão Salgado** e aprovada pela Banca Examinadora formada pelos professores

Patrícia Cabral de Azevedo Restelli Tedesco
Centro de Informática / UFPE

Profa. Bernadette Farias Lóscio
Centro de Informática / UFPE

Profa. Ana Maria de Carvalho Moura
Coordenação de Ciência da Computação / LNCC

Prof. Ronaldo dos Santos Mello
Departamento de Informática e de Estatística / UFSC

Profa. Maria da Conceição Moraes Batista
Departamento de Estatística e Informática / UFRPE

Visto e permitida a impressão.
Recife, 03 de julho de 2014.

Profa. Edna Natividade da Silva Barros
Coordenadora da Pós-Graduação em Ciência da Computação do
Centro de Informática da Universidade Federal de Pernambuco.

Aos meus filhos Erik e Bruno, preciosidades de minha vida.

Agradecimentos

É importante concluir um trabalho e poder agradecer aos que sempre estiveram presentes ou que de alguma forma contribuíram com a sua realização. Assim, dirijo os meus agradecimentos:

- À professora Ana Carolina Salgado, minha orientadora, a quem devo muito pelas discussões e ensinamentos. Fico feliz em ter trabalhado com você e de ter conhecido a pessoa forte, dedicada e comprometida com tudo o que faz. Carol, você é realmente admirável! Muito obrigada pela disponibilidade, paciência e conselhos recebidos ao longo desta jornada.

- À professora Damires Yluska de Souza, minha co-orientadora, com quem muitas vezes compartilhei as minhas ansiedades. Quero dizer que tudo ficou melhor com a sua ajuda. Sempre disponível e paciente não dava pra ficar muito tempo sem te consultar. Agradeço também pelo grande apoio nas correções e revisões dos artigos.

- Aos professores do CIn (Centro de Informática), em especial aqueles que contribuíram ao longo das disciplinas do doutorado, Ana Carolina Salgado, Fernando Fonseca, Patrícia Tedesco e Valéria Times. Meus sinceros agradecimentos.

- À professora Maria da Conceição Batista, a querida Ceça, nos ensinamentos e discussões sobre Qualidade da Informação e a professora Bernadette Lóscio pelos comentários e revisões de artigo.

- Ao amigo Edemberg, parceiro no IFPB e no doutorado, pelas longas conversas durante nossas viagens para UFPE e discussões valorosas sobre este trabalho. A sua companhia tornou mais "leve" esta jornada, principalmente nos momentos em que necessitava de ânimo e descontração.

- Ao meu pai Edvaldo Carlos Freire (*in memoriam*), pelos ensinamentos e formação que guiaram o meu crescimento pessoal. Apesar do pouco tempo que passamos juntos, muitas foram as lições de vida aprendidas. A minha mãe Heide Azevedo Freire pela orientação, força e incentivo ao longo da minha formação pessoal e profissional.

- À minha irmã Sheila Azevedo Freire e aos meus sobrinhos Caio, Bárbara e Ingra pela alegria da convivência e ajuda constante nas mais diferentes etapas da minha vida.

- Ao meu esposo, Laerte Ramos, e meus filhos Erik e Bruno pela paciência, compreensão e amor recebidos de forma incondicional. Em especial a Laerte, pelo apoio e incentivo em todas as etapas deste doutorado.

- Aos alunos do CIn, Nicolle Cysneiros, Pedro Henrique e Bruna Carolina que contribuíram na etapa de codificação deste trabalho.

- Aos colegas do IFPB e aos membros do projeto DINTER IFPB-UFPE realizado com o apoio da CAPES.

- Acima de tudo a Deus, por ter me dado forças e saúde durante a realização desta tese.

Resumo

Os *Peer Data Management Systems* (PDMS) são sistemas que permitem o gerenciamento de dados estruturados e semiestruturados em ambientes Ponto-a-Ponto (P2P). Nestes sistemas, cada ponto corresponde a uma fonte de dados cujo esquema representa os dados que se deseja compartilhar na rede. Pontos estão conectados por meio de mapeamentos (correspondências semânticas entre os esquemas dos pontos) estabelecendo uma vizinhança semântica entre eles.

O processamento de consultas é reconhecido como o principal serviço que um PDMS pode prover. Uma etapa importante deste processo está relacionada ao roteamento da consulta, ou seja, a habilidade do sistema de identificar, selecionar e fazer o encaminhamento da consulta ao melhor conjunto de pontos capazes de respondê-la. A cada encaminhamento a consulta precisa ser reformulada, ou seja, reescrita de acordo com o esquema do ponto destino. Na reformulação, termos (conceitos e/ou propriedades utilizados na formulação da consulta) podem ser perdidos por não possuírem correspondentes exatos no esquema do ponto destino. Neste caso, estratégias de reformulação que usam expansão buscam melhorar a consulta adicionando novos termos com o objetivo de tornar a consulta mais abrangente e evitar a ausência de resultados. Ao longo do roteamento, termos perdidos ou adicionados, a cada reformulação, podem levar à perda semântica da consulta original.

Neste trabalho apresentamos a *SemRouting*, uma abordagem para o roteamento de consultas em PDMS baseada no uso de aspectos semânticos e de qualidade. A abordagem *SemRouting* compreende uma estratégia para identificação e seleção do melhor conjunto de pontos, um modelo para representação das informações semânticas e de qualidade e uma estratégia para análise e preservação da semântica da consulta original durante o roteamento.

Para avaliação da abordagem, experimentos foram realizados e os resultados discutidos e apresentados. A análise dos resultados produzidos nos experimentos mostra que as estratégias adotadas na abordagem *SemRouting* confirmam as hipóteses levantadas nesta tese em relação à preservação semântica da consulta e à seleção do melhor conjunto de pontos durante o roteamento da consulta.

Palavras-chave: Roteamento Semântico de Consulta. Informação Semântica. Qualidade da Informação. *Peer Data Management System*.

Abstract

Peer Data Management Systems (PDMS) are P2P applications that allow structured and semi-structured data management. In these systems, each peer represents an autonomous data source, which exports either its entire data schema or only a portion of it. Such schemas, named exported schemas, represent the data to be shared with the other peers. To enable data sharing, mappings (i.e., correspondences) between elements of the exported schemas are generated and maintained in such a way that peers directly connected are called semantic neighbours.

A key issue in query answering in PDMS regards query routing, i.e., the process of identifying the most relevant peers among the ones available in the network to answer a given submitted query. Each peer individually must decide, based on its local knowledge, to which existing neighbors the query should be forwarded. From the query submission peer, the original query is successively rewritten into queries over the peers, according to the correspondences between the original peer and the target ones. In this process, some of the original query terms may be lost. Query reformulation strategies by means of query expansion have been used in order to enhance queries. However, close terms, which are added to the reformulated queries, may produce a gap between the original query semantics and their expanded forms. In this sense, we argue that these expanded reformulated queries (which are produced with lost or added terms with respect to the original one) may indeed represent a semantic loss of the original query.

In this work, we propose an approach, named *SemRouting*, which combines the use of semantic information and Information Quality (IQ) in order to enhance query routing processes in PDMS. The *SemRouting* is composed by three main strategies: (i) query semantic loss analysis and query semantics preservation, (ii) semantic information and IQ representation model, and (iii) selection of thebest peers to route queries to.

To evaluate our proposal, experiments were carried out. In this work, we discussthe results we have obtained. The results obtained in the experiments show that the strategies adopted in the SemRouting approach confirm the research hypotheses with respect tothe query semantics preservation and the selection of the best set of peers in the query routing processes.

Keywords: Query Routing. Semantic Information. Information Quality. Peer Data Management System.

Lista de Figuras

Lista de Quadros e Tabelas

Lista de Abreviaturas

DHT	*Distributed Hash Table*
OWL	*Ontolgy Web Language*
P2P	*Peer-to-Peer*
PDMS	*Peer-to-Peer Data Management System*
SON	*Semantic Overlay Network*
TTL	*Time-to-live*

Sumário

Capítulo 1

Introdução

O grande volume de informações disponíveis na Web fez surgir uma demanda por sistemas que permitam o compartilhamento e o acesso a estas informações de forma rápida. Neste cenário, diversas soluções surgiram, dentre elas, os Sistemas de Integração de Dados [Halevy *et al.* 2006; Lóscio 2003], os *Peer Data Management Systems* (PDMS) [Karvounarakis *et al.* 2013; Roth 2012; Pires 2009; Halevy *et al.* 2004] e os sistemas *pay-as-you-go* [Hatem *et al.* 2010; Sarma *et al.* 2008; Salles *et al.* 2007]. Em se tratando dos PDMS, a arquitetura definida é baseada na tecnologia P2P (*Peer-To-Peer*).

O termo P2P refere-se ao paradigma da computação distribuída em que cada ponto compartilha recursos e serviços de uma maneira autônoma e descentralizada [Bernstein *et al.* 2002]. Em um sistema P2P, pontos comunicam-se por meio de uma rede lógica (*overlay network*) definida no topo de uma rede física [Doval e O'Mahony, 2003]. Dentre os sistemas P2P, os PDMS são não somente um tipo de sistema, mas também uma área completa relacionada ao gerenciamento de dados em ambientes P2P [Roth e Skritek, 2013; Doan *et al.* 2012].

Em um PDMS, não existe um ponto central de coordenação do sistema, todos os pontos são autônomos e armazenam fontes de dados (estruturadas e semi-estruturadas). Cada ponto disponibiliza um esquema (denominado de esquema exportado) que representa os dados que se deseja compartilhar na rede. Pontos estão conectados entre si por meio de mapeamentos (aqui denominados de correspondências) que expressam o relacionamento semântico existente entre

elementos dos esquemas exportados. Pontos conectados entre si por meio de correspondências são denominados vizinhos semânticos [Gennaro *et al.* 2011; Pires *et al.* 2009]. É neste panorama que se encontra definido o escopo deste trabalho.

1.1. Motivação

Em um PDMS, consultas são respondidas parcialmente com os dados locais (armazenados no próprio ponto) e com os dados que podem ser obtidos a partir de caminhos ou rotas existentes por entre pontos vizinhos (ou seja, conexões estabelecidas entre dois pontos por meio de correspondências semânticas pelos quais a consulta poderá ser encaminhada). Para que outros pontos possam ser consultados, a consulta precisa ser reformulada, ou seja, reescrita de acordo com o esquema dos pontos para os quais a consulta será encaminhada. Na reformulação da consulta, o sistema utiliza as correspondências existentes entre os pontos da rede para identificar os termos (conceitos e/ou propriedades associados ao esquema do ponto) a serem utilizados na formulação da consulta [Souza *et al.* 2009].

Um aspecto importante do processamento de consultas em um PDMS está relacionado ao roteamento da consulta, ou seja, a habilidade do sistema de identificar e selecionar o melhor conjunto de pontos (ou seja, pontos classificados como preferíveis de acordo com critérios de qualidade como, por exemplo, relevância, disponibilidade, atualidade) capazes de responder uma determinada consulta que esteja sendo encaminhada [Ismail *et al.* 2011]. Estratégias por inundação (*flooding*) mostram-se ineficientes devido ao aumento no tráfego de informações na rede e no volume de resultados muitas vezes indesejados [Dedzoe 2013]. Assim, fazer o roteamento tomando como base o conhecimento do ponto em relação aos seus vizinhos e o que foi solicitado na consulta, de maneira eficiente, vem se mostrando uma tarefa importante e necessária [Ismail *et al.* 2011]. Estratégias para o roteamento de consultas vêm sendo desenvolvidas ao longo dos anos com o objetivo de tornar mais eficiente este processo [Ismail *et al.* 2011; Roth 2012; Montanelli *et al*. 2010; Li *et al*. 2007; Herschel e Heese, 2005].

Consideramos que no processo de roteamento é preciso que cada ponto, ao receber uma consulta, realize e coordene as seguintes atividades: (i) execução da consulta (pelo ponto que recebe a consulta); (ii) seleção de próximos pontos

(considerando os vizinhos que podem melhor responder a consulta); (iii) reformulação da consulta (para cada ponto vizinho selecionado); (iv) encaminhamento da consulta (depois de reformulada para cada ponto vizinho selecionado); e (v) integração dos resultados das consultas (do próprio ponto e dos vizinhos para os quais uma versão reformulada da consulta tenha sido encaminhada).

Em relação à atividade de seleção, diferentes aspectos podem influenciar a escolha dos pontos como, por exemplo, a disponibilidade ou a relevância dos dados. Assim, cada ponto precisa decidir, com base no conhecimento local que tem da rede, para qual dos seus vizinhos semânticos a consulta deverá ser encaminhada [Haase *et al.* 2008]. Neste cenário, o uso de informações semânticas provenientes do uso do contexto (conjunto de elementos que caracterizam uma situação num dado momento) [Dey 2001; Vieira *et al*. 2011] e de critérios de qualidade [Hose *et al.* 2008] podem permitir que o processo de seleção de pontos e de definição de rotas para a consulta possa ser realizado de forma mais específica e dinâmica. Neste caso, informações sobre a semântica da consulta (descrita pelo conjunto de termos utilizados na formulação da consulta de acordo com o esquema do ponto onde a mesma será executada), as preferências do usuário em relação à consulta, a disponibilidade do ponto e a análise da qualidade da informação dos pontos vizinhos são elementos que podem ser considerados informações contextuais, durante a atividade de seleção de pontos para o roteamento.

Em um PDMS, o processo de roteamento da consulta é influenciado por fatores que estão associados às características do sistema: dinamicidade do ambiente, heterogeneidade das fontes e o grande número de pontos participantes [Aghamahmoodi *et al.* 2014; Montanelli *et al.* 2011].

Em relação à dinamicidade do ambiente, pontos podem entrar ou sair a qualquer momento ou se mostrarem indisponíveis no momento do encaminhamento da consulta [Souza *et al.* 2009]. Da mesma forma, situações de sobrecarga de processamento ou de atividade de manutenção da rede [Silva *et al*. 2013] podem levar o ponto a uma situação de indisponibilidade momentânea.

Considerando a heterogeneidade das fontes, o sucessivo processo de reformulação da consulta, ao longo do roteamento, pode levar a uma perda semântica [Delveroudis e Lekeas, 2009] da consulta original, ou seja, termos definidos na consulta

original podem não ter correspondentes exatos (i.e., equivalentes) nos esquemas dos pontos destino. Nas estratégias de reformulação que consideram expansão, a adição de termos tem sido proposta com o objetivo de ampliar o conjunto de resultados das consultas e impedir o retorno de respostas vazias, o que algumas vezes, por outro lado, pode gerar perda de precisão dos resultados [Campos *et al.* 2013]. Termos incluídos na consulta reformulada, ao longo do roteamento, podem levar a situações de distanciamento semântico entre a consulta reformulada e a consulta original. Neste caso, tais consultas poderiam produzir resultados inadequados à consulta originalmente solicitada pelo usuário.

A seleção de um grande número de fontes nem sempre garante o melhor resultado às respostas do usuário. O volume de dados, gerado pela atividade de seleção de pontos, pode impactar na qualidade do sistema em relação aos resultados, ao custo de integração e ao tempo de resposta [Dong *et al.* 2012]. Determinadas fontes podem possuir baixa relevância em relação à consulta ou baixa capacidade de resposta se comparadas a outras do sistema.

1.2. Definição do Problema

Diante deste cenário, investigamos o problema do roteamento a partir das seguintes questões:

- Como identificar o melhor conjunto de pontos capazes de responder da melhor forma possível à consulta do usuário?
- Como preservar a semântica da consulta original ao longo do roteamento?
- Como evitar a execução de consultas com alta perda semântica?

Com base nas questões mencionadas, definimos o problema do roteamento de consultas em um PDMS da seguinte forma: Seja P={P1,..,Pn} um conjunto de pontos semanticamente relacionados por meio de correspondências e Q uma consulta submetida em um ponto qualquer de P. Nosso problema é fazer o encaminhamento de Q, por entre os pontos do sistema, de forma que o melhor conjunto de pontos possa ser escolhido considerando a semântica da consulta, as preferências do usuário em relação a consulta, a dinamicidade do ambiente e a análise da qualidade da

informação dos pontos. Ao longo do roteamento deve-se preservar, da melhor forma possível, a semântica original de Q e evitar que reformulações de Q com alta perda semântica sejam executadas pelo sistema.

1.3. Objetivo

Propor uma abordagem, denominada *SemRouting*, que considere aspectos semânticos e de qualidade da informação no processo de seleção de pontos e geração de rotas para o encaminhamento de consultas em um PDMS. A abordagem deverá permitir a preservação e a identificação do nível de perda semântica da consulta original ao longo do roteamento.

Para cumprir o objetivo geral, foram considerados alguns objetivos específicos:

- Investigar como informações semânticas (provenientes do uso de ontologias e do contexto) e de qualidade da informação podem ser utilizadas para tornar mais eficiente o roteamento de consultas em PDMS.
- Verificar como o conhecimento local de cada ponto que compõe o PDMS pode ser expandido a partir do conhecimento proveniente do uso de contexto e de critérios de qualidade;
- Elaborar um modelo para representação das informações semânticas e de qualidade;
- Definir um mecanismo para preservação semântica da consulta original ao longo do roteamento;
- Identificar a perda semântica da consulta original ao longo do roteamento.
- Definir uma estratégia para o roteamento de consultas em um PDMS que preserve a semântica da consulta original ao longo do roteamento e faça a seleção do melhor conjunto de pontos de acordo com o conhecimento local do ponto;

1.4. Hipóteses da Tese

Esta tese possui três hipóteses de pesquisa:

H1. O uso de informações semânticas combinadas à qualidade da informação produz resultados que permitem ao sistema obter um melhor conjunto de pontos capazes de responder a consulta do usuário.

H2. Ao longo do roteamento, a eliminação de termos da consulta reformulada que não possuem correspondência semântica com os termos da consulta original permite preservar a semântica da consulta original.

H3. A análise da perda semântica da consulta reformulada, em relação à consulta original, permite identificar os pontos onde a consulta deve ser executada durante o roteamento.

Para comprovar cada uma das hipóteses descritas, são realizados experimentos utilizando o protótipo do sistema SPEED (*Semantic PEEr Data Management System*) [Pires 2009]. O SPEED é um PDMS que adota uma abordagem semântica baseada em ontologias e informações contextuais [Souza 2009] com o propósito de prover soluções para problemas críticos de gerenciamento de dados em sistemas P2P, tais como: conectividade, mapeamentos e processamento de consultas. Os experimentos apresentam os resultados obtidos com a nossa proposta e avaliam as hipóteses definidas.

1.5. Contribuições Esperadas

Esta tese espera alcançar as seguintes contribuições:

- Especificação e implementação do conjunto de critérios de qualidade (relevância, capacidade de resposta, atualidade, tempo de execução) utilizado para análise da qualidade da informação (QI) do ponto.
- Especificação e implementação de um modelo para representação das informações semânticas e de QI com o objetivo de expandir o conhecimento local do ponto e facilitar o processo de roteamento da consulta em um PDMS.

- Especificação e implementação de uma estratégia para preservação da semântica da consulta original durante o roteamento.
- Especificação e implementação de um mecanismo de parada baseado na análise da semântica da consulta evitando que consultas com alta perda semântica sejam encaminhadas.
- Especificação, implementação e avaliação por meio de experimentos da abordagem definida para o roteamento de consultas.

1.6. Organização da Tese

Este documento está organizado em seis capítulos. Além do presente capítulo, os demais relatam o seguinte:

- O Capítulo 2 mostra a fundamentação teórica desta tese. Definições sobre Ontologia, Contexto e Qualidade da Informação são apresentadas. Define os sistemas para o gerenciamento de dados em ambientes P2P (PDMS) e mostra suas principais características. Aborda técnicas tradicionais para o roteamento de consultas e discute as questões relacionadas à análise da qualidade da informação baseada em múltiplos critérios.
- O Capítulo 3 apresenta algumas estratégias existentes para o roteamento de consultas em PDMS consideradas relevantes ao tema desta tese. O uso de aspectos semânticos e de qualidade são analisados em cada uma das estratégias relacionadas. Ao final uma análise comparativa é realizada de acordo com as questões de pesquisa consideradas neste trabalho.
- O Capítulo 4 propõe a abordagem *SemRouting*. Inicialmente define o problema da perda semântica da consulta original ao longo do roteamento e apresenta o mecanismo de poda para preservação desta semântica. Descreve o uso das informações semânticas e relaciona os critérios de qualidade. Mostra o modelo para representação das informações semânticas e de qualidade exemplificando o seu uso durante a atividade de seleção de pontos. Define o mecanismo de parada e mostra um exemplo completo do uso da abordagem.

- O Capítulo 5 discute aspectos relacionados à implementação, aos resultados alcançados e aos experimentos que comprovam as hipóteses indicadas neste capítulo.
- O Capítulo 6 finaliza o documento apresentando as conclusões, contribuições da pesquisa e indicando alguns trabalhos futuros.

Capítulo 2

Fundamentação Teórica

Este capítulo introduz conceitos associados ao tema desta tese. Para isso inicia apresentando uma visão geral sobre Semântica e Qualidade da Informação. Define os sistemas para o gerenciamento de dados em ambientes P2P (PDMS) e mostra suas principais características. Destaca o uso dos mapeamentos e sua importância na reformulação da consulta. Apresenta técnicas tradicionais para o processo de roteamento de consulta. Relaciona questões associadas à qualidade da informação, ao problema de decisão baseada em múltiplos critérios e ao uso de contexto no cenário de roteamento de consultas. Por fim, faz algumas considerações que servirão como referência para análise e classificação das estratégias de roteamento a serem descritas no Capítulo 3.

2.1. Semântica

Em ambientes distribuídos, dinâmicos e heterogêneos, a semântica vem sendo utilizada como forma de prover o significado dos dados e facilitar a sua interpretação em uma situação específica [Orsi e Tanca, 2012; Bolchini *et al*. 2009]. É possível obter este conhecimento semântico, por exemplo, a partir do uso de *ontologias* e de *contexto* [Souza *et al.* 2009].

2.1.1. Ontologia

Ontologia é uma especificação formal e explícita de uma conceituação compartilhada [Gruber 1993]. Uma ontologia fornece uma meta-informação que descreve a semântica dos dados e é utilizada de forma a facilitar a compreensão comum, o reuso e o compartilhamento de um conhecimento [Euzenat e Shvaiko, 2013].

Com as ontologias é possível dar significado aos dados favorecendo a integração e, consequentemente, o gerenciamento de dados em ambientes distribuídos [Pires *et al.* 2011]. Sendo assim, nesses ambientes, as ontologias têm sido usadas para alguns propósitos, dentre eles [Xiao 2006]: (i) **representação de metadados**: cada fonte de dados é representada por uma ontologia local; (ii) **conceituação global**: uma ontologia global pode ser usada para prover uma visão conceitual sobre os esquemas heterogêneos das fontes; (iii) **suporte às consultas de alto nível**: dada uma ontologia global, os usuários podem formular consultas sem conhecimento específico das diferentes fontes de dados.

Quando múltiplas ontologias são simultaneamente usadas, elas podem ter diferentes tipos de heterogeneidades [Euzenat e Shvaiko 2013]: Sintática (quando as ontologias estão representadas em diferentes formalismos, por exemplo, OWL e F-Logic,); Terminológica (quando ocorrem variações em nomes usados para se referir às mesmas entidades em ontologias diferentes, por exemplo, *Paper vs Artigo);* Conceitual (heterogeneidade semântica, diferenças encontradas na modelagem de um mesmo domínio de interesse) e Semiótica (diferenças relacionadas a como as entidades são interpretadas pelas pessoas).

Para tratar com a heterogeneidade, são utilizados processos de associação ou correspondência (*Matching*) entre os elementos das ontologias. Como resultado deste processo de *matching* é gerado um alinhamento, isto é, um conjunto de correspondências que indicam a associação lógica entre os elementos das ontologias [Euzenat e Shvaiko 2013]. Processos de *matching* contribuem diretamente para operações de integração e consulta das fontes de dados do sistema.

Em PDMS, o uso de ontologias incluem: o desenvolvimento de OPDMS (*Ontology Peer Data Manegement System*), onde ontologias são usadas nas questões relacionadas ao gerenciamento de dados, como representação de esquemas, agrupamentos e mapeamentos [Pires *et al.* 2012, Montanelli *et al.* 2011]; processos de *matching*, usado para estabelecer as associações entre ontologias [Mazak *et al.* 2010; Pires *et al.* 2009] e no processamento da consulta [Souza *et al.* 2009, Sassatelli 2009]. Como forma de tratar a heterogeneidade semântica, pesquisas têm considerado o uso de ontologias como uma forma de dar ao sistema um domínio de referência. Nesse caso, uma ontologia de domínio pode ser usada como referência semântica ou

conhecimento de base (*background knowledge*) para melhorar, por exemplo, processos de *matching* [Souza *et al.* 2011].

2.1.2. Contexto

O conceito de contexto tem sido objeto de investigação há vários anos em algumas comunidades científicas, tais como Lingüística e Psicologia Cognitiva. Na comunidade de Ciência da Computação pode-se observar importantes contribuições para o seu entendimento e formalização, particularmente em trabalhos da área de Inteligência Artificial [Calvi *et al.* 2005]. Várias definições podem ser encontradas na literatura, mas a mais referenciada é aquela apresentada por Dey [2000]:

> *"Contexto é qualquer informação que pode ser usada para caracterizar uma situação de uma entidade. Uma entidade é uma pessoa, um lugar, ou um objeto que é considerado relevante para a interação entre um usuário e uma aplicação, incluindo o próprio usuário e a própria aplicação."*

É importante destacar que a definição de Dey é bastante ampla, e, dependendo da área de aplicação e do uso de contexto, as definições podem se tornar mais específicas, uma vez que o contexto está sempre relacionado a um domínio, um foco ou um ponto de vista [Vieira *et al.* 2011].

Segundo Brézillon [2013], o contexto é considerado como um espaço de conhecimento compartilhado, que é explorado repetidamente pelos participantes em uma interação. O contexto está sempre relacionado a um foco, onde um foco pode ser um passo na execução de uma tarefa ou uma decisão. É ele, o foco, que permite determinar quais elementos deveriam ser instanciados e usados para compor o contexto.

Vieira e seu grupo [2011] definem contexto com base nas definições de Dey e Brézillon e fazem uma clara distinção entre os conceitos de contexto e elemento contextual (CE). **Elemento contextual** é qualquer parte da informação que permite ao agente, um indivíduo ou software, caracterizar entidades de um domínio. O **contexto** é o conjunto de elementos contextuais instanciados necessários à tarefa que está sendo executada na interação entre um agente e uma aplicação. Mais do que isso, os elementos que compõem um contexto têm um relacionamento relevante com a tarefa

que o agente está executando. Pode-se observar que o elemento contextual é estável e pode ser definido em tempo de projeto, enquanto que o contexto é dinâmico e deve ser construído em tempo de execução, quando uma interação ocorre ou quando uma tarefa é realizada.

Contexto é o conhecimento que permite definir o que é ou não relevante em uma dada situação. Compreendendo o contexto, o sistema pode se adaptar e mudar seu comportamento, ou seja, sua sequência de ações, o estilo das interações e o tipo da informação fornecida aos usuários, em circunstâncias diversas [Salgado *et al.* 2009]. É possível representar o contexto a partir de diferentes técnicas como, por exemplo, Grafos Contextuais [Tahir e Brézillon 2013], Mapas de Tópicos [Lee e Segev, 2012] e Ontologias [Bandara *et al*. 2013; Souza *et al.* 2012] .

Dentre as técnicas existentes, ontologias têm sido consideradas como uma forma de expressar a semântica, compartilhar conhecimento e permitir o raciocínio sobre seus elementos [Souza *et al.*2012; Wang *et al.* 2004]. Ontologias permitem formalizar a semântica dos elementos contextuais e a definição de regras de contexto considerando os elementos contextuais [Bandara *et al.*2013; Souza *et al.* 2012].

Em ambientes distribuídos e dinâmicos, é possível fazer uso do contexto para vários processos, incluindo: no processamento da consulta, onde a consulta é expandida a partir do contexto obtido no momento de sua submissão [Souza *et al.* 2009], na integração de resultados, com o objetivo de definir qual parte da informação é importante de acordo com um contexto específico [Orsi e Tanca, 2012], no processo de conciliação de esquemas, para identificar em que contexto os elementos ocorrem e resolver problemas de verificação semântica entre elementos de esquemas distintos [Belian *et al.* 2010] e na identificação de recursos para compartilhamento [Montanelli *et al.* 2011], considerando o contexto para estabelecer a formação da vizinhança semântica entre pontos de um sistema.

2.2. Qualidade da Informação

O grande número de fontes de informação disponíveis na web e a natureza acessível desta informação por um conjunto diversificado de usuários fez crescer o interesse pelos aspectos relacionados à Qualidade da Informação (QI) neste ambiente [Arazy e Kopak, 2011]. Em seu uso, os dados podem apresentar baixa qualidade tanto

por não refletirem a realidade ou por serem mal utilizados e mal entendidos pelos usuários. Mesmo dados precisos, se não estiverem disponíveis e dispostos em tempo hábil para sua utilização por parte do usuário interessado, serão de pouco valor [Batista 2008].

A qualidade da informação (QI) está associada ao conjunto de critérios ou dimensões utilizados para indicar o grau de qualidade global associado à informação em um determinado sistema [Pipino *et al.* 2002]. Existe uma definição comum na área de que a QI está relacionada ao termo "adequacão ao uso" (do inglês, *fitness for use*), ou seja, a informação é considerada adequada ao uso na perspectiva da tarefa que está sendo realizada [Wang e Strong 1996].

A QI é um aspecto multidimensional baseado em um conjunto de critérios ou dimensões. Uma das classificações mais referenciadas é apresentada no trabalho de Wang e Strong (1996), resumida a seguir:

(i) intrínseca – diz respeito à qualidade do dado. Inclui os seguintes critérios: credibilidade, acurácia, objetividade, reputação.

(ii) contextual – qualidade do dado relacionado à tarefa. Inclui os seguintes critérios: valor agregado, relevância, temporalidade, completude, quantidade apropriada.

(iii) representação – está relacionada ao formato e significado do dado. Inclui os seguintes critérios: interpretação, facilidade de entendimento, representação concisa, representação consistente.

(iv) acessibilidade - está relacionada à disponibilidade do dado. Inclui os seguintes critérios: acessibilidade, segurança de acesso.

Como a qualidade é uma condição dependente de seu uso e, muitas vezes, do ponto de vista do usuário, é necessário que possam ser estabelecidos mecanismos de medição desses critérios. Alguns critérios podem ser medidos de forma direta, quantitativamente como, por exemplo, o critério de temporalidade. No entanto, outros critérios são subjetivos a exemplo da reputação cuja medição, em alguns casos, depende do *feedback* do usuário em relação aos resultados obtidos em suas respostas. Sendo assim, para cada critério é preciso estabelecer formas de como avaliar ou medir estes resultados.

De acordo com Malakooti e seu grupo (2006), produzir uma análise da QI de acordo com um conjunto de critérios é considerado um problema de decisão com múltiplos critérios (*MCDM - Multi-Criteria Decision Making*). Diferentes métodos e técnicas para resolução de problemas MCDM são utilizados para análise e classificação de pontos (fontes de dados) [Naumann e Roker, 2000]. Dentre os mais conhecidos podemos destacar: *SAW* (*Simple Additive Weighting*) [Hwang e Yoon, 1981; Naumann 1998], *TOPSIS* (*Technique for Order Preference by Similarity to Ideal Solution)* [Hwang e Yoon, 1981; Naumann 1998] e *AHP* (*Analytical Hierarchy Process*) [Saaty 1988; Naumann 1998].

O método *SAW*, por sua simplicidade, é um dos mais populares e mais utilizados para classificação de fontes. Este método consiste na geração de um valor único para análise da QI considerando a soma ponderada das medidas de qualidade associadas a cada ponto. Para produzir este valor único é preciso realizar três etapas: normalizar as medidas de qualidade entre todos os pontos de forma a torná-las comparáveis entre si, aplicar os pesos relativos à importância de cada critério e somar as medidas ponderadas de cada ponto para que se possa obter o valor único de QI.

Para normalização das medidas é considerado o impacto do critério em relação ao ponto, ou seja, se o critério é de qualidade ou positivo (por exemplo, reputação, relevância) ou se o critério é de custo ou negativo (por exemplo, data de atualização do ponto, tempo de resposta). Quanto maior a medida de um critério positivo, melhor é o ponto em relação ao critério e, quanto maior a medida de um critério negativo, pior é o ponto. Depois de normalizados, os critérios são ponderados e somados produzindo um valor único de QI para cada ponto. Ao final, uma classificação é gerada de acordo este valor.

O método *TOPSIS* consiste em estabelecer uma classificação para os pontos utilizando como referência uma solução ideal e uma solução ideal negativa. Neste caso, terá melhor classificação o ponto que estiver mais próximo da solução ideal e mais distante da solução ideal negativa. A solução ideal representa uma alternativa formada pelo conjunto das melhores medidas para cada critério e a solução ideal-negativa é uma alternativa formada com as piores medidas. Neste método, pesos também determinam a preferência de critérios na composição final da QI.

O método *AHP* consiste em produzir uma classificação para os pontos a partir da análise de uma hierarquia de objetivos. Por exemplo, no caso da QI dos pontos poderíamos produzir uma medida a partir da análise de dois sub-objetivos: qualidade e custo. Neste caso, cada sub-objetivo é avaliado como um fator de decisão. As alternativas de um mesmo sub-objetivo são comparadas em pares para avaliar a importância ou preferência relativa entre sub-objetivos. Pesos associados à importância do critério em relação à fonte também são utilizados para ponderar os resultados. Os resultados das comparações são representados em uma matriz, onde a comparação entre alternativas (medidas de critérios associadas a cada ponto) resulta em um valor entre 1 (igual importância) e 9 (extremamente importante). Resolvendo os sub-objetivos se atinge a classificação final dos pontos em relação ao objetivo inicial que avalia cada fonte a partir dos critérios de qualidade e custo.

Outros métodos podem ser encontrados na literatura [del Pilar Angeles *et al.* 2010; Wenning 2010; Naumann e Roker, 2000]. Como ilustração, na área de redes de comunicação, técnicas de MCDM também vêm sendo utilizadas na identificação dos melhores recursos e rotas para o encaminhamento de mensagens [Omheni *et al.* 2014; Márquez-Barja *et al.* 2011; Wenning 2010].

Particularmente, em um ambiente distribuído e dinâmico, existem (no mínimo) três fatores que podem influenciar a resposta para uma dada consulta do usuário, assim como a sua qualidade [Shvaiko *et al.* 2008; Zaihrayeu 2006]:

- Rede – pontos podem modificar os dados de suas fontes, de seus esquemas, redefinir mapeamentos entre vizinhos, e novos pontos podem entrar ou sair do sistema a qualquer momento. Logo, a mesma consulta submetida em um dado ponto, mas em outro instante, poderá fornecer diferentes respostas com medidas de qualidade distintas.
- Ponto – mapeamentos são estabelecidos de várias formas entre pontos. Portanto, a mesma consulta submetida no mesmo instante, em diferentes pontos, resultará em diversos grafos de propagação da consulta. Consequentemente, os resultados poderão ser diferentes e com medidas de qualidade distintas.

- Consulta – diferentes consultas submetidas no mesmo ponto poderão resultar em diferentes grafos de propagação de consulta e, assim, produzir resultados variados com medidas de qualidade distintas.

Assim, vários trabalhos vêm destacando a necessidade de obter e analisar a QI como forma de melhorar o processamento da consulta ([Roth 2012], [Montanelli *et al* 2011],[Kantere *et al. 2011*], [Zaihrayeu 2006], [Zhuge *et al*. 2005]).

2.3. PDMS

Os *Peer Data Management Systems* (PDMS) são sistemas que permitem o gerenciamento de dados estruturados e semi-estruturados em ambientes P2P [Doan *et al.* 2012]. Cada ponto, nesse sistema, pode agir não só como uma fonte de dados disponível para consulta, mas também como um ponto para integração dos resultados provenientes das consultas executadas em outros pontos do sistema [Doan *et al.* 2012].

Sistemas baseados no paradigma P2P caracterizam-se não somente pela capacidade de descentralização e compartilhamento de recursos. Outras características estão associadas a estes sistemas [Sassatelli 2009]: **escalabilidade**, ou seja, crescimento e distribuição de pontos na rede; **autonomia**, onde pontos podem decidir sobre quais recursos compartilhar; **auto-organização**, isto é, a rede se organiza em função da entrada e saída de pontos, entre outras.

Em um PDMS, pontos comunicam-se por meio de uma rede virtual (lógica), denominada *overlay*, que funciona sobre uma rede física. Assim, a topologia da rede define a forma de comunicação entre os pontos [Doval e O'Mahony 2003, Schlosser *et al.* 2003]. Na Figura 2.1 podemos observar a definição de duas redes *overlays*, formadas por diferentes interconexões de pontos, sobre a infra-estrutura de uma rede física.

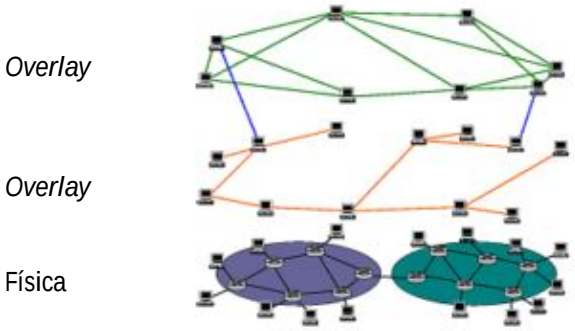


Figura 2.1 - Exemplo de rede física e *overlay* [Pires 2007]

Existem diversas classificações para as topologias P2P [Lv *et al.* 2002, Fiorano 2014, Schollmeier 2001]. De acordo com sua estrutura, as redes P2P podem ser assim classificadas como [Theotokis e Spinellis, 2004]:

- Não-estruturada – a rede não possui um servidor central, assim como o controle sobre a topologia e localização de recurso (por exemplo, dados ou serviços);
- Estruturada – a localização dos recursos é feita de forma mais eficiente. A identificação e endereço dos recursos normalmente são mantidos em tabelas de roteamento distribuído (*Distributed Hash Tables - DHT*);
- Fracamente estruturada – essa categoria está entre as topologias estruturadas e não-estruturadas. É caracterizada por não especificar completamente a localização do recurso armazenado, embora essa localização possa ser obtida por meio de seus algoritmos de roteamento.

Quanto ao nível de centralização, as redes P2P podem também ser classificadas da seguinte forma [Theotokis e Spinellis, 2004]:

- Pura – não possui controle central, todos os pontos executam as mesmas tarefas, tanto como cliente quanto como servidor. Por essa razão cada ponto também é chamado de "*servents*", ou seja, *SERvers+cliENTS*. Cada ponto é responsável por manter informações sobre seus recursos podendo, ao receber uma consulta, respondê-la ou encaminhá-la para outros pontos vizinhos com os quais esteja conectado diretamente.

- Híbrida – existe um servidor central que mantém a descrição do que cada ponto participante da rede dispõe. No caso de consultas, o papel do servidor é identificar e retornar o endereço do ponto que possui o dado desejado ao ponto solicitante.
- *Super-peer* ou parcialmente centralizada – alguns pontos chamados *super-peers (SP)*, gerenciam uma determinada quantidade de pontos conectados a ele formando um agrupamento de pontos. Esses *super-peers* agem como servidores centrais mantendo a descrição e o endereço dos arquivos armazenados em cada ponto do seu agrupamento. Os *super-peers* que compõem a rede são interligados entre si.

Em um PDMS, pontos são fontes de dados heterogêneas, cujo conteúdo que deseja compartilhar está representado por um esquema local (denominado *esquema exportado*) associado a um domínio de interesse específico (por exemplo, Educação, Engenharia, Comunicação). Não existe um esquema global que represente o domínio integrado de todo o conhecimento disponível no sistema [Doan *et al.* 2012].

Nas questões relacionadas ao gerenciamento de dados em PDMS, o trabalho de Sassatelli [2009] destaca alguns aspectos a serem considerados para prover o sistema com funcionalidades que garantam o compartilhamento de dados estruturados, semi-estruturados e semanticamente mais ricos:

- Representação dos dados – o modelo a ser adotado para representação do esquema exportado pelos pontos deve ser capaz de fornecer expressividade semântica dos recursos a serem compartilhados, assim como favorecer o uso de uma linguagem de consulta;
- Compartilhamento de dados – permitir compartilhamento de informações entre os pontos mesmo que as fontes de dados disponíveis possuam diferentes esquemas e representações. O processo de integração adotado precisa ser descentralizado e flexível de forma a atender aos requisitos dos PDMS.
- Processamento de Consulta – uma consulta submetida em um determinado ponto pode ser respondida localmente e/ou ser propagada para os vizinhos diretos daquele ponto. Este é um processo complexo e diversas questões precisam ser abordadas quanto à sua eficiência e eficácia. Em ambientes de

compartilhamento de dados distribuídos, uma camada de mediação decompõe as consultas formuladas a partir de um esquema global e as envia para as fontes de dados distribuídas no sistema. Outra forma de consulta é feita integrando e materializando visões de dados a partir de outras fontes de dados. Entretanto, no cenário P2P, um ponto central que integre resultados pode tornar-se um ponto de falha. Neste sentido, por questões de crescimento do ambiente e para preservar a autonomia dos pontos, o processamento de consultas deve ser descentralizado. Considerando que o número de pontos participantes e a disponibilidade dos dados durante uma consulta podem sofrer variações a qualquer tempo, é necessário que a execução do plano de consulta possa ser feita de forma dinâmica.

- Recuperação dos dados – a consulta submetida em um determinado ponto pode precisar obter os dados que estão localizados nos diversos pontos distribuídos no ambiente. Localizar fontes de dados está entre os principais problemas em ambientes P2P. Em alguns sistemas, índices centralizados ou distribuídos são utilizados para armazenar a localização dos dados. Durante as consultas, os índices são consultados para que o ponto, onde o dado desejado está armazenado, possa ser identificado.

Para permitir o compartilhamento e integração dos dados de forma transparente é preciso que as fontes sejam especificadas e relacionadas por meio de mapeamentos para que o sistema saiba como usar os dados [Doan *et al.* 2012].

2.3.1. Mapeamentos em PDMS

Mapeamentos estabelecem relações entre elementos de esquemas de fontes de dados distintas [Doan *et al.* 2012]. Em PDMS, devido ao problema de heterogeneidade de suas fontes, o uso de um modelo comum para representação dos esquemas exportados pelos pontos facilita a definição dos mapeamentos. Alguns exemplos incluem o uso de modelos baseados em esquemas relacionais, XML[1], RDF[2] e ontologias OWL[3] [Euzenat e Shvaiko, 2013]. Entre estes modelos, as ontologias têm sido utilizadas para descrever a semântica dos esquemas das fontes de dados em

[1] http://www.w3.org/xml/
[2] http://www.w3.org/TR/rdf-primer
[3] http://www.w3.org/TR/owl-ref/

diversos ambientes de integração de dados [Chen *et al*. 2012; Pires *et al.* 2011; Montanelli *et al.* 2011; Li e Vuong, 2007; Xiao, 2006]. Elas permitem tratar com os problemas de integração semântica e de heterogeneidade dos dados, auxiliando no processo de integração dos dados entre as diversas fontes distribuídas [Pires *et al*. 2011].

Existem dois tipos de mapeamentos (Figura 2.2) [Herschel *et al.* 2005]: *mapeamento local* - define correspondências entre os elementos disponíveis na fonte de dados e o esquema exportado pelo ponto e; *mapeamento de esquema* - define as correspondências entre os elementos dos esquemas de dois pontos.

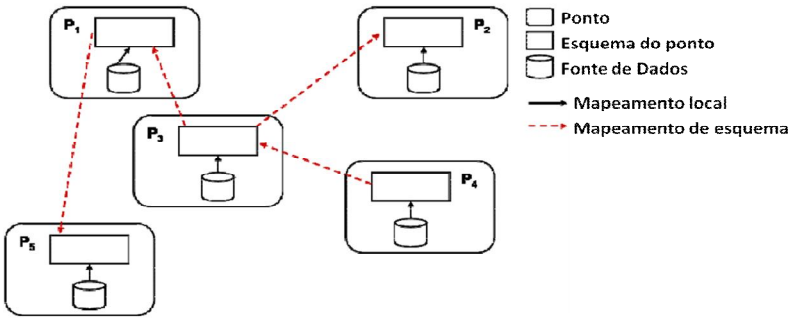


Figura 2.2: Tipos de Mapeamentos em PDMS [Pires 2009]

Para estabelecer os mapeamentos ou correspondências entre os pontos são utilizadas ferramentas para associação de esquemas, alguns deles representados por meio de ontologias. Exemplos são: *SemMatcher* [Pires *et al.* 2009], *OII Harmony* [Seligman *et al.* 2010], *AgreementMaker* [Cruz *et al.* 2009], *H-Match* [Castano *et al.* 2006], *COMA++* [Aumueller *et al.* 2005].

O conjunto de mapeamentos ou correspondências em um PDMS forma sua rede semântica. Pontos conectados por meio de correspondências são denominados *vizinhos semânticos* [Gennaro *et al.* 2011; Pires *et al*. 2009; Montanelli e Castano, 2008]. Assim, ao submeter uma consulta em um ponto P na rede, o sistema pode obter dados relevantes de P e de quaisquer outros pontos desde que a consulta possa ser representada em termos do esquema do ponto onde será executada. Para isto são utilizadas técnicas de reformulação da consulta.

2.3.2. Reformulação da Consulta em PDMS

Os mapeamentos ou correspondências constituem a base para os processos de reformulação da consulta. Neste processo, uma consulta submetida em termos do esquema de um ponto origem é reformulada em uma nova consulta expressa em termos do esquema de um ponto destino, utilizando para isto os diferentes tipos de mapeamentos ou correspondência existentes entre eles [Roth 2012; Kantere *et al.* 2011; Souza *et al.* 2009;].

Em alguns casos, os termos da consulta a ser reformulada nem sempre encontram termos equivalentes no esquema do ponto destino. Além disto, a dinamicidade do ambiente e os caminhos providos pelos mapeamentos costumam gerar longos processos de reformulação. Tais situações costumam dificultar o processamento da consulta, ocasionando em algumas situações perda de informações [Roth 2012].

Encontrar "respostas completas" para uma consulta em um PDMS baseada em uma definição semântica global do sistema é considerado um problema "sem decisão" [Doan *et al.* 2012]. Respostas aceitas pelos usuários em relação as suas consultas costumam ser incompletas. Tais respostas têm sido denominadas de "*good-enough*", ou seja, são consideradas aceitáveis por satisfazerem aos propósitos da consulta submetida em meio ao ambiente dinâmico [Shvaiko *et al.* 2008].

Com o objetivo de melhorar as respostas dadas aos usuários, técnicas para reformulação da consulta por **expansão de termos** [Campos *et al.* 2013; Carpineto *et al.* 2012; Souza *et al.* 2009; Cao *et al.* 2008; Bhogal *et al.* 2007] e **por personalização** [Sharma *et al.* 2014; Alves *et al.* 2013; Campos *et al.* 2013; Tanca *et al.* 2011] têm sido utilizadas em diversas aplicações.

A reformulação da consulta por expansão consiste na adição de termos à consulta que tenham aproximação semântica (ou seja, possuam significado próximo ou similar) com os termos da consulta a ser reformulada [Carpineto e Romano, 2012].

A reformulação da consulta por personalização consiste em prover o usuário com um conjunto de informações relevantes de acordo com critérios e preferências específicas a cada usuário do sistema. Estes critérios podem descrever o domínio de

interesse do usuário, o nível de qualidade dos pontos assim como qualquer outro aspecto que permita personalizar a consulta [Kostadinov 2007].

2.3.3. Roteamento de Consultas em PDMS

Em um PDMS, consultas são respondidas parcialmente com os dados locais (armazenados no próprio ponto) e com os dados que podem ser obtidos a partir de outros vizinhos semânticos. Assim, ao realizar uma consulta é provável que o usuário nem sempre receba respostas exatas e completas. Pontos podem dispor de informações parciais ou até mesmo terem saído da rede o que ocasionaria parcialidade (incompletude) da resposta [Doan *et al.* 2012].

O processo de roteamento de consultas está relacionado à habilidade do sistema de identificar e selecionar um conjunto de pontos capazes de responder a uma dada consulta submetida em um ponto qualquer do sistema. Para que este processo possa ser executado, mecanismos de propagação e de encaminhamento da consulta são necessários e suas implementações podem variar conforme a topologia da rede. De acordo com Ahmed e Boutaba (2011), algoritmos de roteamento podem ser classificados em duas categorias:

- *Uninformed* ou *blind search* – as consultas são enviadas aos pontos sem nenhum conhecimento prévio quanto à capacidade de resposta do ponto em relação à consulta. Assim, quando um ponto recebe uma consulta, ele a encaminha para seus vizinhos até que o resultado seja alcançado. A propagação da consulta é limitada por um mecanismo de *time-to-live* (TTL). O TTL é um valor indicativo do número máximo de saltos para encaminhamento da consulta entre os pontos vizinhos. Quando o TTL atinge zero (TTL=0), o processo de propagação da consulta é interrompido mesmo que nenhum resultado tenha sido alcançado. *Flooding* [Gnutella 2014], *Random walk* [Qin *et al*. 2002] e *Iterative deepening* [Yang e Garcia-Moina, 2002] são algoritmos que pertencem a esta categoria. Apesar da rápida propagação da consulta esta técnica gera um grande volume no tráfego de mensagens na rede.

- *Informed search* – utiliza informações presentes na consulta para identificar os pontos que possam melhor respondê-las. Algoritmos baseados em índices DHT (Distributed Hash Table) [Stoica *et al.* 2001; *Zhao et al.* 2004] e Bloom

Filters [Bloom 1970; Petrakis *et al.* 2004] pertencem a esta categoria. Nestes, consultas por palavras-chave podem ser realizadas com mais facilidade. Uma desvantagem dessa técnica é a dificuldade em manter as estruturas de índices necessárias ao roteamento das consultas e responder às consultas mais complexas.

O processo de roteamento pode ser favorecido pela forma como a vizinhança foi estabelecida pelo sistema. Em um PDMS, pontos podem compartilhar dados de domínios de conhecimentos distintos. A formação de **comunidades semânticas** está relacionada com a capacidade de agrupar dinamicamente os pontos com interesses semelhantes em organizações estruturadas, a partir das correspondências semânticas, favorecendo o processo de roteamento da consulta [Pires *et al.* 2011; Montanelli *et al.* 2011].

A QI no Roteamento de Consultas

Em relação aos processos de roteamento é possível que a decisão da escolha dos melhores pontos para o encaminhamento da consulta seja influenciada por vários critérios de qualidade (por exemplo, requisitos de confiabilidade, relevância, capacidade de resposta, tempo de execução, atualidade). Obter e combinar informações originárias dos pontos em relação à QI pode e deve ser levada em conta na seleção de pontos durante o processo de roteamento [Roth 2012]. A QI possui múltiplas dimensões o que torna difícil a atividade de comparar diretamente os pontos com respeito a um conjunto de critérios e construir uma classificação em relação aos mesmos [Naumann 2001].

O Contexto no Roteamento de Consultas

O roteamento baseado em contexto é uma área de pesquisa referenciada no cenário das redes de comunicação [Sassi *et al.* 2013; Pillac *et al.* 2013; Byrne *et al.* 2011]. Em PDMS o contexto vem sendo utilizado em diferentes trabalhos com o objetivo de melhorar alguns processos, entre eles os relacionados ao processamento da consulta [Alves *et al.* 2013; Souza *et al.* 2011, Tanca *et al.* 2011] e a formação de comunidades semânticas [Montanelli *et al.* 2011].

Em ambientes dinâmicos, o roteamento de consultas baseado no contexto refere-se às estratégias de roteamento que utilizam a informação contextual para determinar rotas que preencham requisitos específicos (por exemplo, requisitos de confiabilidade). Ter conhecimento do contexto significa dizer que uma entidade (por exemplo, o ponto), ao executar uma atividade, está levando em conta seu próprio contexto e o contexto das entidades que interagem com ela (por exemplo, pontos vizinhos, usuário) [Wenning 2010]. Por exemplo, é possível considerar o perfil do usuário (por exemplo, sua localização) e a disponibilidade do ponto no momento de decidir para qual ponto uma consulta deve ser encaminhada.

2.3.4. Perda de Informações no Processamento de Consultas em PDMS

Uma consequência da heterogeneidade das fontes de dados em um PDMS é a perda de informações (do inglês, *information loss*) durante o processamento da consulta [Roth 2012]. À medida que a consulta vai sendo encaminhada por entre os pontos do sistema, a cada reformulação, nem sempre os termos de um ponto origem encontram correspondentes exatos no ponto destino [Roth 2012].

Neste processo, as consultas reformuladas podem ser equivalentes ou apresentarem certo grau de perda semântica entre si. Delveroudis e Lekeas (2007) definem a perda semântica (do inglês, *semantic loss*) de uma consulta Q como sendo a diferença entre os termos existentes em Q e sua versão reformulada Q_{ref} ($Q - Q_{ref}$). Mena e seu grupo (2000) consideram duas medidas para análise da perda de informações: *intensional* (esquema) e *extensional* (dados). A medida *intensional* avalia as mudanças na consulta de acordo com a análise da perda semântica. A medida *extensional* objetiva estabelecer uma classificação dos planos de execução da consulta com relação aos resultados esperados. Para isto, histogramas podem ser utilizados para computar o potencial de contribuição dos mapeamentos [Ismail *et al*. 2013; Roth 2012; Gennaro *et al*. 2011].

2.4. Considerações

Neste capítulo, definimos os principais fundamentos que norteiam este trabalho. Apresentamos conceitos gerais sobre Ontologia, Contexto, PDMS e

Qualidade da Informação. Destacamos a importância da geração das correspondências semânticas e da reformulação da consulta no processo de roteamento. Identificamos o problema de análise da QI baseada em múltiplos critérios no processo de roteamento da consulta. Mostramos os métodos mais conhecidos e utilizados para resolução de problemas de MCDM em relação à seleção de fontes de dados. Por fim, apresentamos o problema da perda de informações em PDMS.

Em relação à perda de informações, serão discutidas com mais detalhes no Capítulo 4, as consequências desta perda e a solução proposta pelo *SemRouting* para preservação da semântica da consulta original ao longo do roteamento.

O próximo capítulo descreve as estratégias para roteamento de consultas mais relevantes ao tema deste trabalho.

Capítulo 3

Estratégias para o Roteamento de Consultas em PDMS

Em um PDMS, devido a não existência de um ponto de centralização de todo conhecimento distribuído e disponível no sistema, torna-se importante observar as estratégias adotadas para o roteamento eficiente das consultas. Cada ponto precisa decidir independentemente dos demais pontos que compõem a rede, para qual dos seus vizinhos a consulta deverá ser encaminhada. Esta decisão normalmente é feita tomando como base apenas o conhecimento local disponível no ponto.

Estratégias para o roteamento de consultas baseadas em semântica vêm sendo desenvolvidas com o objetivo de tornar mais eficiente este processo. Além do uso da semântica, em algumas propostas, aspectos relacionados à qualidade da informação têm sido incorporados às soluções de roteamento na intenção de reduzir o espaço de busca da consulta e melhorar a qualidade das respostas dadas aos usuários.

Este capítulo apresenta algumas estratégias consideradas mais relevantes ao tema deste trabalho. Ao final, estabelece uma análise comparativa entre as estratégias, de acordo com as questões de pesquisa apresentadas no Capítulo 1: identificação e seleção do melhor conjunto de pontos capazes de responder a consulta do usuário; preservação da semântica da consulta original ao longo do roteamento e mecanismo de parada do roteamento com o objetivo de evitar a execução de consultas com alta perda semântica.

3.1. Roteamento de Consultas em PDMS

De acordo com Ismail e seu grupo (2010) o roteamento semântico é um método para o encaminhamento de consultas muito mais focado na natureza da consulta do que na topologia da rede. O roteamento semântico objetiva melhorar as técnicas tradicionais de roteamento priorizando pontos que possam melhor responder à consulta. Entretanto, a identificação e priorização dos pontos não é uma tarefa trivial.

Para reduzir o tráfego de informações, lidar com os aspectos dinâmicos do ambiente e obter melhores resultados à consulta do usuário, alguns PDMS organizam seus pontos de acordo com a similaridade semântica entre eles, formando agrupamentos semânticos ou comunidades semânticas [Ismail *et al.* 2013; Montanelli *et al.* 2011; Pires *et al.* 2011]. Desta forma, o espaço inicial para o roteamento da consulta é reduzido, considerando apenas o conjunto de pontos que formam a comunidade. A seguir, apresentamos algumas estratégias de roteamento de consulta aplicadas a estes PDMS.

3.1.1. MCS-SP

O MCS-SP (*Minimal Covering Shortcut – Super-Peer*) é um sistema que permite a formação de comunidades semânticas em uma topologia de super-pontos (SP) onde cada SP pode pertencer a mais de uma comunidade [Ismail *et al.* 2013]. No MCS-SP, cada ponto pode armazenar dados em diferentes modelos (por exemplo, relacional, objeto ou XML). Para tratar a questão de heterogeneidade das fontes de dados é utilizado um modelo para representação comum dos dados denominado sGraph [Faye *et al.* 2007]. O sGraph corresponde a uma estrutura de grafo onde cada nó representa um elemento associado ao esquema da fonte, e as arestas representam ligações semânticas entre os elementos do grafo. Mapeamentos semânticos estabelecem ligações entre pontos e super-pontos. As ligações entre SP são estabelecidas por meio de índices semânticos.

A organização dos pontos em agrupamentos de super-pontos é feita de acordo com um histórico de sucesso em resposta aos termos das consultas submetidas aos SP. Cada agrupamento forma um hipergrafo, onde cada nó é um SP. Um SP pode pertencer a mais de uma comunidade definindo uma transversal entre os hipergrafos.

Neste cenário, o sistema define caminhos mínimos, ou seja, MCS (*Minimal Covering Shortcut*) entre todas as comunidades para o roteamento semântico de consultas.

A Figura 3.1 exemplifica os conjuntos de MCS em um hipergrafo de *SP* { {SP_1, SP_2, SP_6}, {SP_1, SP_6, SP_8} e {SP_3, SP_7, SP_8, SP_{10}}}. Cada agrupamento tem no mínimo um *SP* em comum usado para fazer o encaminhamento da consulta para outras comunidades. O usuário define preferências em relação à estratégia de roteamento, definindo, se assim desejar, o número de rotas transversais a serem utilizadas em suas consultas. Um componente da arquitetura do ponto denominado de *k-filter* é responsável pela aplicação deste filtro às rotas.

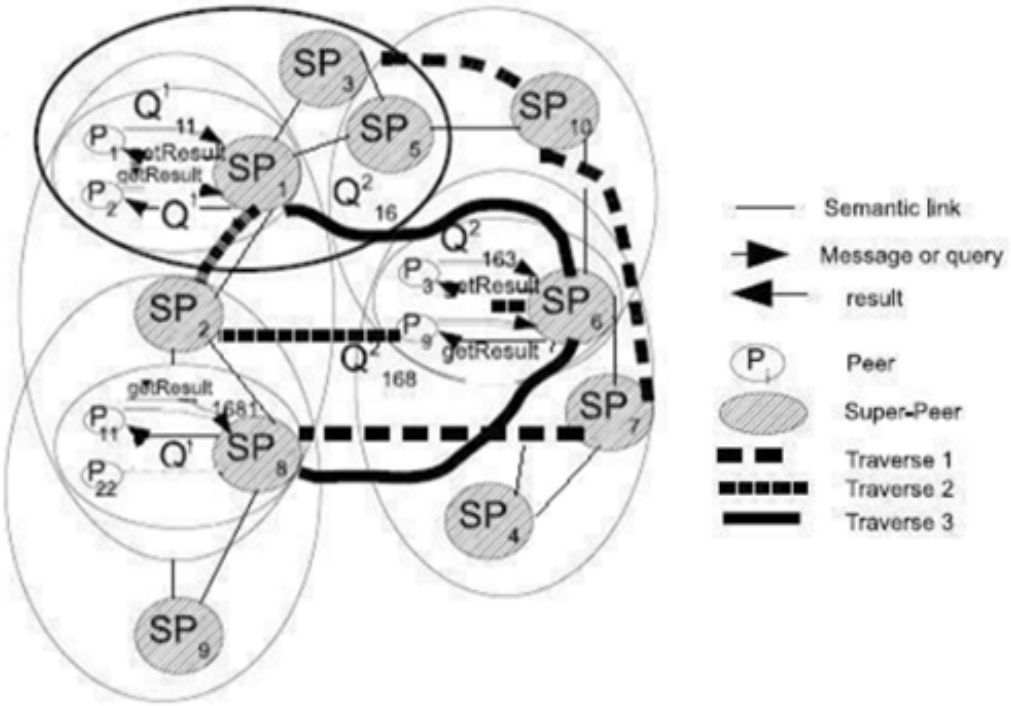


Figura 3.1 - - Roteamento da consulta [Ismail *et al.* 2011]

Na Figura 3.1 assumimos que uma consulta Q1 é submetida em P1. A partir do ponto de submissão, o roteamento da consulta se dará da seguinte forma:

1. O *SP* responsável por P_1 é identificado, no caso, SP_1.
2. SP_1 escolhe a rota transversal {SP_1, SP_2, SP_6} de acordo com preferências do usuário em relação a estratégia de roteamento.

3. Cada SP que recebe a consulta encaminha para seus pontos aptos a responder a consulta e também para aqueles SP existentes em outras rotas transversais
4. No final, o retorno da consulta irá corresponder ao conjunto de pontos relevantes e seus respectivos SP ($(P_2:SP_1),(P_{11}:SP_8)$).

É responsabilidade de cada SP identificar pontos de seu *cluster* que possam responder a consulta. Para isto uma função de cálculo de similaridade mede a capacidade do ponto ou SP em responder a consulta. O mecanismo de parada do roteamento é delimitado por um TTL e pelo tamanho e número de SP em cada MCS utilizado.

3.1.2. Ontozilla

No Ontozilla, pontos representam seus dados por meio de ontologias. Para facilitar o processo de busca, pontos que compõem a rede com mesmo domínio de conhecimento são agrupados em SIG (*Special Interest Groups*) [Joung e Chuang 2009]. Em cada SIG, os pontos empregam um sistema de classificação para agrupar seus interesses em classes, onde cada classe representa um *cluster* definido como uma árvore hierárquica (*cluster tree*) de pontos que abrigam o mesmo conceito. Nesta arvore, *links* estabelecem ligações semânticas entre pontos.

Os SIG são descrições conhecidas por todos os pontos na rede. Cada classe tem uma descrição contendo o nome da classe, o nome SIG, a hierarquia de classificação e algumas anotações da classe. As descrições de SIG e classes são representadas por ontologias. Assim, os relacionamentos entre SIG e/ou classes podem ser inferidos com uma ontologia pré-definida, ou usando alguma técnica de correspondência semântica onde se possa medir a similaridade entre os SIG ou classes de forma a determinar se dois pontos abrigam o mesmo assunto.

As ligações (*links*) semânticas estabelecidas entre os pontos facilitam o roteamento. Um ponto *x* que tem um *link* com um ponto *y*, na prática, equivale a dizer que *x* mantém a descrição de *y* em sua tabela de roteamento, fazendo com que *x* passe a ter uma referência direta com *y.* Os *links* no Ontozilla são classificados da seguinte forma:

- *SIG links* - são usados para conectar pontos de diferentes SIG.
- *Partner links* – estabelecem relacionamento cooperativo entre pontos. Por exemplo, se um ponto *x* frequentemente consulta informações do ponto *y*, é interessante que *x* mantenha um *partner link* com *y*.
- *Twin links* – são usados para conectar pontos em um mesmo *cluster*.
- *Parent links* e *child links* – estabelecem a conexão entre classe e sub-classe em um *cluster-tree*

Para ilustrar, a Figura 3.2 mostra um representação do Ontozilla. Cada agrupamento representado na figura por linhas pontilhadas que cercam *clusters* representa um SIG. Os *links* semânticos entre os pontos (dentro de um *cluster*) estabelecem caminhos para que as consultas possam ser roteadas de forma eficiente aos pontos que contenham as informações solicitadas na consulta. O encaminhamento de consultas é limitado por TTL.

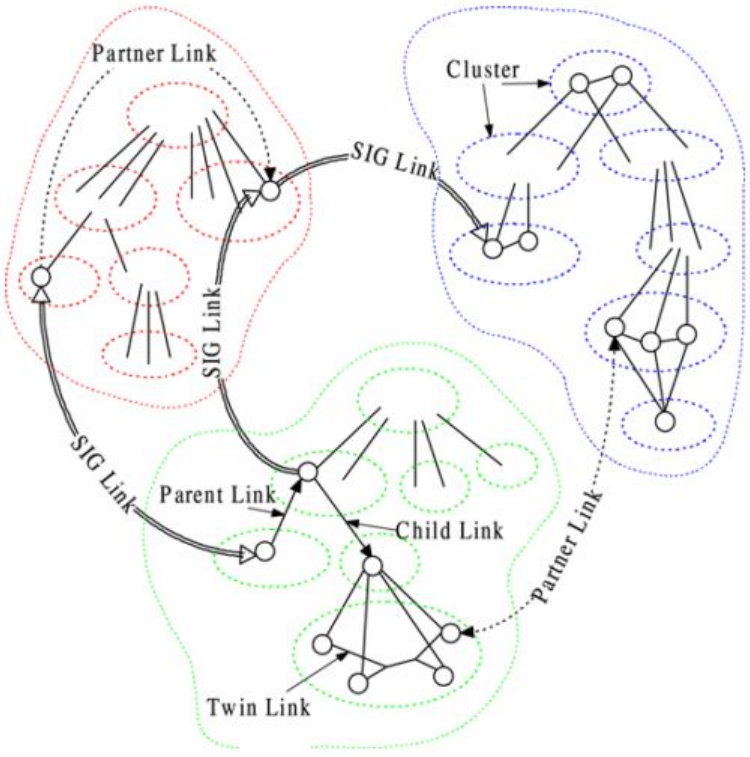


Figura 3.2- Representação do Ontozilla [Joung e Chuang 2009]

O roteamento no Ontozilla é feito conforme a estratégia de roteamento intra-SIG e inter-SIG. O **roteamento intra-SIG** é feito de forma direta. Se um ponto recebe uma consulta que solicita um dado pertencente a sua super-classe, ele encaminha a

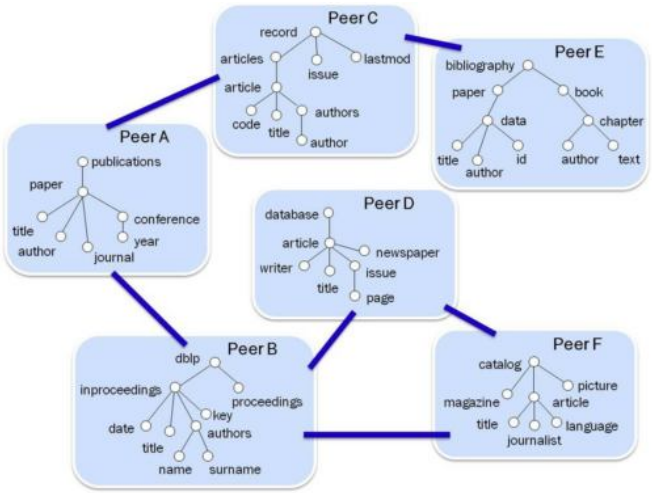
consulta para o *cluster* mais superior (*parent links*). Se a consulta é para dados pertencentes a sua subclasse, ele encaminha a consulta para o *cluster* mais inferior (*child links*). Finalmente, se a consulta é para os dados pertencentes a sua própria classe, ele transmite a consulta ao seu *cluster* através de ligações individuais (*twin links*). Se mesmo assim, o ponto que fez a consulta não estiver satisfeito com o número de respostas, ele pode fazer uma nova consulta agora envolvendo as subclasses ou super-classes das classes que já foram consultadas, fazendo com que o processo se repita. O **roteamento inter-SIG** está relacionado ao pedido de dado fora do SIG do próprio ponto. Quando um ponto necessita de um dado ele procura em sua tabela de roteamento e encaminha a consulta para o ponto correspondente por meio de ligações SIG. Se a tabela de roteamento não contém qualquer ponto correspondente, ele transmite *SIG inquire message*, que são mensagens utilizadas no processo de junção de SIG, dentro de um número de saltos estimados, para que seja feita uma busca do novo SIG. Caso o SIG seja encontrado, a consulta é enviada para ele e, então, roteada conforme o processo intra-SIG.

3.1.3. SUNRISE

O SUNRISE (*System for Unified Network Routing, Indexing and Semantic Exploration*) é um PDMS que integra fontes de dados heterogêneas utilizando uma rede semântica baseada no agrupamento de pontos de dados com mesmo domínio de interesse [Mandreoli *et al.* 2010]. A arquitetura do SUNRISE foi desenvolvida para trabalhar, independente do modelo de dados, com qualquer forma de representação dos esquemas e formulação de consultas. Entretanto, as particularidades do modelo de dados utilizado precisam ser consideradas no desenvolvimento dos módulos do sistema. A rede é organizada em SON (*Semantic Overlay Network*) de forma que cada ponto tem na sua vizinhança pontos semanticamente relacionados.

Cada ponto que compõe o PDMS utiliza o SRI (*Semantic Routing Index*), uma estrutura de dados local utilizada para manter um resumo das informações a respeito do grau de similaridade semântica entre os conceitos armazenados entre um ponto e seus respectivos vizinhos. Assim, um ponto **p** que tenha **n** vizinhos e **m** conceitos em seu esquema, armazena um SRI estruturado como uma matriz com **m** colunas e **n+1** linhas, onde a primeira linha refere-se ao conhecimento sobre o esquema local do ponto **p**, como mostra a letra (b) da Figura 3.3 que utilizou como cenário o PDMS

representado na letra (a) da mesma figura. É possível assim, que cada ponto sintetize, para cada conceito de seu esquema, a aproximação semântica das sub-redes acessíveis a partir de seus vizinhos e, assim, forneça uma informação sobre a relevância dos dados que podem ser alcançados em cada trajeto a ser escolhido.



(a) Cenário de demonstração do SUNRISE

PeerA SRI	paper	title	author	...
PeerB	0.51	0.49	0.37	...
PeerC	0.81	0.86	0.66	...

(b) Parte do Índice de Roteamento Semântico (SRI) do ponto A

Figura 3.3- Exemplo do uso do SRI [Mandreolli *et al.* 2007]

O sistema permite ao usuário explorar os caminhos mais promissores durante uma busca. O usuário estabelece preferências indicando o cenário inicial de consulta: o ponto e o conceito, a condição de parada e a estratégia de roteamento. São condições de parada: (a) máximo de saltos (TTL) e (b) meta de satisfação (uma medida de qualidade do caminho a ser explorado). Caso o valor da meta de satisfação esteja abaixo de um dado *threshold,* a consulta não é encaminhada ao ponto seguinte. Infelizmente, o trabalho não descreve a forma de cálculo da meta de satisfação.

A estratégia de roteamento SRI também vem sendo utilizada em conjunto com a estratégia MRoute [Gennaro *et al.* 2007] com o objetivo de prover o compartilhamento de dados estruturados, semi-estruturados e multimídia em um

PDMS [Gennaro *et al.* 2011] desenvolvido como parte do projeto NeP4B[4] (*Networked Peers for Business*).

3.1.4. H-Link

O H-link é um mecanismo de roteamento semântico para sistemas P2P cujas principais características são: o uso de ontologias para representação do conhecimento dos pontos, o uso de técnicas de correspondência de ontologias para seleção de pontos semanticamente similares e o gerenciamento independente de sua própria ontologia pelos pontos [Montanelli e Castano, 2008]. A idéia chave do H-Link é explorar os resultados das interações durante a fase de descoberta de conhecimento para treinar o comportamento do mecanismo de roteamento. Para alcançar esse objetivo, os pontos são conectados por meio de medidas de confidência baseadas em técnicas de correspondência que acompanham a afinidade semântica entre os conteúdos dos diferentes pontos. Logo, os pontos são organizados em uma SON (*Semantic Overlay Network*) cujos pontos têm conhecimento similar e estão interligados como vizinhos semânticos.

A ontologia do ponto no H-Link provê uma descrição formal do contexto do ponto em termos de: (1) conhecimento local dos dados do ponto que serão compartilhados com os outros pontos, e (2) conhecimento do ponto sobre a rede que é o conhecimento do ponto sobre seus vizinhos semânticos progressivamente descobertos ao longo das interações que ocorrem na rede. É importante destacar que a definição de contexto nesse trabalho está relacionada à representação formal do conhecimento do ponto (por exemplo, uma ontologia) sobre os dados compartilhados.

Para ilustrar esse mecanismo, o autor propõe um exemplo (Figura 3.4), onde cada ponto é independente e se uniu ao sistema por meio de sua própria ontologia. Suponha que o ponto A está interessado em localizar outros pontos que possuam dados semanticamente relacionados ao domínio de **publicação**. Para isso, o ponto A formula uma consulta Q_1 e submete ao sistema contendo os conceitos de interesse *Publication* e *Book* com suas respectivas propriedades *year* e *author*. Ao receber a consulta Q_1 os pontos B, C e D usam o gerador de correspondência semântica (H-Match) [Castano *et al.* 2006] para comparar a descrição dos conceitos na consulta com

[4] Project NeP4B (Networked Peers for Business). http://dbgroup.unimo.it/nep4b.

a ontologia do ponto. Conforme o resultado desse processo, os pontos B e D enviam ao ponto A uma lista ordenada dos conceitos encontrados, e, para cada entrada, o cálculo do valor de afinidade semântica (SA). Como se pode observar na Figura 3.4, apenas os pontos B e D retornam valores enquanto C não responde ao ponto A por não possuir conceitos correspondentes.

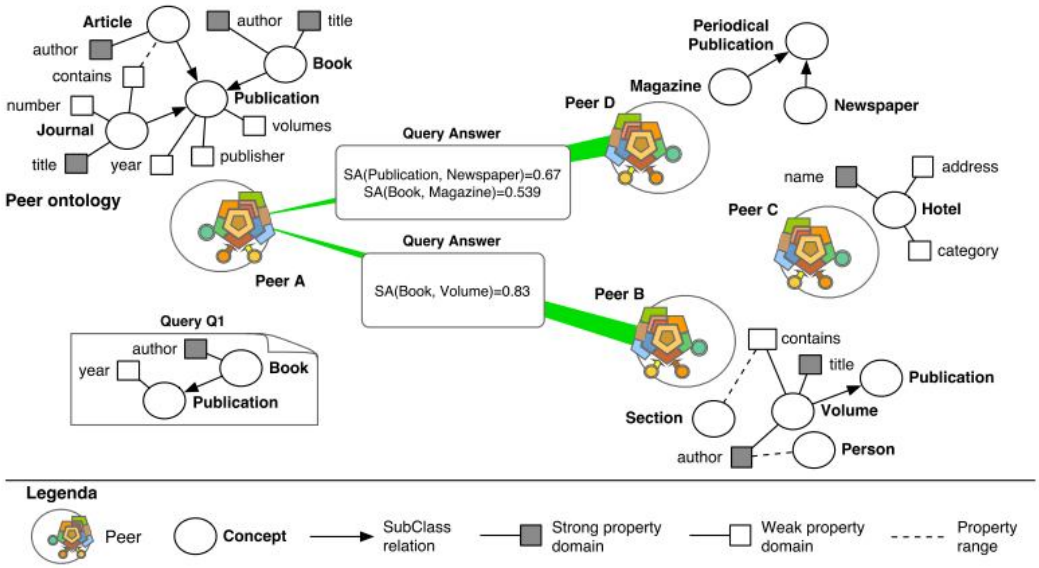


Figura 3.4 - Mecanismo de roteamento H-Link [Castano e Montanelli 2008].

É importante destacar que a resposta fornecida ao ponto A poderá ser utilizada em interações futuras. Quando consultas similares forem feitas, esses resultados contribuirão no processo de identificação dos melhores pontos disponíveis na rede, aptos a responder a consulta [Castano e Montanelli 2008].

Para interromper o roteamento, o H-Link utiliza um sistema de distribuição de crédito que corresponde ao número de respostas esperadas em relação à consulta. O valor do crédito é um parâmetro estimado a partir de interações passadas. Nesse sistema, a distribuição de crédito é feita proporcionalmente de acordo com a classificação de similaridade semântica entre a consulta e os vizinhos selecionados. O número de créditos é decrementado sempre que a consulta é encaminhada para um novo ponto. Ao final, se não existir crédito disponível o mecanismo de propagação da consulta é interrompido e as respostas são retornadas seguindo o caminho inverso da consulta.

3.1.5. ESTEEM

O ESTEEM (*Emergent Semantics and cooperaTion in multi-knowledgE EnvironMents*) tem como objetivo oferecer uma plataforma baseada em comunidade semântica para integração de dados e serviços em uma rede P2P não estruturada [Montanelli *et al.* 2011]. Cada ponto nesse sistema é caracterizado por uma **ontologia do ponto** (descrição dos dados a serem compartilhados), uma **ontologia de serviço** (descrição dos serviços a serem compartilhados), um **contexto corrente** (descreve o perfil do ponto, seu interesse, situação e coordenadas espacial/temporal), **perfil de confiança** (determinado com base no número de queixas disparadas por outros pontos da comunidade, para o qual o ponto tenha sido um provedor de um determinado dado sobre uma transação específica) e de **qualidade dos dados** (computação de métricas de qualidade sobre os dados exportados pelo ponto). Em se tratando dos aspectos de qualidade, cada ponto tem a possibilidade de associar metadados de qualidade para o dado exportado. São critérios utilizados para os dados exportados pelo ponto: *column completeness*, *format consistency*, *accuracy* e *internal consistency* [Batini e Scannapieco 2006].

No ESTEEM, um modelo de contexto geral, a *Context Dimension Tree* (CDT) (Figura 3.5) representa todos os possíveis contextos passíveis de ocorrer em determinada situação. Baseado nesse modelo, o ponto que inicia a formação de uma comunidade semântica poderá associar a CDT a um contexto desejável para a comunidade. Assim, a CDT de um ponto expressa algumas perspectivas (dimensões) que determinam qual porção do dado é interessante em diferentes situações. Por exemplo, *actor*, *situatio*n e *interest topic* são algumas das dimensões mais comuns que podem guiar a seleção de informação ou serviços relevantes à consulta [Montanelli *et al.* 2011].

Uma sub-árvore da CDT determina uma porção do conjunto de dados, especificado como uma visão. Essa visão representa os dados que são relevantes quando o contexto correspondente torna-se corrente. Na Figura 3.5, pode-se observar um exemplo da modelagem CDT para uma aplicação na área médica, onde um médico de um hospital da África Central está interessado em saber sobre estruturas disponíveis para doenças contagiosas em sua região. O objetivo desse médico é encontrar dados e serviços relevantes para o tratamento de pacientes com malária. A

área cinza na figura representa o contexto corrente do ponto de submissão da consulta para formação da comunidade, onde os valores em tempo de execução para os parâmetros **id_name** e **reg_name** são “Malária” e “África Central”.

O ponto de origem da consulta submete seu contexto corrente para seus vizinhos semânticos que por sua vez utilizam um matching de contexto para calcular a similaridade entre o contexto recebido e o local. Os pontos com contexto similar respondem ao ponto que submeteu o CDT inicial. Como consequência, vizinhanças semânticas são estabelecidas por ligações entre os pontos similares. Em geral, as solicitações orientadas a contexto podem antecipar solicitações de dados ou serviços e ajudar os pontos na especificação de consultas mais refinadas. Cada ponto armazena o conjunto de comunidades semânticas em uma lista denominada JSC (*Joined Semantic Communities*) composta pelo identificador da comunidade e seu respectivo manifesto.

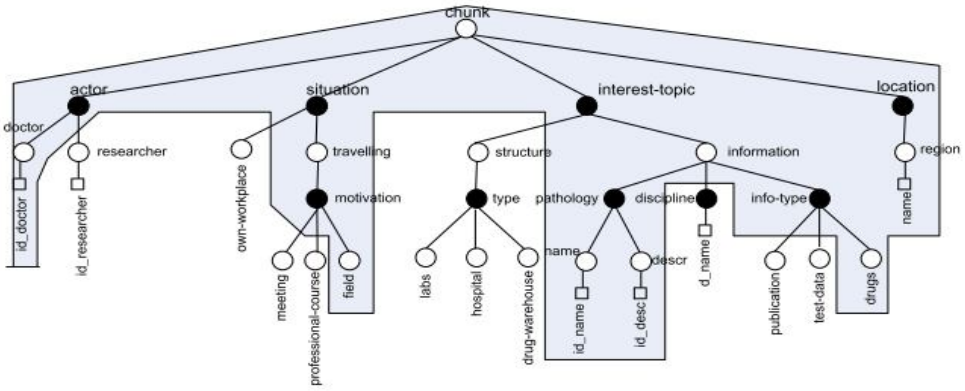


Figura 3.5 - Exemplo de CDT [Aiello *et al.* 2007]

Para responder uma consulta, o módulo de roteamento semântico aplica o mecanismo de roteamento por comunidade [Castano e Montanelli, 2007]. O objetivo deste mecanismo é identificar a partir da estrutura JSC quais comunidades deverão receber a consulta. Em seguida a consulta é propagada nesta comunidade. Um mecanismo de TTL é implementado para evitar ciclos e propagação ilimitada na rede durante a propagação da consulta [Bianchini *et al.* 2009].

3.1.6. Humboldt Peer

A estratégia adotada no PDMS *Humboldt Peer* está voltada para a completude dos resultados da consulta a partir da redução de custo dos planos de consulta [Roth 2012]. O *Humboldt Peer* é um PDMS composto por um conjunto de pontos, onde cada

ponto armazena dados em uma base relacional, associados por meio de mapeamentos. O modelo baseado em custo e benefício tem como objetivo encaminhar a consulta apenas para aqueles pontos que possam fornecer resultados. Para obter essa decisão, cada ponto classifica todos os seus vizinhos de acordo com o potencial de resultados a serem retornados baseados em seus respectivos mapeamentos [Roth e Naumann 2007].

O modelo envolve duas dimensões: *cobertura* e *densidade*. A *cobertura* descreve a proporção do tamanho de um conjunto de tuplas em função do total do número de tuplas armazenadas no PDMS. A medida aplica-se tanto para os dados armazenados como para o conjunto de tuplas retornadas como resultado a uma consulta. No caso das tuplas retornadas, a medida é baseada no número total de tuplas que preenchem os requisitos da consulta. A dimensão *densidade*, por outro lado, descreve o número de valores de atributo para cada resultado em relação aos atributos da consulta. É esperado que alguns atributos retornem com valores nulos criando assim tuplas com resultados incompletos de baixa densidade.

O algoritmo de cálculo da completude revela o impacto da perda de informação de um mapeamento sobre os resultados da consulta em um ponto. Baseado neste cálculo, planos de consulta com estratégias de estimativa de custo (*budget*) são definidas objetivando estabelecer um limite de roteamento (a exemplo do mecanismo de TTL) para mapeamentos alternativos entre pontos.

Ao receber uma consulta, o ponto deve classificar diferentes planos de consulta local de acordo com o potencial de dados a serem retornados e mediante a estimativa recebida. Caso os planos não atendam à consulta eles deverão ser retirados da lista de classificação gerada. A estimativa do potencial de dados a serem retornados é obtida por meio de histogramas multi-dimensionais [Roth e Naumann 2005].

Existem dois tipos de estratégias de estimativa definidas - *Weight* e *Greedy,* descritas a seguir:

- *Weight* – nessa estratégia, o ponto considera o peso de contribuição dos mapeamentos de acordo com a sua vizinhança. Nesse caso, o ponto distribui uma estimativa na proporção inversa à contribuição da informação perdida em determinado mapeamento. Por exemplo, se o uso de

determinado mapeamento provoca uma grande perda de informação, a estimativa será menor em comparação a outros mapeamentos que produzam resultados melhores. Essa estratégia prefere explorar a vizinhança direta de um ponto (com alta relevância semântica e menor perda de informação) em vez de tentar alcançar vizinhos indiretos, ou seja, mais distantes.

- *Greedy* – para maximizar a exploração de mapeamentos essa estratégia destina toda a estimativa para o mapeamento que apresenta melhores resultados, ou seja, menor perda de informação. Essa estratégia promete um retorno de mais dados, para estimativas pequenas, comparado ao tamanho do espaço de busca.

É importante observar que ambas as estratégias são fortemente dependentes da razão entre o valor total da estimativa de custo, o tamanho do espaço de busca e a média do número de mapeamentos existentes no PDMS.

3.1.7. GrouPeer

O GrouPeer é um sistema desenvolvido com o objetivo de permitir a avaliação precisa de consultas por meio de um processo automático de criação, manutenção e agrupamento de grupos semânticos similares em uma rede P2P não estruturada [Kantere *et al.* 2011].

O método de descoberta de pontos remotos e relevantes consiste em propagar, ao longo do caminho percorrido por uma consulta, tanto a consulta original como também sua versão reescrita. Nesse caso, os pontos receberão as consultas e decidirão qual a versão que irão responder de acordo com o resultado obtido a partir de uma função de similaridade. Reformulações sucessivas da consulta podem produzir versões com diferenças semânticas da consulta original. Se a cadeia de mapeamentos utilizados na reescrita é pobre em informação relevante para a consulta (ou seja, partes da consulta não podem ser reformuladas com precisão), isto pode resultar em uma rápida degradação da consulta em poucos saltos do roteamento.

O GrouPeer mantém uma lista dos atributos eliminados durante as reformulações para que eles possam ser utilizados com os esquemas dos próximos pontos para os quais a consulta tenha sido encaminhada. No sistema, pontos decidem

individualmente se respondem a consulta que tenha sido reescrita ou a sua versão original. O mecanismo de parada do roteamento é definido a partir de um valor de TTL previamente estabelecido. Durante o roteamento, a identificação de vizinhos relevantes é feita a partir da avaliação de similaridade entre o esquema do ponto e dos atributos da consulta. Isto implica dizer que um ponto envia uma consulta apenas para os vizinhos cujos esquemas têm o maior valor de similaridade em relação à consulta.

A Figura 3.6 mostra um processo de propagação de uma consulta no GrouPeer. Com base na figura, observa-se que o ponto P2 recebe a consulta original Qorig e a consulta reescrita Qsr_M1,2 , resultante do mapeamento M1,2, entre P1 e P2, e assim sucessivamente entre os demais pontos.

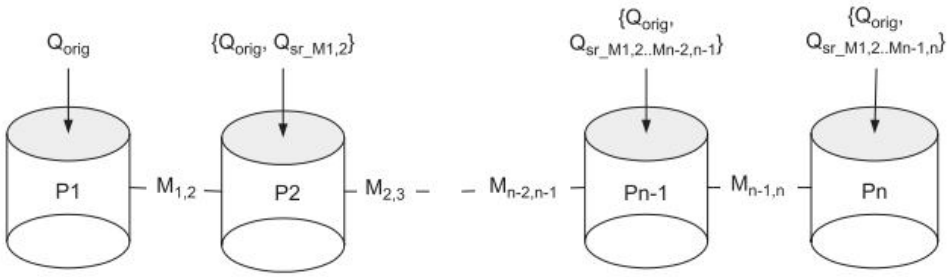


Figura 3.6- Propagação de uma consulta no GrouPeer [Kantere *et al.* 2009]

No ponto destino, as duas consultas, a original (Qorig) e a previamente reescrita (Qsr_previous), serão novamente reescritas de acordo com os mapeamentos para os vizinhos desse ponto. A Qorig será rescrita para Qar e a Qsr_previous para Qsra (Figura 3.7).

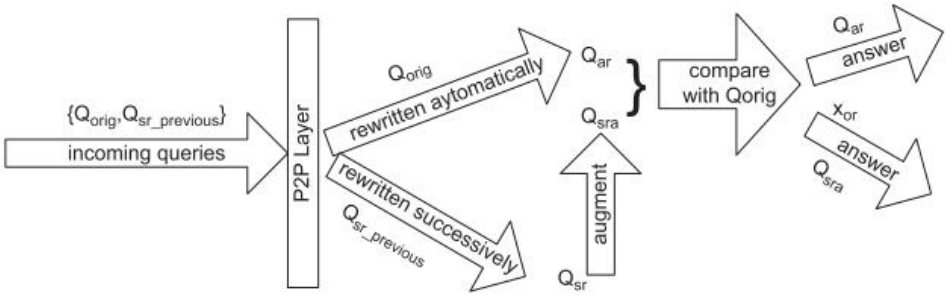


Figura 3.7- Procedimento de resposta da consulta no ponto [Kantere *et al.* 2009]

Após reescritas, as consultas Qar e Qsra são comparadas com Qorig, por meio de uma função de similaridade [Kantere *et al.* 2011]. Depois desta comparação, a

consulta que tiver o maior grau de similaridade com a consulta original, será a consulta a ser respondida pelo ponto.

Ao receber a resposta da consulta, o ponto que submeteu a consulta original indica seu grau de satisfação em função da resposta recebida. Esta avaliação em torno da versão utilizada para a resposta será enviada ao ponto que a respondeu. Dependendo da avaliação, o ponto que submeteu a consulta original pode decidir se ele tem interesse comum com o ponto remoto e convidá-lo a estabelecer novos mapeamentos entre eles.

3.2. Análise dos Trabalhos Relacionados

Em uma análise dos trabalhos apresentados, foi observado que a formação de agrupamentos semânticos em um PDMS vem sendo utilizada como alternativa para reduzir o espaço de busca durante o roteamento de consultas e estabelecer uma vizinhança de pontos com melhor aproximação semântica. Outras características importantes são o uso de informações contextuais, a exemplo das preferências de usuários [Mandreolli *et al.* 2010; Castano e Montanelli, 2008; Montanelli *et al.* 2011] e de critérios de qualidade [Roth 2012] como forma de tornar mais específica a seleção de pontos na rede.

A identificação e seleção de pontos também envolve o uso de índices semânticos [Ismail *et al.* 2011; Joung e Chuang 2009; Mandreolli *et al.* 2010], funções de similaridade semântica de acordo com os mapeamentos para análise dos pontos no nível de esquema [Kantere *et al.* 2011; Mandreolli *et al.* 2010; Castano e Montanelli, 2008; Montanelli *et al.* 2011] e histogramas para análise dos pontos no nível de dados [Roth 2012]. Embora permita o acesso direto aos pontos, o uso de índices necessita de um processo de atualização periódica para garantir o funcionamento do roteamento da consulta em função da dinamicidade do ambiente (entrada e saída de pontos, evolução de esquemas, atualizações de dados e mudanças de vizinhança). Isto significa que qualquer mudança gerada em um ponto deverá ser atualizada e propagada para toda vizinhança deste ponto.

No trabalho de Joung e Chuang (2009) a eficiência e corretude do roteamento da consulta estão baseadas na confiança das ligações semânticas entre os pontos. Pontos registrados na tabela de roteamento podem estar indisponíveis se

considerarmos a dinamicidade do ambiente onde os pontos podem entrar ou sair da rede a qualquer momento. Embora seja positivo fazer uma estimativa do conjunto de resultados para avaliar o nível de contribuição do ponto em relação à consulta, a geração e manutenção de histogramas tem um custo elevado em relação ao cálculo computacional e de atualização dos dados já apontado por outros autores [Doan *et al*. 2012; Akbarinia *et al.* 2007].

Quanto ao mecanismo de parada, Joung e Chuang (2009) e Montanelli *et al.* (2011) utilizam um mecanismo de TTL para limitar o número de encaminhamentos da consulta na rede. O trabalho do Humboldt Peer [Roth 2012] utiliza um plano de estimativa de custo (*budget*) associado ao cálculo da completude das respostas. Na estratégia do H-Link utilizada [Castano e Montanelli, 2008], um valor baseado em crédito é estabelecido em função do número de respostas esperadas pelo ponto. Mecanismos limitados por TTL tendem a limitar o encaminhamento da consulta para possíveis pontos promissores. No caso do Humboldt Peer, a estratégia de estimativa de custo *Weight* apresenta como problema o fato de que o custo estabelecido inicialmente pode levar os pontos a gastar tudo que dispõe em mapeamentos que possuem alta perda de informação. O mesmo pode acontecer com a estratégia *Greedy* que investe toda a sua estimativa de custo em um único mapeamento esperando obter melhores resultados. A estratégia de SRI utilizada nos trabalhos de Mandreolli *et al.* (2010) e Gennaro *et al.* (2011) apresenta como critério de parada o uso de TTL e de uma meta de satisfação que corresponde à medida de qualidade da rota a ser explorada. Embora outros trabalhos relacionados à estratégia SRI tenham sido avaliados, não encontramos referência à forma de cálculo desta meta de satisfação.

Apesar das estratégias analisadas oferecerem soluções que minimizam a perda de informações em um PDMS a partir da eliminação de pontos que não possam contribuir com a consulta, pouco se trata, com exceção do Groupeer [Kantere *et al.* 2011], do problema da preservação semântica da consulta original ao longo do roteamento. A estratégia do Groupeer objetiva evitar que as respostas dadas aos usuários sejam prejudicadas em função da perda de informações durante o sucessivo processo de reescrita da consulta. Assim, cada ponto selecionado, ao responder uma consulta, escolhe entre reescrever a consulta original ou utilizar a sua versão reescrita

encaminhada pelo ponto anterior. Neste processo, termos eliminados em pontos anteriores podem ser utilizados durante o processo de reescrita da consulta original.

3.3. Quadro Comparativo

Nesse capítulo as principais abordagens relacionadas ao tema deste trabalho foram apresentadas. O Quadro 3.1 apresenta um comparativo das principais características observadas em cada sistema, destacando as questões de pesquisa apresentadas no Capítulo 1: identificação e seleção do melhor conjunto de pontos capazes de responder a consulta do usuário; preservação da semântica da consulta original ao longo do roteamento e mecanismo de parada do roteamento (evitando o enacaminhamento de consultas com alta perda semântica). O significado de cada característica é apresentado a seguir:

- Estratégia – identifica o nome da estratégia de roteamento ou nome do PDMS;
- Representação dos esquemas – formato de representação dos esquemas das fontes de dados pertencentes a cada ponto no PDMS
- Identificação de Pontos – indica qual tipo de estrutura (índices ou mapeamentos) é utilizada para acessa os pontos
- Seleção de Pontos – tipo de informação utilizada para selecionar pontos para o encaminhamento da consulta. Campos preenchidos com o valor "Não" indicam que a estratégia não usa o tipo de informação especificada na coluna.
    - Informação Contextual – indica de onde é proveniente a informação utilizada
    - Critério de Qualidade – nome do critério de qualidade
- Preservação da Consulta - indica como a estratégia trata a questão da preservação da consulta original ao longo do roteamento. Campos preenchidos com o valor "Não" indicam que a estratégia não trata a preservação.
- Mecanismo de Parada – indica o tipo de informação utilizada para interromper o roteamento da consulta.

- Outras Características - informações adicionais e relevantes ao comparativo entre os trabalhos.

3.4. Considerações

Este capítulo apresentou estratégias para o roteamento semântico de consultas em PDMS. Em seguida, foi realizada uma análise comparativa utilizando como referência as questões de pesquisa que norteiam este trabalho. De acordo com esta análise, concluímos que as estratégias consideram o roteamento por comunidades semânticas com o objetivo de melhorar o alcance das consultas no PDMS e fazem a seleção de pontos de duas formas: considerando a similaridade do ponto em relação à consulta, ou o nível de contribuição esperada em termos de resultados a serem obtidos pela consulta.

Apesar de ser uma necessidade em ambientes dinâmicos com muitas fontes, nenhuma das abordagens considera o uso de contexto e da análise de múltiplos critérios de qualidade, de forma combinada, para classificação e eliminação das fontes de dados com o objetivo de obter o melhor conjunto de resultados. No caso do ESTEEM, a CDT (*Context Dimension Tree*) é utilizada na formação dos agrupamentos em torno de um contexto comum e não como critério para o roteamento. A preservação da consulta é um aspecto não considerado dentre os trabalhos relacionados. Outro ponto a ser observado está relacionado ao mecanismo de parada. Nos trabalhos apresentados não existem critérios semânticos definidos para evitar o encaminhamento de consultas que não estejam preservadas. As soluções apresentadas estão baseadas em número de saltos (TTL) ou estimativa de créditos.

O próximo capítulo apresenta a proposta deste trabalho que une aspectos semânticos e de qualidade na definição de uma abordagem para o roteamento de consultas em PDMS.

Quadro 3.1 - Quadro Comparativo das Estratégias de Roteamento

Estratégia	Representação dos Esquemas	Identificação de Pontos	Seleção de Pontos		Preservação da Consulta	Mecanismo de Parada	Outras Características
			Informação Contextual	Critério de Qualidade			
MCS-SP	sGraph	Mapeamentos e índices semânticos	Preferência do usuário	Não	Não	TTL	Baseado em hypergrafo; threshold de vizinhaça
Ontozilla	Ontologias	Mapeamentos e índices semânticos	Não	Não	Não	TTL	SIG (*Special Interest Groups*)
SUNRISE	Ontologia, relacional, XML	Mapeamentos e índices semânticos	Preferência do usuário	Não	Não	Meta de satisfação; TTL	SRI + MRoute [Gennaro *et al.* 2007] em PDMS dados multimídia [Gennaro *et al.* 2011]
H-Link	Ontologia	Mapeamentos semânticos;	Contexto do ponto	Não	Não	Baseado em crédito	O contexto está relacionado ao dado e vizinhança do ponto
ESTEEM	Ontologia	Mapeamentos semânticos	CDT (*Context Dimension Tree*)	Não	Não	TTL	Metadados de qualidade em resultados; Qualidade +CDT na formação das comunidades
Humboldt Peer	Relacional	Mapeamentos semânticos	Não	Cobertura e Densidade	Não	Baseado em estimativa de custo	Histogramas multi-dimensionais para estimativa de custo
Groupeer	Relacional	Mapeamentos semânticos	Não	Degradação da consulta	Manutenção de conceitos perdidos	TTL	Opção entre reescrita da consulta reformulada e consulta original

Capítulo 4

A Abordagem SemRouting

Um dos principais serviços que um PDMS pode oferecer é o de permitir a realização de consultas em seus diversos pontos. Entretanto, consultar um grande número de pontos nem sempre garante a satisfação do usuário em relação à resposta a suas consultas. Pontos com informações não relacionadas à consulta, aumento do tempo de resposta, geração de sobrecarga desnecessária na rede e elevado custo para integração dos resultados, muitas vezes irrelevantes, são situações decorrentes deste processo que podem impactar esse nível de satisfação [Doan *et al*. 2012]. Além disso, devido à heterogeneidade das fontes, a consulta original ao ser reformulada e roteada pode rapidamente sofrer perda semântica [Delvedouris e Lekeas, 2007] em relação a sua formulação inicial.

Para tratar com estes problemas, a abordagem *SemRouting*, descrita neste capítulo, combina informações semânticas e de qualidade da informação (QI) com o objetivo de preservar a semântica da consulta original, ao longo do roteamento, e obter o melhor conjunto de pontos capazes de respondê-la em um PDMS.

Este capítulo está organizado da seguinte forma: a Seção 4.1 introduz o ambiente para o roteamento considerado na definição da abordagem; a Seção 4.2 define o problema da perda semântica da consulta no roteamento; a Seção 4.3 mostra

informações semânticas e de qualidade para o roteamento de consultas; a Seção 4.4 propõe a abordagem *SemRouting;* a Seção 4.5 mostra um exemplo completo de aplicação da *SemRouting*; a Seção 4.6 faz um comparativo entre as abordagens apresentadas no Capítulo 3 e a *SemRouting*. Por fim, a Seção 4.7 apresenta as considerações do capítulo.

4.1. Visão Geral do Ambiente

A abordagem *SemRouting* pressupõe um PDMS com pontos semanticamente relacionados dois-a-dois por meio de correspondências (associações semânticas entre esquemas de pontos). Ontologias são usadas para representar os esquemas dos dados exportados por cada ponto da rede. Neste ambiente, pontos conectados diretamente por meio de correspondências são denominados "vizinhos semânticos" [Gennaro *et al.* 2011; Pires *et al*. 2009; Montanelli e Castano, 2008].

Em um PDMS, consultas submetidas em um ponto são respondidas com os dados armazenados no próprio ponto e/ou por dados que podem ser obtidos em outros pontos. Devido à heterogeneidade das fontes, a consulta, ao ser encaminhada de um ponto a outro do sistema, precisa ser reformulada, ou seja, reescrita de acordo com o esquema do ponto destino.

Dado este ambiente, na definição da *SemRouting*, dois aspectos são considerados importantes e estão fortemente relacionados: as **correspondências semânticas** existentes entre os esquemas dos pontos vizinhos, e a estratégia de **reformulação da consulta** utilizada pelo sistema. Para a geração das correspondências semânticas, a *SemRouting* adota a estratégia definida no *SemMatcher* [Pires *et al.* 2009] e, para reformulação da consulta, a *SemRef* [Souza *et al.* 2009]. A seguir passamos a descrever cada uma destas estratégias.

4.1.1. SemMatcher

O *SemMatcher* [Pires *et al*. 2009] é um processo de associação semântica que estabelece correspondências entre ontologias de acordo com a similaridade semântica entre elas. Para tal, uma ontologia de domínio é utilizada como "*background knowledge*".

Ontologias de domínio (OD) contêm elementos (conceitos e propriedades) de um determinado domínio de conhecimento (por exemplo, Educação ou Saúde). O *SemMatcher* usa esta OD para identificar a semântica do relacionamento entre elementos de duas ontologias. Inicialmente, os elementos, de cada uma das ontologias, são mapeados para elementos equivalentes na OD. Em seguida, as correspondências semânticas são inferidas baseadas no relacionamento semântico existente entre os elementos na OD. Por exemplo, se dois conceitos x e y, de ontologias distintas, possuem conceitos equivalentes k e z na OD, sendo k e z conceitos semanticamente relacionados, é possível inferir que x e y possuem uma relação semântica. Assim, o *SemMatcher* é capaz de identificar qual tipo de relacionamento semântico existe entre k e z e, consequentemente, especificar o tipo de correspondência que pode ser estabelecida entre x e y [Souza 2009].

A Figura 4.1 fornece uma visão geral deste processo para especificação da semântica das correspondências entre duas ontologias O_1 e O_2. Neste cenário, x, um conceito de O_1, e y, um conceito de O_2, possuem conceitos equivalentes k e z na OD (O_1:x ≡ Do:k e O_2:y ≡ Do:z). Considerando que k é um subconceito de z na OD, é possível inferir que a mesma relação ocorre entre x e y. Assim, pode-se afirmar que O_1:x é um subconceito de O_2:y (O_i:x ⊑ O_i:y)

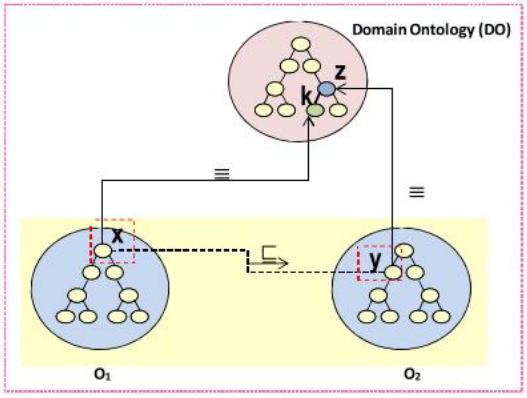


Figura 4.1- Especificando correspondências entre ontologias usando uma OD [Souza 2009]

Considerando estes aspectos, uma correspondência semântica é definida como uma das seguintes expressões [Souza 2009]:

1. O_i:x ≡ O_i:y, é uma correspondência *isEquivalentTo*
2. O_i:x ⊑ O_i:y, é uma correspondência *isSubConceptOf*
3. O_i:x ⊒ O_i:y, é uma correspondência *isSuperConceptOf*

4. O_i:x $\triangleright$ O_i:y, é uma correspondência *isPartOf*
5. O_i:x $\triangleleft$ O_i:y, é uma correspondência *isWholeOf*
6. O_i:x $\approx$ O_i:y, é uma correspondência *isCloseOf*
7. O_i:x $\perp$ O_i:y, é uma correspondência *isDisjointWith*

Onde *x* e *y* são elementos (conceitos ou propriedades) pertencentes às ontologias O_1 e O_2 que representam, neste trabalho, respectivamente, ontologias de pontos vizinhos semanticamente relacionados.

Para cada tipo de correspondência o *SemMatcher* atribui um valor entre 0 e 1, que reflete o grau de proximidade semântica entre conceitos e que se apresenta definido da seguinte forma [Souza *et al.* 2009] :*isEquivalentTo* (1.0), *isSubConceptOf* (0.8), *isSuperConceptOf* (0.8), *isCloseOf* (0.7), *isPartOf* (0.3), *isWholeOf* (0.3) e *isDisjointWith* (0.0). O grau de proximidade varia, do mais próximo, a equivalência (*isEquivalentTo*), ao mais distante, a disjunção (*isDisjointWith*).

Como ilustração da geração obtida pelo *SemMatcher* apresentamos, a seguir, um exemplo que considera uma ontologia de domínio (Figura 4.2) sobre organismos vivos.

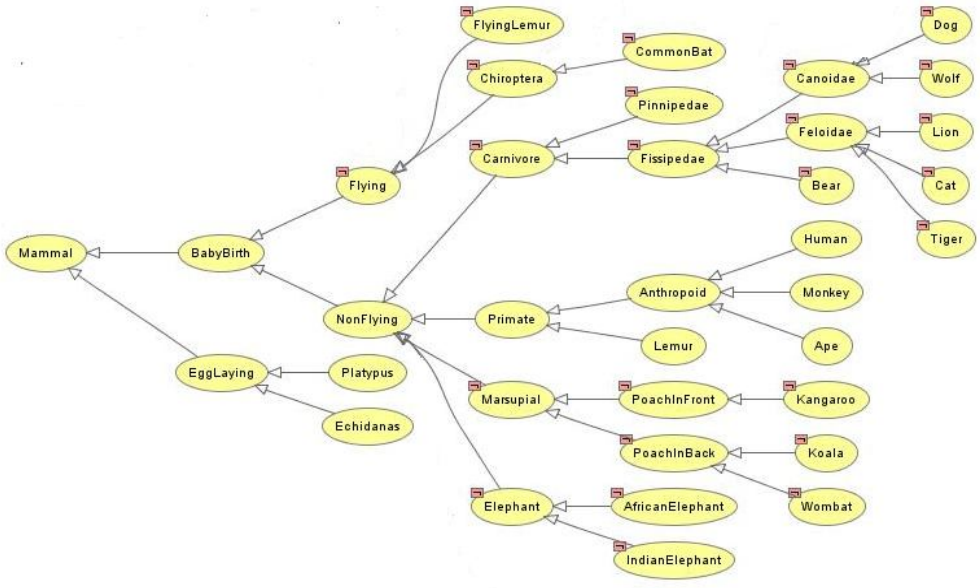


Figura 4.2 - Fragmento da ontologia de domínio sobre organismos vivos.

Exemplo 1: Seja P1 e P2 dois pontos cujos esquemas estão representados por duas ontologias O1 e O2, respectivamente. Neste cenário, o *SemMatcher* foi utilizado para identificar as correspondências semânticas entre O1 e O2. A Figura 4.3 mostra parte

do alinhamento produzido, onde podemos observar alguns tipos de correspondências, por exemplo, o termo *Human* em O1 possui um termo equivalente em O2. Além disto, outro tipo de correspondência, menos usual, foi identificada (*isCloseTo*) para o termo *Human,* no caso, *Monkey*. A correspondência *isCloseTo* indica que os elementos *Human* e *Monkey* são elementos irmãos e possuem o mesmo conceito ancestral na ontologia de domínio (Figura 4.2).

Ontology 1	Correspondence	Ontology 2
Anthropoid	isSubConceptOf	Primate
Anthropoid	isSuperConceptOf	Monkey
Anthropoid	isSuperConceptOf	Human
Human	isEquivalentTo	Human
Human	isCloseTo	Monkey
Lemur	isSubConceptOf	Primate
Monkey	isEquivalentTo	Monkey
Primate	isEquivalentTo	Primate
Primate	isSubConceptOf	NonFlying
Primate	isDisjointWith	Rodent
Primate	isDisjointWith	Elephant
Primate	isDisjointWith	Marsupial

Figura 4.3 - Fragmento de um Alinhamento produzido pelo *SemMatcher*

4.1.2. SemRef

A *SemRef* [Souza *et al.* 2009] é uma estratégia para reformulação de consultas baseada em semântica que, com o objetivo de evitar reformulações vazias (quando termos da consulta que está sendo reformulada não encontra correspondentes exatos no ponto destino) e de melhorar o conjunto de resultados da consulta submetida, faz a reformulação da consulta entre dois pontos de forma enriquecida, combinando técnicas de expansão [Carpineto *et al.* 2012] e de personalização [Tanca *et al.* 2011] da consulta.

Para expandir a consulta com termos semanticamente relacionados, o reformulador *SemRef* utiliza as correspondências semânticas. Em relação à personalização da consulta, preferências do usuário são consideradas como forma de estabelecer se o usuário deseja o enriquecimento e, caso afirmativo, o tipo de enriquecimento a ser alcançado no momento da reformulação da consulta. Além disso, o usuário pode definir se deseja obter respostas por meio do enriquecimento ou se somente respostas exatas. Caso opte pelo enriquecimento, quatro tipos de enriquecimento podem ser produzidos pelo *SemRef* em uma dada consulta Q:

- **Approximate -** inclui conceitos que estão próximos dos conceitos de Q (com base na correspondência *iscloseto*);
- **Specialize -** inclui conceitos que são sub-conceitos dos conceitos de Q (com base na correspondência *isSubConceptTo*);
- **Generalize -** inclui conceitos que são super- conceitos dos conceitos de Q (com base na correspondência *isSuperConceptTo*);
- **Compose -** inclui conceitos que são parte-de ou todo-de conceitos de Q (com base na correspondência *PartOf e WholeOf*);.

De acordo com o modo de reformulação (exato ou enriquecido) e o tipo de enriquecimento estabelecido pelo usuário, por meio das variáveis de enriquecimento (*approximate*, *specialize*, *generalize*, *compose*), dois tipos de consultas reformuladas podem ser produzidas:

(i) uma **exata**, onde apenas as correspondências do tipo equivalência são consideradas e,

(ii) uma **enriquecida**, onde todas as outras correspondências, além da equivalência, são consideradas.

No *SemRef* as consultas podem ser formuladas considerando os elementos providos pela ontologia do ponto de submissão da consulta usando SPARQL[5] ou ALC/DL [Baader *et al*. 2003]. Para ilustrar a reformulação, considerando o cenário do exemplo 1, suponha que um usuário formule uma consulta *Q = {Human}* em *P1* e solicite todos os possíveis tipos de enriquecimento (A*pproximate*, *Specialize*, *Generalize*, *Compose*). Considerando o ponto P2 como destino, o *SemRef* irá produzir as seguinte reformulações para P2:

- Reformulação Exata, na forma *Qexact = {Human}*, onde apenas as correspondências de equivalência são consideradas.
- Reformulação Enriquecida, na forma *Qenriched={Monkey}*, onde são incluídos termos obtidos dos demais tipos de correspondências (*isSubConceptOf, isSuperConceptOf*, *isCloseTo*, *isPartOf, isWholeOf,*

[5] 1 http://www.w3.org/TR/rdf-sparql-query/

isDisjointWith). Neste caso, observamos que *Monkey* é um termo obtido a partir da correspondência *isCloseTo*.

Considerando a existência de mais pontos, este processo de reformulação irá se repetir sempre que a consulta for encaminhada entre pares de pontos vizinhos do sistema. Ao longo do roteamento, a perda ou adição de termos pode provocar a perda semântica da consulta original. A próxima seção apresenta e define este problema.

4.2. O Problema da Perda Semântica da Consulta Original no Roteamento

Como comentado anteriormente, fazer uma consulta em um PDMS significa receber informações do próprio ponto e de outros pontos, desde que existam caminhos semânticos (conexões estabelecidas entre os pontos por meio de correspondências semânticas) pelos quais a consulta poderá ser encaminhada.

Neste trabalho, utilizamos a definição de consulta apresentada em Souza (2009). Assim, uma consulta Q expressa de acordo com a ontologia de um ponto P_i segue a seguinte forma em lógica descritiva $Q = Q_1 \sqcup Q_2 \sqcup ... \sqcup Q_M$, onde $Q_i = t_1 \sqcap t_2 \sqcap ... \sqcap t_n$, e t_j é um termo associado a um elemento (conceito ou propriedade) da ontologia de P_i. Ao longo do roteamento, a consulta, a cada encaminhamento de um ponto a outro do sistema, precisa ser reformulada, ou seja, reescrita de acordo com o esquema do ponto para o qual está sendo dirigida. Neste processo, consideramos que a consulta reformulada deve preservar da melhor forma possível a semântica da consulta original. Por semântica da consulta formulamos a seguinte definição:

Definição 1 (**Semântica da Consulta**): Seja P um ponto, O uma ontologia que representa o esquema de P, Q uma consulta submetida em P e $T=\{t_i\}$, i=1..n, o conjunto de termos de Q onde cada t_i está relacionado a um elemento (conceito ou propriedade) de O. Dizemos que a semântica da consulta Q é dada pelo conjunto de termos $\{t_1, ..., t_n\} \in T$ que compõem Q.

Entretanto, devido à heterogeneidade das fontes, a consulta, ao ser reformulada, pode não encontrar termos equivalentes no esquema do ponto destino

gerando reformulações com alterações semânticas em relação à semântica da consulta original. Neste cenário, a sucessiva perda de termos, ao longo do roteamento, pode produzir respostas inadequadas à consulta original submetida pelo usuário [Kantere *et al.* 2011]. Para melhorar o conjunto de resultados e evitar reformulações vazias, estratégias de reformulação por expansão [Carpineto e Romano, 2012] fazem a adição de novos termos semanticamente relacionados com a consulta a ser expandida em cada ponto.

Ao longo do roteamento, consideramos que as alterações semânticas geradas na consulta reformulada, em função da adição (expansão da consulta com termos sem correspondência semântica com os termos da consulta original) ou ausência de termos (perda de termos equivalentes aos termos da consulta original), podem levar à perda semântica da consulta reformulada em relação a sua semântica original. Neste trabalho, definimos a perda semântica da consulta, gerada ao longo do roteamento, da seguinte forma:

Definição 2 (Perda Semântica da Consulta): Seja $P=\{P_1, ...,P_n\}$ um conjunto de pontos relacionados por meio de correspondências semânticas, Q a consulta original submetida pelo usuário em um ponto P_i, $T=\{t_1,...,t_n\}$ um conjunto de termos utilizados na sua formulação, Qref uma versão reformulada de Q e $Tr=\{tr_1,..., tr_n\}$ o conjunto de termos de Qref. Dizemos que a perda semântica de Qref ocorre, quando reformulada de P_i para P_j, sse $(\nexists tr \mid (tr \equiv t)) \vee (\exists tr: \exists t \mid \neg (tr \sqsupset_{\rightarrow} t) \wedge \neg (tr \rhd_{\rightarrow} t) \wedge \neg (tr \lhd_{\rightarrow} t) \wedge \neg (tr \approx_{\rightarrow} t) \wedge \neg (tr \sqsubset_{\rightarrow} t) \wedge \neg (tr \equiv_{\rightarrow} t))$.

Para mostrar a perda semântica da consulta original, ao longo do roteamento, apresentamos um exemplo definido em um PDMS, formado pelos pontos P1, P2, P3 e P4, que compartilha informações sobre organismos vivos. Neste sistema, cada ponto disponibiliza uma ontologia que representa os dados a serem compartilhados com os demais pontos.

Exemplo 2: Para o cenário descrito, suponha que um usuário tenha submetido uma consulta Q, no ponto P1, com o objetivo de obter dados sobre primatas semelhantes ao homem. Utilizando a ontologia do ponto P1 o usuário submete a consulta **Q=Anthropoid** (Figura 4.4).

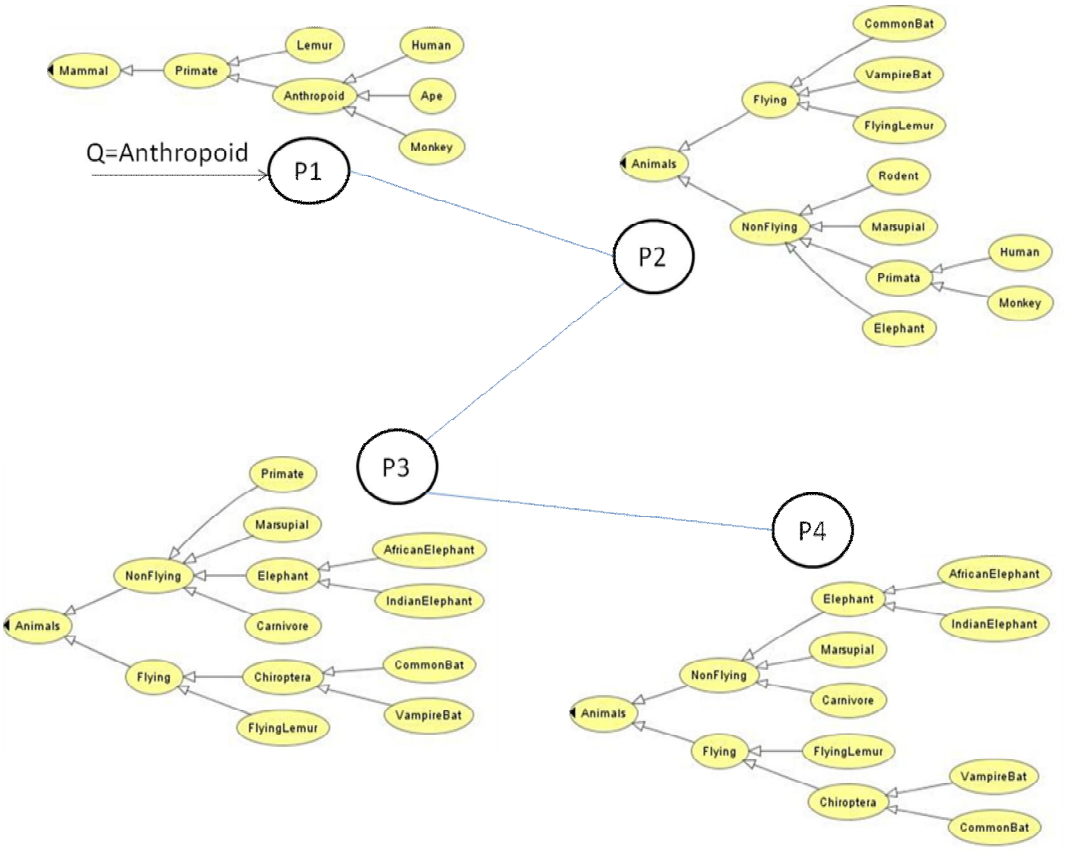


Figura 4.4- Cenário do Sistema para o Exemplo 2

Sabendo que o usuário solicitou o enriquecimento da consulta por meio das variáveis *approximate*, *specialize* e *generalize*, e, considerando o conjunto de correspondências existentes entre os pontos, são produzidas as seguintes reformulações, a cada encaminhamento da consulta:

1. Do ponto P1 para P2

Para a reformulação de *Q* em Q_{12} serão consideradas as correspondências definidas entre as ontologias dos pontos *P1* e *P2*. A Figura 4.5 mostra as correspondências relacionadas ao conceito **Anthropoid**. De acordo com o enriquecimento definido pelo usuário, são consideradas na reformulação as correspondências do tipo *isEquivalentTo*, *isSubConceptOf, isSuperConceptOf* e *isCloseTo*. É possível observar na Figura 4.5 que não existe um conceito equivalente a **Anthropoid**. Logo, como foi solicitado o enriquecimento da consulta, *Q* será reformulada na consulta Q_{12} com os conceitos **Q_{12}= Primate, Monkey, Human.**

Ontology 1	Correspondence	Ontology 2
Anthropoid	isSubConceptOf	Primate
Anthropoid	isSuperConceptOf	Monkey
Anthropoid	isSuperConceptOf	Human
Human	isEquivalentTo	Human
Human	isCloseTo	Monkey
Lemur	isSubConceptOf	Primate
Monkey	isEquivalentTo	Monkey
Primate	isEquivalentTo	Primate
Primate	isSubConceptOf	NonFlying
Primate	isDisjointWith	Rodent
Primate	isDisjointWith	Elephant
Primate	isDisjointWith	Marsupial

Figura 4.5 - Fragmento das correspondências entre as ontologias de P1 e P2

2. Do ponto *P2* para *P3*

Estando em *P2*, para encaminhar a consulta Q_{12} para *P3*, será realizada uma nova reformulação utilizando para isto as correspondências entre as ontologias dos pontos *P2* e *P3* (Figura 4.6). Da mesma forma, é possível observar que não existem conceitos equivalentes aos solicitados em **Q_{12}= Primate, Monkey, Human**. Logo, o enriquecimento da consulta será realizado e Q_{12} será reformulada gerando a consulta **Q_{23}= Primate, NonFlying**. A partir deste ponto é possível já observar um distanciamento semântico da consulta reformulada com relação à consulta original. **NonFlying** é um conceito que envolve subconceitos como **Elephant**, **Marsupial** e **Carnivore** (Figura 4.6) que não atendem ao que foi solicitado originalmente na consulta. Entretanto, apesar de **NonFlying** ser semanticamente distante do conceito originalmente solicitado, **Primate** possui uma relação semântica mais próxima, o que poderia justificar um possível encaminhamento da consulta de *P2* para *P3*.

Ontology 1	Correspondence	Ontology 2
Flying	isEquivalentTo	Flying
Flying	isSuperConceptOf	FlyingLemur
Flying	isSuperConceptOf	Chiroptera
Flying	isDisjointWith	NonFlying
FlyingLemur	isEquivalentTo	FlyingLemur
FlyingLemur	isSubConceptOf	Flying
FlyingLemur	isDisjointWith	Chiroptera
Marsupial	isEquivalentTo	Marsupial
Marsupial	isSubConceptOf	NonFlying
Marsupial	isDisjointWith	Primate
Marsupial	isDisjointWith	Elephant
Marsupial	isDisjointWith	Carnivore
NonFlying	isEquivalentTo	NonFlying
NonFlying	isSuperConceptOf	Primate
NonFlying	isSuperConceptOf	Elephant
NonFlying	isSuperConceptOf	Marsupial
NonFlying	isSuperConceptOf	Carnivore
NonFlying	isDisjointWith	Flying
Primate	isEquivalentTo	Primate
Primate	isSubConceptOf	NonFlying
Primate	isDisjointWith	Elephant
Primate	isDisjointWith	Marsupial
Primate	isDisjointWith	Carnivore
Rodent	isSubConceptOf	NonFlying

Figura 4.6- Correspondências entre as ontologias de P2 e P3

3. Do ponto P3 para P4

Fazendo mais uma vez o encaminhamento da consulta de *P3* para *P4* e, consequentemente a sua reformulação utilizando as correspondências semânticas entre *P3* e *P4* (Figura 4.7), é possível observar a perda semântica total da consulta. A consulta **Q_{23}= Primate, NonFlying,** ao ser reformulada para *P4,* perde o conceito **Primate** e produz o enriquecimento de Q_{23} a partir do conceito **NonFlying**. Logo, Q_{23} será reformulada em Q_{34} com os conceitos **Q_{34}=NonFlying, Elephant, Marsupial, Carnivore.** Comparando os conceitos obtidos na reformulação podemos observar não apenas a perda, mas o distanciamento semântico da consulta original obtido com as sucessivas reformulações. Em $Q_{34,}$ o conceito **NonFlying** está relativamente distante do conceito original, sendo um conceito mais amplo ao que foi originalmente solicitado. Por outro lado, **Elephant**, **Marsupial** e **Carnivore** são disjuntos de **Anthropoid** e, por esta razão, não atendem à consulta original.

Ontology 1	Correspondence	Ontology 2
Elephant	isDisjointWith	Carnivore
Flying	isEquivalentTo	Flying
Flying	isSuperConceptOf	FlyingLemur
Flying	isSuperConceptOf	Chiroptera
Flying	isDisjointWith	NonFlying
FlyingLemur	isEquivalentTo	FlyingLemur
FlyingLemur	isSubConceptOf	Flying
FlyingLemur	isDisjointWith	Chiroptera
IndianElephant	isEquivalentTo	IndianElephant
IndianElephant	isSubConceptOf	Elephant
Marsupial	isEquivalentTo	Marsupial
Marsupial	isSubConceptOf	NonFlying
Marsupial	isDisjointWith	Elephant
Marsupial	isDisjointWith	Carnivore
NonFlying	isEquivalentTo	NonFlying
NonFlying	isSuperConceptOf	Elephant
NonFlying	isSuperConceptOf	Marsupial
NonFlying	isSuperConceptOf	Carnivore
NonFlying	isDisjointWith	Flying
Primate	isSubConceptOf	NonFlying
Primate	isDisjointWith	Elephant
Primate	isDisjointWith	Marsupial
Primate	isDisjointWith	Carnivore
VampireBat	isEquivalentTo	VampireBat
VampireBat	isSubConceptOf	Chiroptera

Figura 4.7- Correspondências entre as ontologias dos pontos P3 e P4

No exemplo, é possível observar que a perda semântica ocorre sempre em relação aos termos da consulta original. Se considerarmos apenas a consulta corrente, aquela que chega ao ponto para uma nova reformulação, é possível observar que todos os termos expandidos possuem correspondentes semânticos adequados à consulta corrente que foi reformulada. Outro ponto importante que deve ser analisado é como identificar esta perda semântica a cada novo encaminhamento da consulta.

4.3. Informação Semântica e de Qualidade para o Roteamento de Consultas

Em um PDMS, devido à ausência de um controle centralizado, ao receber uma consulta, cada ponto precisa decidir, baseado no conhecimento local que tem da rede, para qual dos seus vizinhos semânticos a consulta deverá ser encaminhada. Consideramos que a informação semântica em torno deste processo de decisão e seleção é importante se desejamos expandir o conhecimento do ponto em relação aos

seus vizinhos e tornar mais específica a seleção dos pontos para o encaminhamento da consulta. De forma complementar, medidas de qualidade, produzidas em função da análise da qualidade da informação (QI) relacionada ao ponto, podem melhorar este conhecimento ajudando a classificar os melhores pontos e produzir as melhores rotas para a consulta [Freire 2012].

Dizemos que a rota da consulta é dada pelo conjunto de pontos capazes de responder à consulta, ao longo do roteamento, partindo do ponto origem (ponto de submissão da consulta) até os pontos destinos, conforme descrição a seguir.

Definição 3 (*Rota da Consulta*): Considerando $P=\{P_1,...,P_n\}$ um conjunto de pontos e Q uma consulta submetida em P_i, dizemos que $RQ_{ik}=\{P_i,...,P_k\}$ é uma rota para a consulta Q do ponto P_i até o ponto P_k, se existe um caminho, estabelecido por meio de correspondências semânticas, entre pares de pontos partindo de P_i até P_k que possa responder a Q.

Consideramos que no processo de roteamento é preciso que cada ponto, ao receber uma consulta, realize e coordene as seguintes atividades: (i) execução da consulta (pelo ponto que recebe a consulta); (ii) seleção de pontos (considerando os vizinhos que podem melhor responder a consulta); (iii) reformulação da consulta (para cada ponto vizinho selecionado); (iv) encaminhamento da consulta (depois de reformulada para cada ponto vizinho selecionado); e (v) integração dos resultados das consultas (do próprio ponto e dos vizinhos para os quais uma versão reformulada da consulta tenha sido encaminhada).

Durante a atividade de seleção, diferentes aspectos podem influenciar a escolha de pontos e a geração de rotas para a consulta. Neste cenário, o uso de informações contextuais (conjunto de elementos que caracterizam uma situação) [Dey 2001] e de critérios de qualidade [Hose *et al.* 2008] podem permitir que o processo de seleção de pontos e de definição de rotas para a consulta seja realizado de forma mais específica e dinâmica.

O contexto associado à atividade de seleção de pontos é definido para cada ponto que recebe a consulta, a partir do conjunto de elementos contextuais associados à consulta e a cada vizinho semântico. Alguns destes elementos são definidos explicitamente, por meio de preferências do usuário em relação à consulta,

outros são dinamicamente percebidos no ambiente, a exemplo da disponibilidade do ponto, ou calculados, a exemplo de medidas de qualidade associadas ao ponto. Preferências do usuário definem o tipo de enriquecimento esperado a cada reformulação e podem priorizar os critérios de qualidade que são utilizados na classificação dos melhores pontos. Os critérios de qualidade utilizados para produzir as medidas de QI serão detalhados na Seção 4.4.2.

4.3.1. Um Modelo para Representação da Informação Semântica e de Qualidade

Para permitir o uso da informação semântica e de qualidade da informação (QI) é importante definir como estes conceitos serão representados. O modelo para representação das informações associadas ao roteamento foi desenvolvido com base no metamodelo apresentado em Souza *et al.* (2012).

O metamodelo provê construtores que podem combinar a informação semântica (por exemplo, informações obtidas por meio de ontologias e informação contextual), com medidas obtidas a partir de critérios de QI (por exemplo, relevância). Combinando tais conceitos, é possível produzir o conhecimento semântico relacionado às atividades que compõem um determinado processo (por exemplo, a seleção de pontos). Os metaconstrutores (metaconceitos) definidos no metamodelo podem ser reusados em outros modelos para fins específicos. Neste trabalho, foi desenvolvido um modelo para o roteamento de consultas em PDMS. A Figura 4.8 mostra o modelo proposto com seus principais conceitos e relacionamentos. Os conceitos e relacionamentos identificados pelo prefixo *meta* pertencem ao metamodelo e estão sendo reutilizados.

Semantic_Information e *Information_Quality* são subconceitos que estão relacionados aos tipos de informações (conceito *Information*) representadas pelo modelo. Uma entidade de domínio (*Domain_Entity*) refere-se a qualquer objeto do mundo real que descreve o domínio de interesse da aplicação. No roteamento identificamos três entidades de domínio: o ponto (*peer*), o usuário (*user*) e a consulta (*query*). *Query_Configuration* e *Reformulation_Mode* representam os tipos de preferências (*User_Preferences*) estabelecidas pelo usuário (*User*) e que estão relacionadas à consulta (*Query*).

Definimos o roteamento da consulta (*Query_Routing*) como um subconceito de processo (*meta:Process)*. O contexto (*Context*) é composto pelo conjunto de elementos contextuais (*ContextualElements*) que tornam as ações relacionadas a uma atividade (*Activity*) mais específicas. *Peer_Selection*, *Query_Forwarding*, *Results_Integration*, *Query_Execution* e *Query_Reformulation* são atividades coordenadas pelo processo de roteamento da consulta (*Query_Routing*). Entretanto, a ênfase deste trabalho está na atividade de seleção de pontos (*Peer_Selection*). Na *SemRouting*, elementos contextuais são utilizados durante a atividade de seleção de pontos para o encaminhamento da consulta. Existem os elementos cujos valores são pré-definidos, como as preferências do usuário em relação à consulta (*Query_Configuration* e *Reformulation_Mode*), os que são calculados dinamicamente, a exemplo das medidas de qualidade (*Quality_Measure e GlobalIQ*) e o elemento que identifica a disponibilidade (Availability) do ponto obtido no momento da atividade de seleção de pontos.

O conceito *Information_Quality* é definido pela composição de critérios de qualidade (*Quality_Criterion)*. Critérios de qualidade fornecem medidas (*Quality_Measure*) que são utilizadas como um tipo de elemento contextual (*ContextualElement*). As medidas de qualidade são utilizadas para geração de uma medida global de qualidade, definida no modelo como *GlobalIQ*. O *GlobalIQ* representa uma medida única da qualidade de informação associada ao ponto. Assim como *QualityMeasure*, o *GlobalIQ* também está representado como um elemento contextual (*ContextualElement)*. Outra visão mais detalhada do modelo pode ser encontrada no Apêndice A.

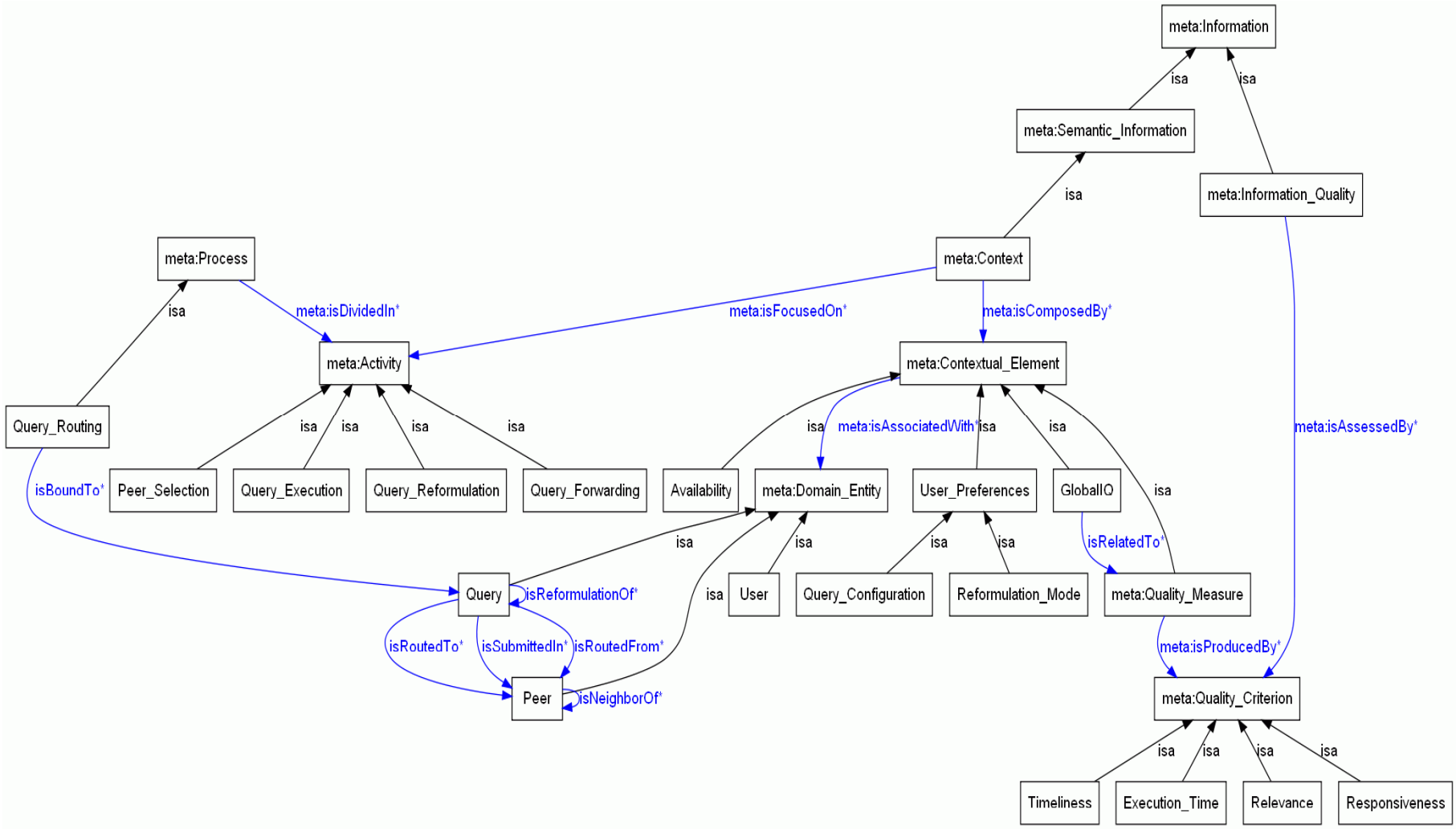


Figura 4.8- Modelo para Representação das Informações do Roteamento

4.3.2. Exemplificando o Modelo

Para exemplificar o uso do modelo, definimos um cenário composto por seis pontos P1, P2, P3, P4, P5 e P6. Cada um dos pontos armazena dados relacionados ao domínio de "educação". A Figura 4.9 mostra o cenário de vizinhança dos pontos instanciados, com base no modelo. Para simplificar a exibição do cenário, as propriedades relacionadas ao ponto P1 foram omitidas.

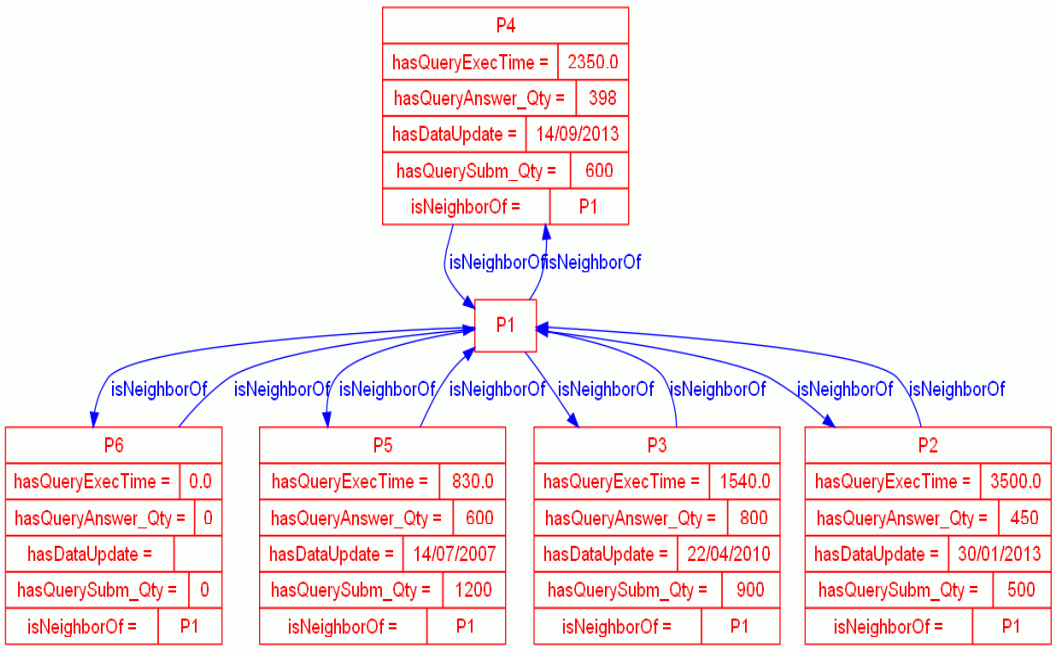


Figura 4.9 Cenário de Pontos

Neste cenário um usuário "Ana", submete em P1 uma consulta Q1, composta pelos termos {Professor, Manual}. Preferências do usuário "Ana" em relação à consulta Q1 foram definidas, ou seja, o tipo de enriquecimento e pesos associados aos critérios de qualidade a serem utilizados durante o roteamento de Q1. A Figura 4.10 mostra a instanciação no modelo da etapa correspondente à submissão da consulta.

Estando em P1, cada ponto vizinho passa a ser identificado como um ponto candidato e, para cada um deles, serão geradas medidas de qualidade para análise da QI do ponto. Em função das medidas, uma medida única (*GlobalIQ*) é produzida para análise da QI dos pontos candidatos. A forma de cálculo desta medida será melhor apresentada na Seção 4.4.3.

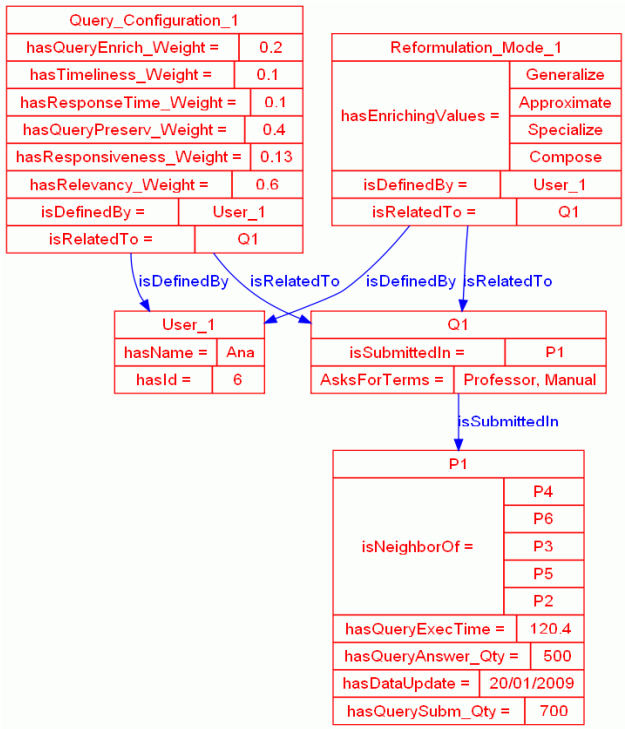


Figura 4.10- Submissão da Consulta Q1

A Figura 4.11 mostra a instanciação da atividade de seleção no momento de escolha dos pontos para encaminhamento da consulta de acordo com o *GlobalIQ*. Neste cenário, de acordo com um determinado *threshold* de qualidade , P1 fará a escolha dos melhores pontos para o encaminhamento de Q1. Todo este processo de seleção e classificação dos pontos será descrito a partir da Seção 4.4 que descreve a abordagem *SemRouting*.

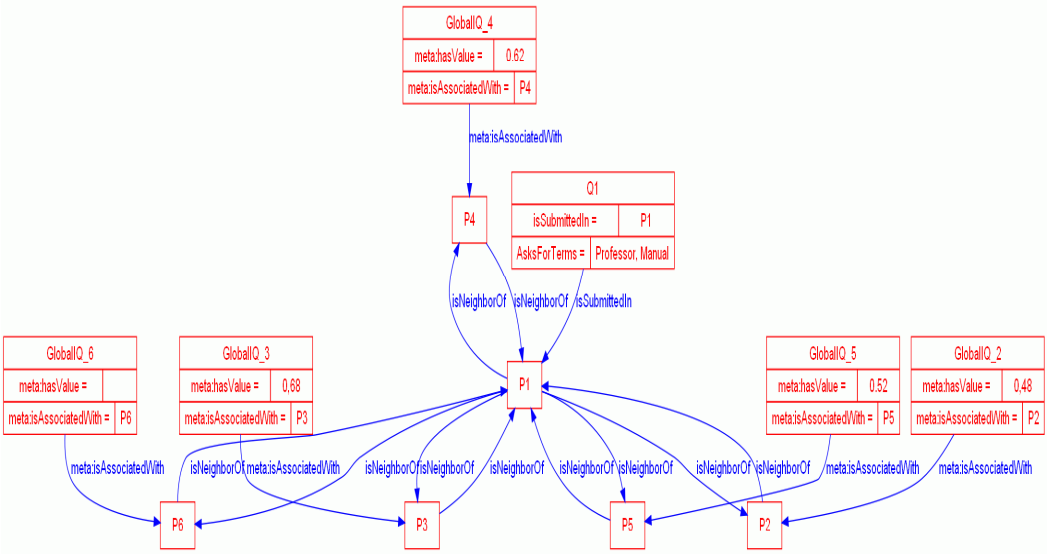


Figura 4.11- Valores de GlobalIQ na Seleção de Pontos

O modelo permite representar o conhecimento local do ponto com informações relacionadas aos seus vizinhos semânticos. Considerando a vizinhança do ponto, as informações semânticas e de qualidade, definimos o conhecimento local do ponto para o roteamento de consultas em um PDMS da seguinte forma:

Definição 4 (Conhecimento Local do Ponto): Dado um PDMS formado por um conjunto de pontos $\{P_1..P_n\}$ semanticamente relacionados por meio de correspondências, dizemos que o conhecimento local de P_i é definido como uma tripla $<PV, Q_{pv}, S>$ onde, PV corresponde aos vizinhos de P_i, Q_{pv} as informações de qualidade associadas a cada ponto em PV e S um conjunto de informações semânticas de P_i (correspondências semânticas entre P_i e cada P_j em PV e o contexto das atividades realizadas em P_i).

Em relação ao conhecimento local do ponto, dados armazenados referentes ao contexto das decisões tomadas pelo ponto, em atividades de seleção anteriores, poderão ser futuramente utilizados para que o ponto venha a inferir novo conhecimento que o ajude a tomar decisões de forma mais rápida.

4.4. A Abordagem SemRouting

A abordagem *SemRouting*, proposta nesta tese, combina aspectos semânticos e de qualidade com o objetivo de definir o roteamento selecionando o melhor conjunto de pontos capazes de responder a consulta e preservando da melhor forma possível a semântica da consulta original. As próximas subseções descrevem os elementos que compõem toda a abordagem.

4.4.1. Estratégia para Análise e Preservação Semântica

Na *SemRouting*, a estratégia para análise e preservação semântica da consulta original consiste na realização de duas etapas: (i) criar uma referência semântica para a consulta original, no momento de sua submissão; (ii) realizar a poda semântica, eliminando da consulta reformulada termos que não estejam presentes na referência semântica.

(i) Criando a Referência Semântica

A Referência Semântica (RS) é uma estrutura definida com base na ontologia de domínio (OD) que armazena informações sobre os termos da consulta original e, para cada termo, os termos associados na ontologia de domínio (OD) que possuem relação semântica direta com os termos originais. Definimos *Relação Semântica Direta* e *Referência Semântica* da seguinte forma:

Definição 5 (Relação Semântica Direta): Sejam O_1 e O_2 ontologias, dizemos que um termo (conceito ou propriedade) T_1 em O_1 tem relação semântica direta com T_2 em O_2, ou seja, $T_1 \overrightarrow{RSD} T_2$, se existe um relacionamento definido entre T_1 e T_2 por meio de uma correspondência Co, onde Co ∈ {*isEquivalentTo, isSubConceptOf, isSuperConceptOf, isCloseTo, isPartOf, isWholeOf*}.

Na definição de RSD, a correspondência *isDisjointWith* não é considerada por definir uma ausência de relação semântica entre termos.

Definição 6 (Referência Semântica): Seja Q a consulta original submetida no ponto P_i, de acordo com sua ontologia local O_i, $T=\{t_{1..tn}\}$ o conjunto de termos de Q e OD uma ontologia de domínio do sistema. Uma Referência Semântica (RS) é definida como um conjunto de triplas da seguinte forma: RS = {<t_i, t_j, w>}, onde t_i é um termo de Q ∈ O_i, t_j é um termo ∈ OD e $t_i \overrightarrow{RSD} t_j$; w é o peso associado a cada correspondência semântica existente entre t_i e t_j.

Para geração da RS, é preciso considerar a consulta original, a ontologia de domínio e as preferências estabelecidas pelo usuário em relação ao tipo de enriquecimento esperado. O algoritmo que gera a RS é mostrado na Figura 4.12.

Inicialmente, a consulta (Q) do usuário (linha 2), o alinhamento (*A*) produzido entre a ontologia do ponto de submissão da consulta e a OD (linha 3) e o conjunto de variáveis (*VEnriq*) relacionadas ao tipo de enriquecimento solicitado pelo usuário (linha 4) são passados como parâmetros de entrada. Para cada termo da consulta original (linha 7) e, considerando cada tipo de enriquecimento solicitado pelo usuário (linha 9), o algoritmo busca em *A* de acordo com *VEnriq* (que estabelece qual tipo de correspondência será considerada) o termo *tc* existente na OD que tenha RSD com o termo *t* da consulta original (linha 11). Enquanto existirem termos relacionados (linha 12), identifica-se o tipo de correspondência que estabelece a RSD entre *tc* e *t* (linha

14). Em seguida, o peso associado ao tipo de correspondência é obtido (linha 15) e os valores inseridos na RS (linha 16). A linha 17 indica a busca em *A* por um novo termo dentro da mesma condição estabelecida por *VEnriq*. Ao final do processo, a RS formada por um conjunto de triplas compostas pelo termo *tc* da consulta original, termo *t* obtido em *A*, e o peso associado à correspondência semântica identificada a partir da RSD entre *tc* e *t* é retornada (linha 21).

Manter os pesos associados a cada correspondência identificada entre os termos da consulta original e os termos da OD na RS, garante a identificação de termos originais recuperados em reformulações posteriores e a avaliação correta da similaridade entre os termos da consulta reformulada em relação aos termos originais.

```
1. Algoritmo Cria_Referencia_Semantica;
2. Entrada Q: consulta original;
3.          A: alinhamento;  // entre a ontologia do ponto de submissão de Q e a OD
4.          VEnriq: <approximate, generalize, specialize,
            compose:lógico>
5. Saída RS: referência semântica;
6. Inicio
7.    Para cada termo t em Q faça
8.    Inicio
9.      Para cada(VEnriq = TRUE) faça
10.     Inicio
11.       Encontre em A o termo tc RSD t de acordo com VEnriq;
12.       Enquanto (tc ≠ vazio) faça
13.       Inicio
14.         corresp←obtem tipo de correspondência que define tc RSD t;
15.         w←obtém peso associado à corresp;
16.         Inclui <t, tc, w> em RS;
17.         Encontre em A o próximo termo tc RSD t de acordo com
            VEnriq;
18.       Fim
19.     Fim
20.   Fim
21.   Retorne RS;
22. End;
```

Figura 4.12- Algoritmo Cria_Referência_Semântica

No cenário do Exemplo 2, descrito na Seção 4.2, observamos que do ponto P1 para o ponto P2, o termo **Primate** obtido na reformulação, por ser um subconceito de **Anthropoid**, tem peso 0.8. Do ponto P2 para P3, o mesmo conceito **Primate** foi mantido por ter um correspondente equivalente no esquema do ponto P3. Entretanto, em relação ao termo **Antrhropoid** da consulta original a sua

correspondência semântica continua sendo de subconceito. Assim, para preservar o valor semântico do termo em relação ao termo original, o valor de correspondência deve ser mantido em 0.8 e não 1.0. Apesar de não constar no exemplo, o próprio termo original **Anthropoid** poderia ter sido recuperado em reformulações posteriores a partir, por exemplo, de uma correspondência *isCloseTo*. Neste caso, em vez de ter valor igual a 1.0 indicando equivalência com o termo original (que indica uma recuperação do termo original na reformulação), teria peso 0.7, o que seria incorreto.

(ii) Realizando a Poda Semântica

O processo de retirada de termos da consulta reformulada que alteram a semântica da consulta original é denominado na *SemRouting* de Poda Semântica. Definimos Poda Semântica da seguinte forma:

Definição 7 (Poda Semântica): Seja Q a consulta original submetida em um ponto P e Qref uma versão reformulada de Q ao longo do roteamento. Dizemos que a poda semântica é o processo de retirada de termos de Qref que não estão presentes na RS.

Na poda semântica, a retirada de termos da consulta reformulada é realizada logo após a etapa de reformulação da consulta. Como mostra a Figura 4.13, a consulta Q é reformulada gerando Qref. Em seguida, a poda semântica é realizada usando a RS para remover de Qref todos os termos que não estejam presentes na RS, gerando uma consulta Qref' mais próxima semanticamente dos termos originalmente solicitados.

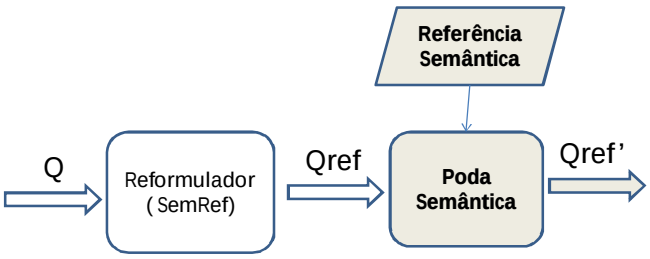


Figura 4.13- Preservação Semântica

Para ilustrar o mecanismo de Poda Semântica, considere novamente o exemplo descrito na Seção 4.2. Inicialmente, é gerado no ponto de submissão da consulta (P1) o alinhamento entre a ontologia de P1 e a OD sobre organismos vivos. A Figura 4.14, mostra um fragmento deste alinhamento onde **Ontology 1** identifica a ontologia de P1 e **Ontology 2** a ontologia de domínio.

Ontology 1	Correspondence	Ontology 2	Measure
Anthropoid	isEquivalentTo	Anthropoid	1
Anthropoid	isSubConceptOf	Primate	0,8
Anthropoid	isSuperConceptOf	Ape	0,8
Anthropoid	isSuperConceptOf	Monkey	0,8
Anthropoid	isSuperConceptOf	Human	0,8
Anthropoid	isCloseTo	Lemur	0,7
Ape	isEquivalentTo	Ape	1
Ape	isSubConceptOf	Anthropoid	0,8
Human	isEquivalentTo	Human	1
Human	isSubConceptOf	Anthropoid	0,8

Figura 4.14- Fragmento do alinhamento entre Ontologia de P1 e OD

Considerando o alinhamento e as preferências estabelecidas pelo usuário (*approximate*, *specialize* e *generalize*), a referência semântica para a consulta **Q=Anthropoid**, seria produzida da seguinte forma:

RS={<Anthropoid, Anthropoid, 1,0>, <Anthropoid, Primate, 0,8>, <Anthropoid, Ape, 0,8>, <Anthropoid, Monkey, 0,8>, <Anthropoid, Human, 0,8>, <Anthropoid, Lemur, 0,7>}.

No cenário descrito na Seção 4.2, **Q=Anthropoid** é submetida em P1 e reformulada em $\mathbf{Q_{12}}$**= Primate, Monkey, Human.** De acordo com a RS gerada, todos os termos estão presentes na RS e não serão eliminados de Q_{12}. A consulta Q_{12} é então encaminhada para P2.

Em P2, $\mathbf{Q_{12}}$ é reformulada em $\mathbf{Q_{23}}$**= Primate, NonFlying**. Neste caso, o mecanismo de poda irá remover de $\mathbf{Q_{23}}$ o termo NonFlying por não estar presente na RS. Removido o termo, $\mathbf{Q_{23}}$**= Primate** é então encaminhada para P3.

Em P3, $\mathbf{Q_{23}}$ é reformulada em $\mathbf{Q_{34}}$**=NonFlying, Elephant, Marsupial, Carnivore**. Nesta reformulação, todos os termos adicionados não estão presentes na RS e, em consequência, são eliminados de $\mathbf{Q_{34}}$. Assim, por não existir uma reformulação próxima aos termos da consulta origianl, o roteamento para P4 não será realizado.

Ao longo do roteamento, observamos que a estratégia de preservação semântica garante não só a preservação semântica da consulta reformulada em relação à consulta original, como também a correta interpretação dos termos obtidos em relação aos termos originais. Assim, o grau de similaridade semântica obtido, para cada termo da consulta reformulada será utilizado, posteriormente, para avaliar a relevância de um ponto em relação à consulta.

4.4.2. Critérios de Qualidade

Considerando uma consulta Q, no momento da seleção dos pontos, critérios de qualidade são usados para eliminar e classificar os melhores pontos, dentre o conjunto de pontos relevantes, para o encaminhamento da consulta. Quatro critérios foram definidos para avaliação dos pontos na rede: **Relevância**, **Capacidade de Resposta**, **Atualidade** e **Tempo de Execução**. O critério de relevância tem como objetivo eliminar pontos que não tragam resultados relevantes à consulta. Os demais critérios foram definidos de forma arbitrária para estabelecer, junto com a relevância, medidas que possam ser utilizadas para classificação dos melhores pontos dentre os relevantes.

Em função da importância do critério de Relevância para a abordagem fizemos sua definição independente dos demais critérios. A seguir apresentamos as definições e as métricas associadas à avaliação dos critérios apresentados.

4.4.2.1. Relevância

Refere-se à medida que verifica a capacidade que um ponto tem para contribuir com respostas com relação aos termos de uma dada consulta. Para o roteamento, medir a relevância significa evitar que consultas reformuladas, com alta perda semântica, sejam encaminhadas durante o roteamento. O cálculo da relevância considera duas medidas: *preservação da consulta* (PC) e *enriquecimento da consulta* (EC). Assim, definimos relevância da seguinte forma:

Definição 8 (Relevância) - Dada uma consulta Q submetida originalmente em um ponto P_i e Q_{ref} uma versão reformulada de Q de acordo com o esquema de P_j, dizemos que a relevância do ponto P_j em relação a Q_{ref} é calculada com base nas seguintes fórmulas:

- **Preservação da Consulta (PC) -** produz uma medida entre [0..1] de acordo com o número de termos originais presentes em Qref que possuem termos equivalentes em P_i.

$$PC_{Q_{ref}P_j} = \frac{|t_{eq}|}{|t_{orig}|} \qquad (1)$$

Onde,

$|t_{eq}|$, número de termos originais presentes em Q_{ref} com equivalência em P_j;

$|t_{orig}|$, número de termos de Q;

- **Enriquecimento da Consulta (EC) –** Calcula o valor do enriquecimento de Q_{ref} baseado nos termos originais de Q. Para isto, uma função RS_sim identifica na RS o peso associado à correspondência entre o termo expandido e o termo original. Considerando que o valor máximo de enriquecimento produzido em função das correspondências semânticas é de 0.8, a medida gerada para EC deverá estar no intervalo entre [0..0.8].

$$EC_{Q_{ref}P_j} = \frac{\sum_{k=1}^{nto}(\sum_{l=1}^{nte_k} RS\_sim(to_k, te_l))/nte_k}{nto} \qquad (2)$$

Onde,

|nto|, número de termos originais de Q;

to_k, k-termo original de Q;

te_l, l-termo expandido em Q_{ref};

RS_sim(to_k,te_l), valor associado ao grau de similaridade semântica entre to_k e te_l.

|nte_k|, número de termos expandidos em Q_{ref} por to_k (termo original de Q);

Assim, a relevância de P_j em relação a Q_{ref} ($Relev_{P_jQ_{ref}}$) é dada pela soma entre as medidas resultantes da aplicação das fórmulas (1) e (2):

$$Relev_{P_jQ_{ref}} = PC_{Q_{ref}P_j} + EC_{Q_{ref}P_j}$$

O valor da relevância pode variar entre [0, 1.8]. O valor de 1.8 indica a relevância máxima para reformulações em que todos os termos originais estão preservados e o máximo enriquecimento foi obtido em um dado ponto. A Figura 4.15 mostra o algoritmo que calcula a relevância de um ponto em relação à consulta.

O algoritmo recebe como entrada a consulta reformulada *Qref* e a referência semântica *RS* (linhas 2 e 3). Como a relevância é sempre calculada tomando como base os termos originais da consulta, os valores correspondentes ao número de termos originais na *RS* e na Q_{ref} são obtidos (linhas 6 e 7) e armazenados em *NTo* e *NToP*, respectivamente. Para cada termo original *Torig* em RS (linha 9) a variável que acumula o enriquecimento é inicializada (linha 11). Em seguida, é obtido o conjunto dos termos enriquecidos por termo original *Tenriq* presentes em *Qref* (linha 12).

Para cada termo expandido presente em *Tenriq* (linha 14) o algoritmo obtém na *RS* o peso semântico associado à relação semântica direta (RSD) entre o termo original

e o termo expandido armazenado na *RS* (linha 16). A soma de todos os valores associados ao enriquecimento por termo original é acumulado em *VTenriq* (linha 17). Ao terminar a leitura do conjunto de termos de *Tenriq*, a média aritmética do enriquecimento obtido por *Torig* é computada e armazenada na variável *EC* que acumula o enriquecimento da consulta (linha 20). Todo o processo se repete até que não existam termos originais a serem analisados. Ao final (linhas 22, 23 e 24), os valores de *PC* e *EC* são calculados sempre em função do número de termos originais (*NTo*) e o cálculo da relevância é computado pela adição dos valores de *PC* e *EC*.

```
1. Algoritmo Calcula_Relevância;
2. Entrada   Qref: consulta reformulada
3.           RS: Referência Semântica;
4. Saída relevância: valor da relevância;
5. Inicio
6.      NTo ← #termos originais na RS;
7.      NToP ← #termos originais preservados em Qref de acordo com RS;
8.      EC←0;
9.      Para cada termo original Torig em RS faça
10.     Inicio
11.       VTenriq←0;
12.       Tenriq ← obtém termos enriquecidos por Torig em Qref;
13.       NTe ← |Tenriq|;
14.       Para cada termo t em Tenriq faça
15.       Inicio
16.         peso ← obtém na RS o peso semântico associado à t;
17.         VTenriq ← VTenriq + peso;
18.       Fim
19.       VTenriq ← VTenriq / NTe;
20.       EC← EC + VTenriq;
21.     Fim
22.     PC ← NToP / NTo;
23.     EC ← EC / NTo;
24.     relevancia ← PC + EC;
25.   Fim
```

Figura 4.15- Algoritmo Calcula_Relevância

4.4.2.2. Demais Critérios

Para definição dos critérios de Atualidade, Capacidade de Resposta e Tempo de Execução utilizamos como referência os trabalhos de Batista (2008), Bouzeghoub e Peralta (2004), Wang e Strong (1996) e Naumann e Rolker (2000).

- **Capacidade de Resposta**

Refere-se ao grau de participação do ponto em resposta às consultas submetidas a ele em um determinado intervalo de tempo. Para obter o grau de participação é realizado um cálculo do percentual de vezes em que as consultas são respondidas pelo ponto em um determinado intervalo de tempo.

$$CRep_{P_i} = \frac{QA_{P_i}(t)}{QS_{P_i}(t)}$$

Onde,

t é o intervalo de tempo usado para medir o critério,

$QA_{P_i}(t)$ é o número de consultas respondidas por P_i no intervalo de tempo t;

$QS_{P_i}(t)$ é o número de consultas submetidas em P_i no intervalo de tempo t.

- **Atualidade**

Refere-se à medida, em dias, que corresponde a quão recente são os dados do ponto considerando a data da última atualização do mesmo.

$$Atual_{P_i} = DataCorrente - DataUltAt_{P_i}$$

Onde, DataCorrente refere-se à data atual do sistema;

$DataUltAt_{P_i}$ é a data da última atualização do ponto P_i.

- **Tempo de Execução**

Refere-se ao tempo médio de execução das consultas submetidas ao ponto. Pode ser calculado em função da média do tempo gasto em resposta às consultas submetidas ao ponto em um determinado intervalo de tempo.

$$TExec_{P_i} = \frac{\sum_{k=1}^{qn} TE_{Q_k}(t)}{QS_{P_i}(t)}$$

Onde,

qn é o número de consultas submetidas em P_i;

$TE_{Q_k}(t)$ é o tempo de execução da consulta Q_k no ponto P_i no intervalo de tempo t;

$QS_{P_i}(t)$ é o número de consultas submetidas em P_i no intervalo de tempo t.

No roteamento, as medidas de QI produzidas pelos critérios de Relevância, Capacidade de Resposta, Atualidade e Tempo de Execução serão utilizadas de forma eliminatória e classificatória. Em uma primeira análise, pontos com relevância abaixo de um dado *threshold* são eliminados. Em seguida, combinando a relevância e os demais critérios será possível estabelecer uma classificação que priorize os melhores pontos dentre os pontos relevantes. O algoritmo para geração das medidas de qualidade e eliminação dos pontos não relevantes é descrito na Figura 4.16.

Inicialmente o algoritmo recebe como entrada a consulta *Q*, o conjunto de pontos candidatos *PC*, ou seja, pontos vizinhos do ponto corrente (ponto que recebe *Q*), as preferências do usuário *PrefUsr* em relação a *Q* e a referência semântica *RS* (linhas 2 a 5). A saída do algoritmo é o conjunto das medidas de qualidade relacionadas a cada ponto em *PC* (linha 6).

No ponto corrente, para cada *p* em *PC* (linha 8), *Q* é reformulada em *Q'* de acordo com *PrefUsr* (linha 10). Para preservar a semântica da consulta a poda semântica é aplicada a *Q'*, ou seja, são retirados de *Q'* os termos que não existem na *RS*, ficando o resultado armazenado em *Qref* (linha 11). Em seguida a relevância de *p* é calculada e armazenada em *relevanciaP* (linha 12). Se *relevanciaP* for maior ou igual a θ (limite semântico para encaminhamento de *Qref*) (linha 13) então o algoritmo passa a calcular as medidas de qualidade associadas a *p*. Na linha 15 o valor de relevância de *p* é armazenado na primeira posição do vetor *mq*. Na linha 16 o algoritmo obtém no ponto corrente, os valores utilizados para calcular as medidas de qualidade de cada *p* relevante. Após o cálculo das medidas de qualidade (*capacidade de resposta*, *tempo de execução* e *atualidade*) do ponto *p*, a consulta reformulada *Qref* e o vetor contendo as medidas de qualidade são incluidos em *QPR* (linha 20) e armazenados no ponto corrente (linha 21). A saída deste algoritmo (*QPR*) corresponde ao conjunto de pontos relevantes, reformulações e suas respectivas medidas de qualidade (linha 24). Na próxima seção, veremos como estes dados serão utilizados para geração de uma medida única para análise da QI do ponto.

```
1.  Algoritmo Medidas_Qualidade
2.  Entrada   Q: consulta;
3.            PC: conjunto de pontos;
4.            PrefUsr: preferências do usuário em relação a Q;
5.            RS: referência semântica;
6.  Saída     QPR: conjunto de <p:ponto;
                                Qref: consulta;
                                mq:vetor[1..4] of real>;
7.  Inicio
8.    Para cada p em PC faça
9.    Inicio
10.     Q'← reformula (Q, PrefUsr);
11.     Qref ← Remove de Q' conceitos que não existem na RS;
12.     relevanciaP ← Calcula_Relevancia (Qref, RS);
13.     Se relevanciaP >= θ Então
14.     Inicio
15.       mq[1] ← relevanciaP;
16.       Obtem (hasQueryExecTime, hasQuerySubmQty,
          hasQueryAnswQty, hasDataUpdate) de p no ponto corrente;
17.       mq[2] ← hasQueryAnswQty/hasQuerySubmQty;
18.       mq[3] ← hasQueryExecTime /hasQuerySubmQty;
19.       mq[4] ← DataAtual()- hasDataUpdate;
20.       Inclui <p, Qref, mq> em QPR;
21.       Armazena medidas de qualidade (mq) de p no ponto
          corrente;
22.     Fim
23.   Fim
24.   Retorna QPR;
25. Fim
```

Figura 4.16– Algoritmo Medidas_Qualidade

4.4.3. Classificando Pontos com Múltiplos Critérios

Na *SemRouting*, a identificação dos melhores pontos dentre os pontos relevantes para o encaminhamento da consulta é um problema de decisão com múltiplos critérios (*MCDM - Multi-Criteria Decision Making*). A partir do conjunto dos pontos relevantes (identificado com o algoritmo Medidas_Qualidade, Seção 4.4.4.2), todos os critérios de qualidade são utilizados para que se possa estabelecer uma classificação dos pontos, para os quais a consulta deverá ser encaminhada.

Neste trabalho, utilizamos o método SAW (*Simple Additive Weighting*) [Hwang e Yoon, 1981] para análise de decisão com múltiplos critérios. Para sua escolha, foram consideradas as análises obtidas nos trabalhos de Batista (2006), Zhu e Buchmann (2002) que fazem um comparativo entre os métodos SAW (Simple Additive Weighting), AHP (Analytic Hierarchy Process), TOPSIS (Technique for Order Preference by Similarity

to Ideal Solution) e DEA (Data Envelopment Analysis). Diante dos resultados apresentados nos trabalhos, e, considerando o número de pontos e de critérios a serem analisados, fizemos a escolha pelo método SAW por se tratar de um método simples, que pondera sobre os critérios definidos e faz uma classificação baseada na influência de cada critério na análise global da informação.

Para análise global da informação, cada ponto da lista de pontos relevantes terá associado um vetor de qualidade com quatro dimensões, onde cada dimensão representa um critério específico (Relevância, Capacidade de Resposta, Atualidade e Tempo de Execução).

$$VQ(P_i) = (c_1, c_2, c_3, c_4)$$

$$VQ(P_i) = (Relev_{P_iQ}, CRep_{P_i}, Atual_{P_i}, TExec_{P_i})$$

Onde VQ(Pi) é o vetor de qualidade associado ao ponto P_i;

$Relev_{P_iQ}$ é o valor de relevância de P_i em relação a Q;

$CRep_{Pi}$.é a capacidade de resposta de P_i;;

$Atual_{P_i}$ é a atualidade de P_i;

$TExec_{P_i}$ é o tempo médio para execução de consultas em P_i.

Como cada critério foi calculado usando uma unidade própria de medida, é preciso então normalizar os valores de forma a gerar uma medida única e comparável. Dependendo do tipo de critério, se positivo ou negativo, diferentes fórmulas serão aplicadas. Neste trabalho, são considerados critérios positivos a Relevância e a Capacidade de Resposta por serem critérios cujos valores influenciam de forma positiva o resultado da QI, ou seja, quanto maior o valor, mais indicado é o ponto. Em relação aos critérios negativos consideramos a Atualidade e o Tempo de Execução. Neste caso, de forma contrária aos critérios positivos, quanto maior o valor obtido para este critério, maior será a influência negativa no cálculo da QI global, ou seja, menos indicado é o ponto. Segundo o método, os valores calculados para cada critério são normalizados utilizando as seguintes fórmulas [Naumann *et al*. 1999]:

$S_{ml} = \frac{c_{ml} - \min(c_{ml})}{\max(c_{ml}) - \min(c_{ml})}$, para os critérios positivos (*Relevância, Capacidade de Resposta*)

$S_{ml} = \frac{\max(c_{ml}) - c_{ml}}{\max(c_{ml}) - \min(c_{ml})}$, para os critérios negativos (*Atualidade, Tempo de Execução*)

Onde, S_{ml} é o valor normalizado do critério l em relação ao ponto P_m;

c_{ml} é valor original do critério l em relação ao ponto P_m;

max (c_{ml}) é o maior valor calculado para o critério l entre todos os pontos avaliados;

min (c_{ml}) é o menor valor calculado para o critério l entre todos os pontos avaliados.

Depois de normalizado, os valores calculados para S_{ml} estarão no intervalo entre 0 e 1. O método SAW requer também um vetor de pesos associados a cada critério de qualidade W = (w_1, w_2, w_3, w_4) onde cada peso está relacionado ao grau de importância individual correspondente a cada critério. Este vetor terá seus valores definidos de acordo com as preferências do usuário estabelecidas no momento da submissão da consulta. A soma dos pesos (w_i) deve ser igual a 1. Assim, será gerado para cada ponto um valor de qualidade global (GlobalIQ), utilizando a seguinte fórmula:

$$\text{GlobalIQ}_{P_m} = \sum_{i=1}^{4} w_i * S_{mi}$$

O algoritmo apresentado na Figura 4.17 mostra os passos a serem seguidos de forma a obter a classificação dos pontos para encaminhamento da consulta. Inicialmente, o algoritmo recebe como entrada (linhas 3 e 4): *QPR*, um conjunto de tuplas contendo o ponto p para encaminhamento da consulta, a consulta reformulada *Qref* a ser encaminhada para p, as respectivas medidas de qualidade *mq* de *p* calculadas, mas não normalizadas, e as preferências do usuário. Na linha 8, um vetor com os pesos associados a cada critério de qualidade (*relevância*, *capacidade de resposta*, *tempo de resposta*, *atualidade*) é obtido de acordo com as preferências do usuário definidas para a consulta. O índice de cada vetor, *mq* e *PesoC*, está definido de forma associada a cada um dos critérios de qualidade da seguinte forma: *(i=1) relevância*, *(i=2) capacidade de resposta*, *(i=3) tempo de execução* e *(i=4) atualidade*. Inicialmente, para cada ponto *p* pertencente a *QP* (linha 9), cada uma das medidas de qualidade presentes em *mq* será normalizada (linha11). A normalização é realizada para que seja possível gerar uma medida única de qualidade (*GlobalIQ*) por ponto. Para isto, os valores de mínimo e máximo correspondentes à medida em *QP* serão identificados (linhas 13 e 14). Dependendo do critério, se *positivo* ou *negativo* (linhas 15 e 17), diferentes fórmulas serão aplicadas para normalização das medidas (linhas 16

e 18), resultando em um valor armazenado em *Vmedida*. O valor de *Vmedida* é ponderado de acordo com o peso associado ao critério *pesoC[i]*, em seguida, acumulado em *GlobalIQ* a cada *mq[i]* de *p* que tenha sido gerada (linha 17). Os valores de *GlobalIQ* para cada *p* são armazenados no conjunto de pontos relevantes *PR* (linha 21). O conjunto *PR* é ordenado de acordo com o *GlobalIQ* (linha 23). Por fim, o algoritmo retorna o conjunto de pontos relevantes *PR* onde cada linha deste conjunto guarda informações sobre o ponto relevante *p*, consulta reformulada a ser encaminhada para *p* e a medida de *GlobalIQ* (linha 24).

```
1. Algoritmo Classifica_Pontos_Relevantes;
2. Entrada
3.    QPR: conjunto de <p:ponto; Qref: consulta; mq:vetor[1..4] of
      real>;   // ponto relevante, consulta reformulada e medidas de
                // qualidade não normalizadas.
4.    PrefUsr: Preferências do usuário;
5. Saída
6.    PR: conjunto de triplas <p:ponto; Qref: consulta;
      GlobalIQ:real>;
7. Inicio
8.    PesoC ← obtém o conjunto de pesos associados aos critérios de
      acordo com PrefUsr;  // PesoC:vetor[1..4] of real;
9.    Para cada ponto QP.p disponível faça
10.   Inicio
11.      Para i de 1 até 4 faça
12.      Inicio
13.         min ← menor valor de mq[i] em QP;
14.         max ← maior valor de mq[i] em QP;
15.         Se (i=1) ou (i=2) Então //critério positivo
16.            Vmedida←(mq[i]- min)/(max-min);
17.         Se (i=3) ou (i=4) Então //critério negativo
18.            Vmedida←( max- mq[i])/(max-min);
19.         GlobalIQ ← GlobalIQ + (PesoC[i] * Vmedida);
20.      Fim
21.      Inclui <QP.p, QP.Qref, GlobalIQ> em PR;
22.    Fim
23.    Ordena PR por GlobalIQ;
24.    Retorna PR
25. Fim
```

Figura 4.17- Algoritmo Classifica_Pontos_Relevantes

4.4.4. Mecanismo de Parada da SemRouting

Para evitar o encaminhamento de consultas que não atendam ao que foi solicitado na consulta original, ou que o usuário fique esperando por um longo tempo as respostas a sua consulta, foram definidas duas condições de parada do roteamento:

- **Baseada em semântica** – pontos com valor de relevância abaixo de um dado limite semântico não serão candidatos ao encaminhamento da consulta.
- **Baseada em saltos** – a consulta será encaminhada ao longo da rede até alcançar um dado valor de TTL (*Time-To-Live*). Inicialmente, o valor de TTL poderá ser estimado utilizando para isto uma média do número de saltos realizados durante o roteamento nas consultas submetidas anteriormente no sistema. Esta informação poderá ser obtida a partir das informações representadas no modelo descrito na Seção 4.3.1. À medida que a consulta vai sendo encaminhada, o TTL é decrementado de 1 até que atinja o valor zero.

Na *SemRouting*, sempre que uma das condições de parada é alcançada, o conjunto de resultados da consulta armazenados no ponto corrente é devolvido seguindo o caminho inverso até o ponto que encaminhou a consulta. Definimos como conjunto de resultados a união de todos os resultados das consultas retornados pelos vizinhos diretos do ponto corrente.

4.4.5. Seleção dos Melhores Pontos

Nas seções anteriores foram descritos os elementos que compõem a abordagem *SemRouting*. Nesta seção, apresentaremos o algoritmo central (Figura 4.18), de nome semelhante ao da abordagem, que faz a integração destes elementos.

Inicialmente o algoritmo recebe como entrada a consulta *Q*, o valor de *TTL* que delimita o tamanho máximo de saltos para o roteamento, as preferências estabelecidas pelo usuário para a consulta (*PrefUsr*), a referência semântica da consulta original (*RS*) e o valor limite para classificação das consultas consideradas relevantes no sistema (linhas 2 a 6). O valor de retorno do algoritmo (*R*) corresponde ao conjunto de resultados das consultas executadas em cada ponto integrados no ponto corrente (ponto que recebe a consulta para o roteamento) que deverá ser devolvido seguindo o caminho inverso até o ponto que encaminhou a consulta (linha 7). Por simplicidade e por estar fora do escopo deste trabalho, consideramos que a etapa de integração de resultados corresponde à união do conjunto de resultados retornados pelos pontos relevantes armazenados no ponto corrente.

Para cada consulta que chega ao ponto corrente, *R* é inicializado (linha 9), *Q* é executada localmente e seu resultado armazenado em *R* (linha 10). Se o valor de *TTL* for diferente de zero, significa que ainda é possível ao ponto fazer, de acordo com os critérios locais de seleção, o encaminhamento de *Q* para seus vizinhos semânticos diretos (linha 11). Para isto, o algoritmo passa a buscar algumas informações: obtém os pontos candidatos ao roteamento a partir de seus vizinhos diretos (linha 13) e guarda este resultado em *PC*; obtém as medidas de qualidade do conjunto de *PC* identificando os pontos relevantes e guarda o resultado em *QPR* (linha 14). No algoritmo, *QPR* é o conjunto de triplas contendo <*p, Qref, MedidasQualidade*> onde *p* é o ponto relevante, *Qref* a consulta reformulada de acordo com a ontologia de *p*, e *MedidasQualidade* é um vetor contendo o valor dos critérios (*relevância, capacidade de resposta, tempo de execução, atualidade*) associados à *p.* Em seguida, conhecendo as medidas de qualidade dos pontos relevantes (*QPR*) e as preferências do usuário em relação à consulta (*PrefUsr*) é gerada uma classificação para o conjunto de pontos relevantes (*PR*) para o encaminhamento da consulta onde é produzida uma medida única para análise da QI de *p* (linha 15). Assim, *PR* é um conjunto de triplas na forma <*p, Qref, GlobalIQ*> onde *p* é o ponto relevante, *Qref* a consulta reformulada de acordo com a ontologia de *p*, e *GlobalIQ* a medida de qualidade global associada a *p*.

Em *Medidas_Qualidade* (linha 14) os pontos que não possuíam relevância em relação à consulta foram eliminados. Como os critérios são também utilizados de forma classificatória, um limite de qualidade, considerando a mediana do conjunto de valores de *GlobalIQ*, associado a cada ponto, será utilizado para identificar os melhores pontos dentre os pontos classificados para o encaminhamento da consulta (linha 16). A razão para esta escolha é que, dentre as medidas de tendência central, a mediana é a que melhor identifica este conjunto de valores por resultar em valores que se aproximam de faixas. A mediana é capaz de identificar o valor central de uma amostra mesmo que estes valores sejam irregulares, ou seja, discrepantes entre si. Em seguida, para cada ponto p disponível no conjunto de pontos relevantes *PR* e com *GlobalIQ* maior do que o valor da mediana do conjunto (*med*) (linha 17), a *SemRouting* envia mensagem para o ponto *PR.p* com parâmetros para um novo roteamento (linha 19). Ao enviar a mensagem, o valor de *TTL* é decrementado (*TTL-1*). Para cada mensagem encaminhada, o ponto corrente fica aguardando até que o tempo máximo de espera (*MAX_WAIT_TIME*) seja atingido ou até que algum resultado seja retornado (linha 20).

O valor de *MAX_WAIT_TIME* pode ser estimado a partir do valor médio do critério de qualidade *tempo de resposta* representado no modelo de roteamento. Todo resultado retornado pelos pontos relevantes que receberam a consulta será armazenado em *R* (linha 22).

```
1. Algoritmo SemRouting (Q, TTL, PrefUsr, RS, θ): R;
2. Entrada    Q: consulta;
3.            TTL: número de saltos;
4.            PrefUsr: preferências do usuário em relação a Q;
5.            RS: referência semântica;
6.            Θ: threshold de relevância;
7. Saída      R: conjunto de resultados de Q;
8. Inicio
9.     R ← Ø;
10.    R ← executa Q local;
11.    Se (TTL ≠ 0) então
12.    Inicio
13.        PC ← vizinhos semânticos do ponto corrente;
14.        QPR ← Medidas_Qualidade (PC, Q, PrefUsr, RS);
15.        PR ← Classifica_Pontos_Relevantes (QPR, PrefUsr);
16.        med ← calcular mediana de GlobalIQ em PR;
17.        Para cada (PR.p disponível) e (PR.GlobalIQ > med) faça
18.        Inicio
19.            Encaminha mensagem (PR.Qref, TTL-1, PrefUsr, RS, θ)
               para execução de SemRouting em PR.p, e
20.            Aguarda  (resultado de PR.p) ou
                        (tempo de espera >(TTL-1)*MAX_WAIT_TIME)
21.            Se (existe resultado de PR.p) então
22.                Inclui o resultado da execução em R;
23.        Fim;
24.    Fim;
25.    Retorna R;
26. Fim.
```

Figura 4.18 - Algoritmo *SemRouting*

Finalizado o tempo de espera de encaminhamento das mensagens ou caso todos os pontos em *PR* tenham retornado seus respectivos resultados, o algoritmo retornará *R* para o ponto do qual recebeu a consulta (linha 25).

4.5. Exemplo de Uso

Para ilustrar toda a abordagem *SemRouting*, considere um cenário do domínio de Educação composto por dez pontos semanticamente relacionados (Figura 4.19).

A ontologia de domínio utilizada é a UnivCsCMO.owl, o valor de TTL é igual a 3 e o limite semântico 0,5 (Θ). O valor de TTL estabelece uma das condições de parada

do roteamento. Cada encaminhamento da consulta de um ponto origem para um ponto destino indica um salto. Assumimos no exemplo que, após reformuladas, as consultas passam pelo processo de poda semântica para preservar a semântica da consulta original.

Neste cenário, um usuário submete uma consulta *Q = {Professor, Magazine}* no ponto P1 e, com o objetivo de alcançar mais resultados, seleciona todas as variáveis de enriquecimento para a consulta (*generalize, specialize, approximate, compose*) e define a classificação em relação aos critérios de qualidade estabelecendo assim os elementos contextuais associados à consulta. Para cada critério, consideramos os seguintes pesos: preservação igual a 0.4, enriquecimento igual a 0.2, relevância igual a 0.6, capacidade de resposta igual a 0.16, atualidade igual a 0.13 e tempo de execução igual a 0.11. De acordo com os elementos contextuais que definem o modo de reformulação da consulta, a referência semântica de Q é gerada em P1 da seguinte forma:

> ***RS***=*{<**Professor**, Professor, 1.0>, <**Professor**, Faculty, 0.8>, <**Professor**, AssociateProfessor, 0.8>, <**Professor**, FullProfessor, 0.8>, <**Professor**, AssistantProfessor, 0.8>, <**Professor**, Lecturer, 0.7>, <**Professor**, PostDoc, 0.7>, <**Magazine**, Magazine, 1.0>, <**Magazine**, Periodical, 0.8>, <**Magazine**, Journal, 0.7>, <**Magazine**, Article, 0.3>}*

Em P1, *Q* é expandida em *Q1={Professor, Magazine, Periodical}*, executada localmente e seu resultado armazenado no próprio ponto. Executada a consulta, P1 passa a buscar os melhores vizinhos para o encaminhamento da consulta. Assim, P1 obtém seus vizinhos semânticos (P2, P3 e P4) e, para cada vizinho, denominado daqui por diante de ponto candidato, reformula Q de acordo com *PrefUsr* e passa a gerar as medidas de qualidade de cada ponto candidato de acordo com as métricas definidas na Seção 4.4.2. Para o cálculo da relevância dos pontos candidatos são consideradas as reformulações obtidas do ponto origem P1 e de cada um dos pontos destinos (P2, P3, P4). As reformulações são geradas a partir do conjunto de correspondências semânticas existentes entre as ontologias dos pares de ponto (origem, destino). Neste cenário, os valores de relevância são calculados da seguinte forma:

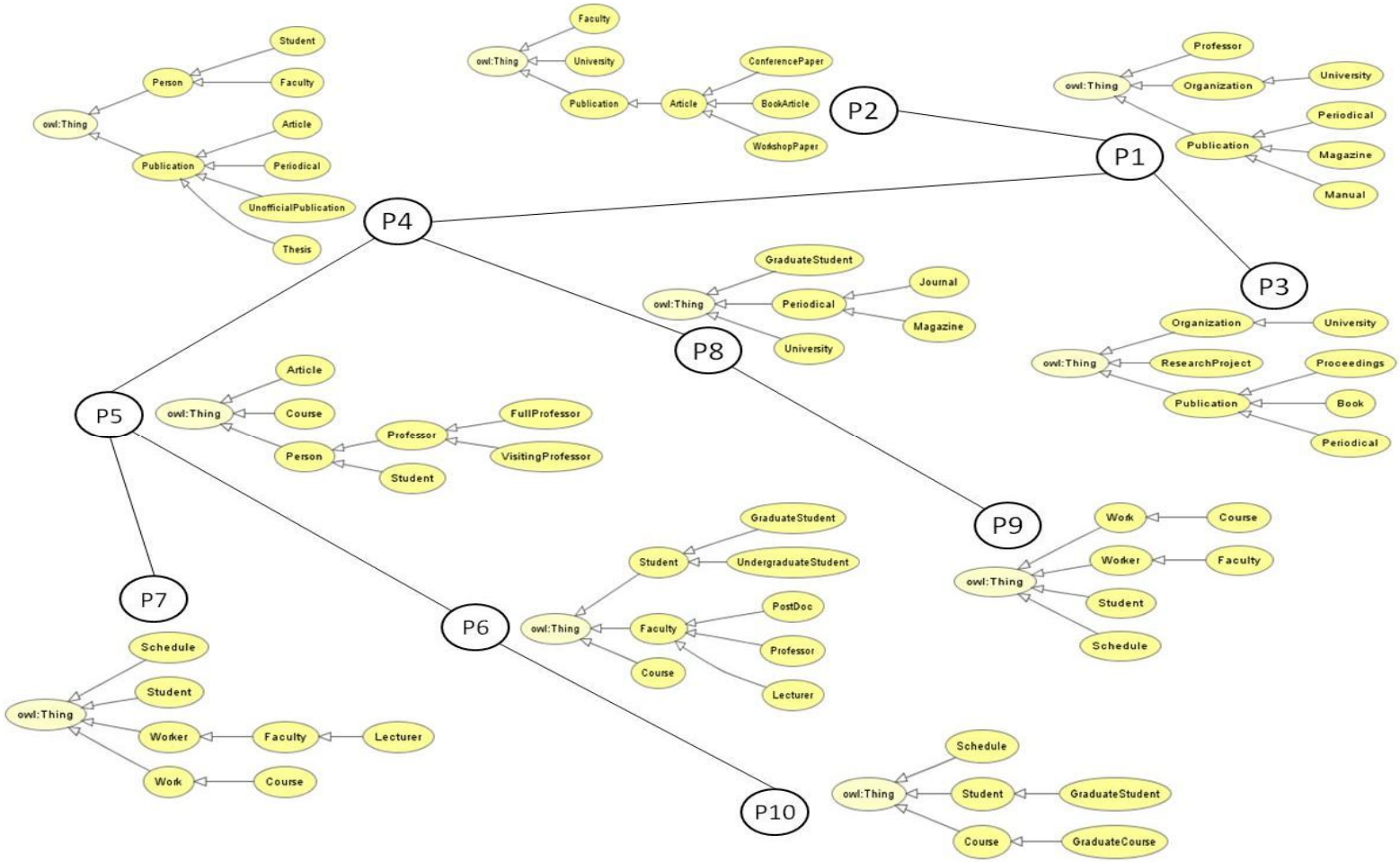


Figura 4.19– Cenário Exemplo

- **De P1 para P2:** De acordo com o alinhamento existente entre P1 e P2 (Figura 4.20), Q é reformulada em Q12={Faculty, Article}.

Ontology 1	Correspondence	Ontology 2	Measure
Magazine	isWholeOf	Article	0,3
Manual	isSubConceptOf	Publication	0,8
Manual	isCloseTo	Article	0,7
Organization	isSuperConceptOf	University	0,8
Organization.orgName	isEquivalentTo	University.orgName	1
Periodical	isSubConceptOf	Publication	0,8
Periodical	isCloseTo	Article	0,7
Professor	isSubConceptOf	Faculty	0,8
Publication	isEquivalentTo	Publication	1
Publication	isSuperConceptOf	Article	0,8
Publication.publicationAuthor	isEquivalentTo	Article.publicationA...	1
Publication.publicationDate	isEquivalentTo	Publication.publicati...	1
Publication.pubTitle	isEquivalentTo	Publication.pubTitle	1
University	isEquivalentTo	University	1

Figura 4.20- Alinhamento entre as ontologias de P1 e P2

Assim, para obter a relevância de P2 em relação a Q são calculadas as seguintes medidas:

- **Preservação da Consulta**

$$PC_{Q_{12}P_2} = \frac{|T_{eq}|}{|T_{orig}|} = \frac{0}{2} = 0$$

- **Enriquecimento da Consulta**

$$EC_{Q_{12}P_2} = \frac{\sum_{k=1}^{nto}(\sum_{l=1}^{nte_k} RS\_sim(to_k, te_l))/nte_k}{nto} = \frac{(0.8/1) + (0.3/1)}{2} = 0.55$$

- **Relevância**

$$Relev_{P_2Q} = PC_{QP_2} + EC_{QP_2} = 0 + 0.55 = 0.55$$

- **De P1 para P3:** De acordo com o alinhamento existente entre P1 e P3 (Figura 4.21**) Q é** reformulada em Q13={ Lecturer }

Ontology 1	Correspondence	Ontology 2	Measure
Manual	isSubConceptOf	Publication	0,8
Manual	isCloseTo	Book	0,7
Manual	isCloseTo	Proceedings	0,7
Organization	isEquivalentTo	Organization	1
Organization	isSuperConceptOf	University	0,8
Periodical	isSubConceptOf	Publication	0,8
Periodical	isCloseTo	Book	0,7
Periodical	isCloseTo	Proceedings	0,7
Professor	isCloseTo	Lecturer	0,7
Publication	isEquivalentTo	Publication	1
Publication	isSuperConceptOf	Book	0,8
Publication	isSuperConceptOf	Proceedings	0,8
University	isEquivalentTo	University	1
University	isSubConceptOf	Organization	0,8

Figura 4.21 - Alinhamento entre as ontologias de P1 e P3

Assim, para Q13={ Lecturer}, temos os seguintes cálculos:

- **Preservação da Consulta**

$$PC_{Q_{12}P_2} = \frac{|T_{eq}|}{|T_{orig}|} = \frac{0}{2} = 0$$

- **Enriquecimento da Consulta**

$$EC_{Q_{12}P_2} = \frac{\sum_{k=1}^{nto}(\sum_{l=1}^{nte_k} RS\_sim(to_k, te_l))/nte_k}{nto} = \frac{0.7}{2} = 0.35$$

- **Relevância**

$$Relev_{P_2Q} = PC_{QP_2} + EC_{QP_2} = 0 + 0.35 = 0.35$$

- **De P1 para P4:** De acordo com o alinhamento existente entre P1 e P4 (Figura 4.22), Q é reformulada em Q14={Faculty,Periodical, Article}.

Ontology 1	Correspondence	Ontology 2	Measure
Magazine	isSubConceptOf	Periodical	0,8
Magazine	isWholeOf	Article	0,3
Manual	isSubConceptOf	Publication	0,8
Manual	isCloseTo	UnofficialPublication	0,7
Manual	isCloseTo	Thesis	0,7
Manual	isCloseTo	Periodical	0,7
Manual	isCloseTo	Article	0,7
Organization	isDisjointWith	Person	0
Periodical	isEquivalentTo	Periodical	1
Periodical	isSubConceptOf	Publication	0,8
Periodical	isCloseTo	UnofficialPublication	0,7
Periodical	isCloseTo	Thesis	0,7
Periodical	isCloseTo	Article	0,7
Professor	isSubConceptOf	Faculty	0,8
Professor.emailAddress	isEquivalentTo	Person.emailAddress	1
Publication	isEquivalentTo	Publication	1
Publication	isSuperConceptOf	UnofficialPublication	0,8
Publication	isSuperConceptOf	Thesis	0,8
Publication	isSuperConceptOf	Periodical	0,8
Publication	isSuperConceptOf	Article	0,8
Publication	isDisjointWith	Person	0
Publication.publicationAuthor	isEquivalentTo	Publication.publicati...	1
Publication.publicationDate	isEquivalentTo	Publication.publicati...	1
Publication.pubTitle	isEquivalentTo	Publication.pubTitle	1

Figura 4.22- Alinhamento entre as ontologias de P1 e P4

Assim, para Q14={Faculty,Periodical, Article} temos os seguintes cálculos:

- **Preservação da Consulta**

$$PC_{Q_{12}P_2} = \frac{|T_{eq}|}{|T_{orig}|} = \frac{0}{2} = 0$$

- **Enriquecimento da Consulta**

$$EC_{Q_{13}P_3} = \frac{\sum_{k=1}^{nto}(\sum_{l=1}^{nte_k} RS\_sim(to_k, te_l))/nte_k}{nto} = \frac{(0.8/1) + ((0.8 + 0.3)/2)}{2} = 0{,}67$$

- **Relevância**

$$Relev_{P_2Q} = PC_{QP_2} + EC_{QP_2} = 0 + 0.67 = 0.67$$

A Tabela 4.1 mostra um resumo dos resultados das reformulações de Q em P1 de acordo com a ontologia de cada um dos pontos candidatos, os valores de preservação (PC), o enriquecimento (EC) e a relevância.

Tabela 4.1 - Resumo das Reformulações e das Medidas obtidas em P1

Ponto		Consulta		PC	EC	Relevância
Origem	**Destino**	**Recebida**	**Reformulada**			
P1	P2	{Professor, Magazine}	{Faculty, Article}	0	0,55	0,55
P1	P3	{Professor, Magazine}	{Periodical}	0	0,35	0,35
P1	P4	{Professor, Magazine}	{Faculty,Periodical, Article}	0	0,67	0,67

Analisar medidas de qualidade em relação aos pontos candidatos no *SemRouting* é calcular a relevância e eliminar pontos que estejam abaixo de um dado limite semântico. No caso, P3 será eliminado por estar abaixo de 0.5. Além disto, significa classificar os pontos relevantes de acordo com os demais critérios de qualidade definidos para o processo (atualidade, tempo de execução e capacidade de resposta). Para cálculo destes critérios, cada ponto mantém um conjunto de informações relacionadas à sua vizinhança semântica.

A Tabela 4.2 mostra o conhecimento do ponto em relação a seus vizinhos para cada critério de qualidade utilizado pelo *SemRouting*: preservação da consulta (PC), enriquecimento da consulta (EC), capacidade de resposta (CR), atualidade (AT), tempo de execução (TE). Nesta tabela estão apresentados os valores calculados e normalizados de cada critério. Como o objetivo desta seção é exemplificar o processo, os valores de atualidade, tempo de execução e capacidade de resposta foram

estimados. Para o critério de relevância, foram utilizados os valores reais de cálculo para cada reformulação (Tabela 4.1).

Tabela 4.2- Medidas Normalizadas e não Normalizadas dos vizinhos semânticos de P1

	Não Normalizados					Normalizados				
Ponto Candidato	**PC**	**EC**	**CR**	**AT**	**TE**	**PC**	**EC**	**CR**	**AT**	**TE**
P2	0	0,55	0,90	376,92	7,78	0,00	0,00	1,00	0,00	0,00
P4	0	0,67	0,66	149,92	5,90	0,00	1,00	0,00	1,00	1,00

Após normalizadas, as medidas são utilizadas para gerar o GlobalIQ. Para o cálculo do GlobalIQ, os valores normalizados são ponderados de acordo com o contexto estabelecido pelo usuário a partir das preferências associadas à configuração da consulta (*Query_Configuration*). Em seguida, a mediana de GlobalIQ é gerada e os pontos com valor maior ou igual à mediana são selecionados para encaminhamento da consulta. A Tabela 4.3 mostra o valor da mediana e os valores de GlobalIQ produzidos para cada ponto candidato. Desta forma, Q será reformulada e encaminhada para os pontos P2 e P4.

Tabela 4.3- Medidas associadas aos pontos vizinhos de P1 no instante da seleção de pontos

Ponto Candidato	Global_IQ	Relevante para o Roteamento
P2	0,16	**Não**
P4	0,44	**Sim**
Mediana	**0,30**	

Para cada ponto que recebe uma consulta, todo o processo anteriormente descrito se repete. Para P2, como o valor de Global_IQ é menor do que o valor da mediana a consulta não será encaminhada.

Em P4, com TTL=2, a consulta poderá ser encaminhada ao ponto P5 ou P8 dependendo do valor de relevância do ponto. A Tabela 4.4 mostra os valores de PC, EC e Relevância de acordo com as reformulações realizadas entre os pontos P5 e P8 durante o roteamento.

Tabela 4.4 - Resumo das Reformulações e das Medidas obtidas em P4

Ponto		Consulta		PC	EC	Relevância
Origem	Destino	Recebida	Reformulada			
P4	P5	{Faculty, Periodical, Article}	{Professor, Article}	0,5	0,15	0,65
P4	P8	{Faculty, Periodical, Article}	{Magazine, Periodical, Journal}	0,5	0,37	0,87

Calculada a relevância, o valor de GlobalIQ é gerado a partir das medidas de qualidade de P5 e P8. A Tabela 4.5 mostra os valores não normalizados e normalizados das medidas de qualidade.

Tabela 4.5- Medidas Normalizadas e não Normalizadas dos vizinhos semânticos de P4

	Não Normalizados					Normalizados				
Ponto Candidato	PC	EC	CR	AT	TE	PC	EC	CR	AT	TE
P5	0,5	0,15	200	200	0,93	0,00	0,00	1,00	1,00	0,00
P8	0,5	0,37	0,7	300	0,5	0,00	1,00	0,00	0,00	1,00

Após a normalização, o valor de GlobalIQ é calculado para que possa ser utilizado na seleção dos melhores pontos em P4. De acordo com Tabela 4.6, a consulta recebida por P4 será reformulada e encaminhada para os pontos P5 e P8.

Tabela 4.6- Medidas associadas aos pontos vizinhos de P4 no instante da seleção de pontos

Ponto Candidato	Global_IQ	Relevantes para o Roteamento
P5	0,30	Sim
P8	0,31	Sim
Mediana	0,30	

Em P5, com TTL igual a 1, uma nova escolha precisa ser feita com relação aos vizinhos de P5, no caso, P6 e P7. A Tabela 4.7 mostra os valores de relevância calculados para P6 e P7.

Tabela 4.7- Resumo das Reformulações e das Medidas obtidas em P5

Ponto		Consulta		PC	EC	Relevância
Origem	Destino	Recebida	Reformulada			
P5	P6	{Professor, Article}	{ Professor, Faculty, PostDoc, Lecturer}	0,5	0,36	0,86
P5	P7	{Professor, Article}	{Faculty, Lecturer}	0	0,37	0,37

Como podemos observar, o valor de relevância de P7 está abaixo do limite semântico para encaminhamento da consulta (que é de 0.5). Assim, P7 não receberá a consulta reformulada. Neste caso, a consulta será reformulada e encaminhada apenas para o ponto P6 sem que seja preciso realizar uma avaliação dos demais critérios de P6. Esta situação ocorre porque não existe um processo de escolha a ser feito. Agora em P6, com TTL igual a zero, apesar da existência do ponto vizinho P10, a consulta não será mais reformulada e encaminhada. Assim, P6 executa a consulta recebida localmente e o seu resultado é enviado no caminho inverso ao ponto que encaminhou a consulta.

Em P8, com TTL igual a 1, a consulta é reformulada e encaminhada para o único vizinho de P8, no caso, P9, dependendo da relevância de P9 (Tabela 4.8). Como a relevância de P9 está abaixo do limite semântico de 0,5 o encaminhamento da consulta nesta rota é interrompido.

Tabela 4.8- Resumo das Reformulações e das Medidas obtidas em P4

Ponto		Consulta		PC	EC	Relevância
Origem	**Destino**	**Recebida**	**Reformulada**			
P8	P9	{Magazine, Periodical, Journal}	{Article}	0	0,15	0,15

Durante o roteamento da consulta submetida em P1, três rotas foram percorridas pela consulta original: **P1-P2**; **P1-P4-P5-P6** e **P1-P4-P8**. Neste cenário com 10 pontos, 6 foram escolhidos. Importante observar que as decisões de escolha para encaminhamento da consulta sempre estiveram baseadas no conhecimento local do ponto em relação aos seus vizinhos. Assim, a cada consulta a ser encaminhada, o ponto realiza a seleção dos melhores caminhos a serem seguidos analisando o contexto corrente definido a partir dos elementos contextuais associados ao ponto.

Neste exemplo, ilustramos diversas situações previstas na *SemRouting*: a preservação da consulta, a relevância do ponto, as condições de parada e a atividade de seleção de pontos. É importante destacar que todos os elementos contextuais identificados na atividade de seleção (medidas de qualidade e preferências do usuário) são armazenados e mantidos por cada ponto em relação aos seus vizinhos semânticos. O objetivo é expandir o conhecimento local do ponto para que a ação de seleção de pontos durante o roteamento possa ser realizada de forma mais específica. Além disto, a definição dos valores de TTL e tempo médio de execução de consultas poderão

ser obtidos futuramente a partir deste conjunto de informações armazenadas em cada ponto do sistema.

4.6. Análise Comparativa

No Capítulo 3, algumas estratégias de roteamento foram apresentadas e analisadas de acordo com as questões de investigação desta tese. Nesta seção, fazemos uma análise comparativa destas estratégias com a nossa abordagem. O Quadro 4.1 resume as principais características apontadas em cada estratégia, incluindo a *SemRouting*.

Assim como os demais trabalhos apresentados, o *SemRouting* procura eliminar pontos que não possam contribuir com a consulta. Como no trabalho do Humboldt Peer, também eliminamos pontos que ofereçam baixa contribuição em relação aos resultados esperados. Para isto, estabelecemos um limite semântico baseado na definição de diferentes tipos de correspondências semânticas [Souza *et al.* 2009]. Considerando a dinamicidade do ambiente, a *SemRouting* constrói as rotas a partir da análise do contexto que cerca a atividade de seleção de pontos. Neste caso, as preferências do usuário, as medidas de qualidade e a disponibilidade do ponto são consideradas como fatores que influenciam diretamente a escolha dos pontos para o encaminhamento da consulta. No ESTEEM e H-Link as informações contextuais também são utilizadas na seleção de pontos mas não envolvem medidas de qualidade. No caso do ESTEEM a qualidade é uma medida que pode ser incorporada aos resultados da consulta como informação de proveniência.

Em relação à preservação semântica da consulta original, ao longo do roteamento, o *SemRouting* define uma RS que é utilizada para remover termos que não estejam semanticamente relacionados com a consulta original. Além disto, a RS pode ser utilizada para identificar o valor semântico associado a cada termo na consulta reformulada e termos originais recuperados em reformulações futuras. Entre todas as estratégias, apenas o GrouPeer oferece uma estratégia, diferente da nossa, que analisa a consulta reformulada de acordo com a consulta original. Enquanto o GrouPeer faz o encaminhamento da consulta escolhendo a melhor reformulação, o SemRouting faz o encaminhamento da consulta buscando preservar a semântica da consulta original a cada reformulação.

Quadro 4.1 - Quadro Comparativo entre as Abordagens Apresentadas e a *SemRouting*

Estratégia	Representação dos Esquemas	Identificação de Pontos	Seleção de Pontos		Preservação da Consulta	Mecanismo de Parada
			Informação Contextual	Critério de Qualidade		
MCS-SP	sGraph	Mapeamentos e índices semânticos	Preferência do usuário	Não	Não	TTL
Ontozilla	Ontologias	Mapeamentos e índices semânticos	Não	Não	Não	TTL
SUNRISE	Ontologia, relacional, XML	Mapeamentos e índices semânticos	Preferência do usuário	Não	Não	Meta de satisfação; TTL
H-Link	Ontologia	Mapeamentos semânticos;	Contexto do ponto	Não	Não	Baseado em crédito
ESTEEM	Ontologia	Mapeamentos semânticos	CDT (Context Dimension Tree)	Não	Não	TTL
Humboldt Peer	Relacional	Mapeamentos semânticos	Não	Cobertura e Densidade	Não	Baseado em estimativa (*Weight* e *Greedy*)
Groupeer	Relacional	Mapeamentos semânticos	Não	Degradação da consulta	Manutenção de conceitos perdidos	TTL
SemRouting	**Ontologia**	**Correspondências Semânticas**	**Preferências do usuário; medidas de qualidade; disponibilidade do ponto**	**Relevância (preservação da consulta + enriquecimento); Capacidade de resposta; Tempo de Execução; Atualidade**	**Poda Semântica de acordo com uma Referência Semântica definida**	**TTL; limite semântico (avaliação da relevância)**

Quanto ao mecanismo de parada, consideramos que o roteamento deve levar em conta critérios semânticos que evitem o encaminhamento de consultas com alta perda semântica. Dentre os trabalhos apresentados, o SUNRISE parece ter um mecanismo similar ao nosso. Entretanto, nos trabalhos do SUNRISE não foi definida a função de cálculo da meta de satisfação para uma melhor comparação.

4.7. Considerações

A abordagem *SemRouting* objetiva expandir o conhecimento local do ponto, combinando o uso de informações semânticas e de qualidade da informação (QI) para seleção de pontos que possam contribuir com melhores resultados de acordo com a consulta original submetida pelo usuário. Para isto, critérios de qualidade foram definidos e suas métricas especificadas para serem usadas na avaliação da QI do ponto. A partir das medidas obtidas em cada critério, foi gerada uma medida de avaliação global para cada ponto.

Verificamos que o uso da semântica na *SemRouting* acontece em diferentes etapas da abordagem: no uso das correspondências semânticas, no cálculo da relevância, no mecanismo de parada do roteamento e no uso do contexto durante a atividade de seleção de pontos. Em relação ao contexto, destacamos que ele é definido a partir das preferências do usuário, da disponibilidade e das medidas de qualidade associadas ao ponto. Discutimos que o contexto armazenado durante a atividade de seleção dos pontos poderá ser utilizado posteriormente para definição de constantes utilizadas no sistema, a exemplo do tempo médio gasto pelos pontos durante a execução de consultas e o tamanho médio alcançado pelas rotas. No primeiro caso, a média do tempo de resposta pode ser utilizada para definir a constante MAX_WAIT_TIME para evitar a espera prolongada por resultados após o encaminhamento da consulta para um ponto vizinho. Em relação ao tamanho médio das rotas, esta informação pode ser usada para estimar o valor inicial de TTL. O próximo capítulo descreve os aspectos de implementação e os resultados obtidos com os experimentos relacionados à *SemRouting*.

Capítulo 5

Implementação e Experimentos

Neste capítulo apresentamos aspectos relacionados à implementação e aos experimentos da abordagem *SemRouting*. Para isto organizamos o capítulo da seguinte forma: a Seção 5.1 descreve o PDMS SPEED utilizado como ambiente para realização dos experimentos; a Seção 5.2 discute a implementação do *SemRouting*; a Seção 5.3 mostra a avaliação experimental realizada; a Seção 5.4 apresenta as considerações finais do capítulo.

5.1. O Sistema SPEED

O SPEED (*Semantic PEEr-to-Peer Data Management System*) é um PDMS baseado em semântica que tem como objetivo prover soluções para problemas de gerenciamento de dados em um ambiente P2P, como conectividade, mapeamentos e processamento de consultas [Pires *et al.* 2011]. No SPEED, existem três tipos de pontos: **ponto de dados**, **ponto de integração** e **ponto semântico**. Um **ponto de dados** representa uma fonte de dados (estruturada ou semi-estruturada) que compartilha dados com outros pontos no sistema. Ontologias são utilizadas para representar os esquemas das fontes em cada ponto de dados. **Pontos de Integração** também são pontos de dados com maior poder computacional e funcionalidade adicional de integração com outros pontos de dados. Um ***Cluster* Semântico** representa o agrupamento de pontos de dados com similaridade semântica em torno de um ponto de integração. **Pontos Semânticos** identificam os *clusters* dentro de uma comunidade. A Comunidade Semântica consiste no agrupamento de *clusters*

pertencentes a um mesmo domínio de conhecimento (por exemplo, “educação”) [Pires *et al*. 2011]

O SPEED utiliza na sua arquitetura diferentes tipos de topologias de rede: estruturada (DHT - *Distributed Hash Table*), não estruturada e super-ponto. A DHT é responsável por organizar os pontos semânticos e permitir que um ponto, ao entrar no sistema, identifique outros pontos com características comuns em uma comunidade semântica. Dentro de uma comunidade os pontos estão organizados em uma topologia de rede de super-pontos. Entre os super-pontos o SPEED tem definida uma rede não estruturada. A Figura 5.1 ilustra a arquitetura geral do SPEED.

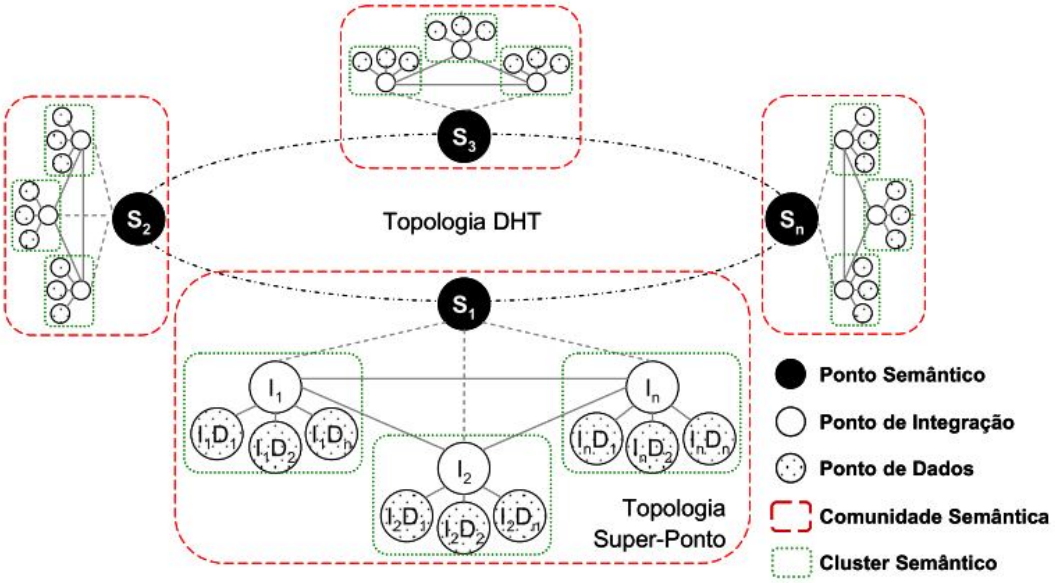


Figura 5.1- Arquitetura do SPEED [Pires 2009]

O SPEED utiliza os módulos *SemMatcher* [Pires *et al.* 2009] e *SemRef* [Souza *et al.* 2009] descritos no Capítulo 4. O *SemMatcher* é utilizado, no processo de formação da rede, para produzir as correspondências e gerar uma medida de similaridade semântica entre ontologias de pontos vizinhos. A partir desta medida, utilizando um *threshold* de *cluster* e outro de vizinhança, o sistema define a composição dos *clusters* semânticos e a vizinhança entre os pontos de integração. A interface para submissão da consulta está integrada ao reformulador de consultas *SemRef* [Souza *et al.* 2009]. O SPEED adota um modelo de programação distribuído baseado em troca de mensagens. Antes deste trabalho, o sistema não possuía uma estratégia para o roteamento da consulta.

5.2. Implementação da Abordagem SemRouting

Os algoritmos que definem a abordagem *SemRouting* foram implementados em *Java* e integrados ao módulo de consulta do SPEED. As informações semânticas e de qualidade utilizadas pelo *SemRouting* foram armazenadas em um SGBD relacional (*Oracle* 11g). Considerando que estamos trabalhando com regras simples nas decisões associadas à seleção dos pontos e que não estamos inferindo e produzindo conhecimento a partir das informações armazenadas, optamos por fazer o gerenciamento dos dados em um SGBD Relacional, pela sua capacidade de armazenamento e facilidade de gerenciamento.

Para submissão da consulta, o SPEED dispõe de uma interface (Figura 5.2) a partir da qual é possível: (i) formular a consulta de acordo com conceitos presentes na ontologia usando SPARQL ou DL e (ii) definir as preferências do usuário em relação ao enriquecimento da consulta na reformulação. Para o roteamento, inserimos mais uma preferência (posicionada na interface no quadro *Quality Criteria*) que estabelece os pesos relacionados aos critérios de qualidade. Como podemos observar, existe um valor previamente definido pelo administrador do sistema que define uma prioridade para classificação dos pesos. Este valor só será alterado caso seja de interesse do usuário alterar esta classificação.

Atualmente, o módulo de consulta não implementa operações de restrição às consultas nem em SPARQL nem em DL. Neste caso, é importante destacar que a ausência desta funcionalidade não tem impacto nos algoritmos que compõem a abordagem. No *SemRouting*, a preservação da consulta original e a seleção de pontos, ao longo do roteamento, estão relacionadas ao conjunto de termos (conceitos ou propriedades da ontologia) usados na formulação da consulta que, por definição, representam sua semântica.

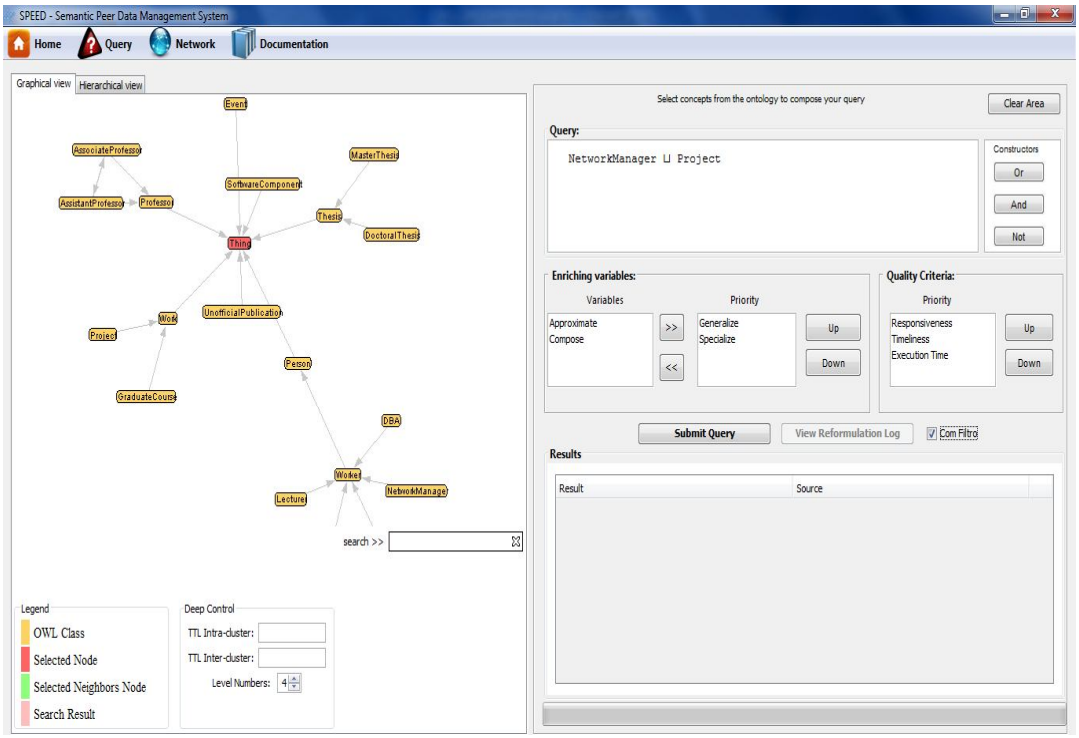


Figura 5.2– Interface de Consulta do SPEED

5.3. Experimentos

Como a abordagem *SemRouting* pressupõe um ambiente de rede pura, consideramos o roteamento da consulta no nível dos pontos de integração do SPEED. Ontologias, pertencentes ao domínio de Educação, representam o esquema de cada ponto. Cada ontologia tem em média 25 conceitos. Os experimentos foram executados em um computador com processador Core I5 e 8GB de memória.

Os experimentos foram realizados para verificação das hipóteses apresentadas no Capítulo 1. Ao todo foram realizados dois experimentos. As próximas subseções apresentam cada um deles de acordo com a seguinte organização:

- Cenário – apresenta características e dados de configuração do ambiente.
- Descrição do Experimento – relata a sequência de atividades realizadas para produção dos resultados.
- Análise dos Resultados – mostra e analisa os resultados obtidos.
- Verificação das Hipóteses – verifica as hipóteses de acordo com os resultados obtidos.

5.3.1. Experimento 1

Este experimento, também apresentado em Freire *et al*. (2014), verificou a hipótese H2 considerando dois aspectos: (i) a preservação semântica da consulta original ao longo do roteamento; e (ii) o grau de preservação semântica desta consulta. O primeiro aspecto foi avaliado utilizando o coeficiente de *Jaccard* [Han *et al.* 2012], o segundo, a partir da análise da precisão dos resultados obtidos em duas diferentes configurações do experimento [Rijsbergen, 1979]. Esta medida será melhor definida na etapa de Descrição do Experimento.

- **Cenário**:

O cenário corrente constitui uma rede formada por um conjunto de cinco pontos (P2178, P2278, P2378, P2478, P2578) conectados de forma sequencial (Figura 5.3). Para análise do valor mínimo de relevância do ponto, para o qual a consulta poderá ser encaminhada, utilizamos o valor de *threshold* de 0.5 (ver Apêndice A).

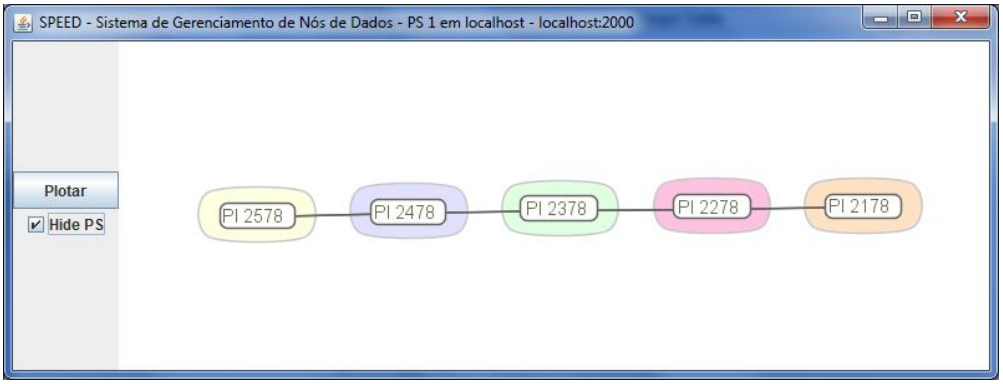


Figura 5.3 – Cenário de Distribuição dos Pontos para o Experimento 1

- **Descrição do Experimento**:

Inicialmente, uma consulta *Q={GraduateStudent* ∪ *Workshop}* foi submetida no ponto P2178 e todas as variáveis de enriquecimento foram selecionadas (*Approximate, Generalization, Specialization, Compose*). Considerando a consulta submetida, o experimento foi executado de duas formas diferentes. Na primeira, denominada de *QRwithout_Pruning*, a consulta foi reformulada sem considerar se os termos expandidos eram semanticamente relacionados com os termos da consulta original, ou seja, sem fazer uso do mecanismo de poda semântica. Na segunda execução,

chamada de *QRpruning*, a mesma consulta foi roteada sempre considerando a consulta original, ou seja, utilizando o mecanismo de poda.

Em *QRwithout_Pruning*, a consulta foi roteada de forma direta utilizando *flooding* [Gnutella 2014]. Neste caso, não foram consideradas medidas para avaliação da relevância de cada ponto. Os pontos eram consultados sempre que fosse possível produzir uma reformulação a partir da consulta corrente.

Em *QRpruning*, de acordo com o contexto de submissão da consulta (preferências do usuário em relação ao modo de enriquecimento e semântica da consulta original), a referência semântica (RS) foi gerada com a seguinte composição de termos e associação de pesos:

RS={<GraduateStudent, GraduateStudent, 1.0>, <GraduateStudent, Student, 0.8>, <GraduateStudent, PhdStudent, 0.8>, <GraduateStudent, MasterStudent, 0.8>, <GraduateStudent, ResearchProject, 0.3>, <GraduateStudent, Course, 0.3>, <Workshop, Workshop, 1.0>, <Workshop, Event, 0.8>, <Workshop, Exhibtion, 0.7>, <Workshop, Conference, 0.7>, <Workshop, Meeting, 0.7>, <Workshop, Lecture, 0.7>}

Usando a RS, a cada encaminhamento da consulta entre o ponto origem e o ponto destino, o valor de relevância do ponto foi calculado para que fosse possível identificar se ele poderia ou não responder a consulta. Na Figura 5.4, apresentamos um dos *logs* de reformulação (entre os pontos P2378 e P2478) produzidos ao longo do roteamento. Na figura, o campo *Reformulated Query* corresponde à consulta que foi reformulada, *Reformulated Query After Prunning* mostra a consulta produzida após ter sido aplicada a poda semântica e *Pruned Terms indicam* os termos eliminados da consulta reformulada que não possuem relação semântica direta com os termos da consulta original. Os valores de preservação (*QPreserv*), enriquecimento (*QEnrich*) e da semântica da consulta (*Query Semantic Value - QSV*) aparecem calculados na mesma figura. O valor de QSV é utilizado para indicar a relevância do ponto em relação à consulta reformulada.

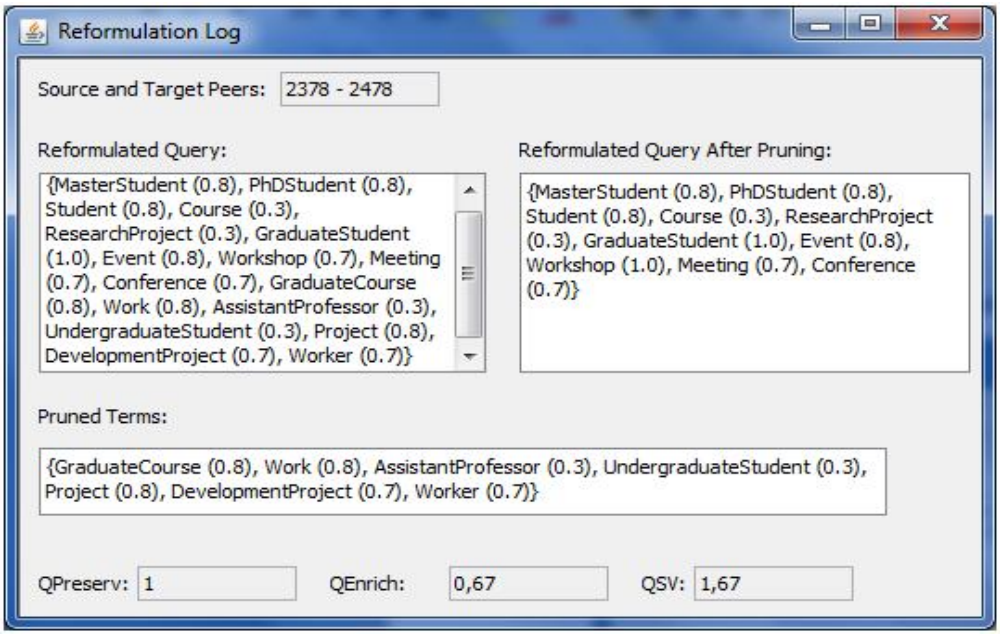


Figura 5.4– Log de Reformulação entre os pontos P2378 e P2478

A Figura 5.5 mostra uma síntese dos resultados obtidos a cada reformulação entre os pontos P2178 e P2278; P2278 eP2378; P2378 e P2478; P2478 e P2578.

Source Peer - Target Peer	...	Query Reformulated	...	QPreserv	QEnrich	QVS
2178 - 2278	...	{Student (0.8), GraduateStudent (1.0), MasterStudent (0.8), PhDStudent (0.8), Lecture (0.7), Conference (0.7), Event (0.8)}		0,5	0,77	1,27
2278 - 2378	...	{Student (0.8), MasterStudent (0.8), PhDStudent (0.8), GraduateStudent (1.0), Course (0.3), ResearchProject (0.3), Meeting (0.7), Conference (0.7), Event (0.8), Lecture (0.7)}	...	0,5	0,66	1,16
2378 - 2478	...	{Student (0.8), GraduateStudent (1.0), Course (0.3), ResearchProject (0.3), MasterStudent (0.8), PhDStudent (0.8), Event (0.8), Meeting (0.7), Workshop (1.0), Conference (0.7)}	...	1	0,67	1,67
2478 - 2578	...	{GraduateStudent (1.0), MasterStudent (0.8), PhDStudent (0.8), Student (0.8), Course (0.3), ResearchProject (0.3)}	...	0,5	0,3	0,8

Figura 5.5– Síntese das Reformulações de Consultas do ponto P2178 à P2578.

- **Análise dos Resultados**:

Para geração dos dados que avaliam a preservação semântica da consulta ao longo do roteamento, consideramos que a RS (Referencia Semântica) seria utilizada como um referencial dos termos esperados para as consultas reformuladas. Como a RS é definida a partir da ontologia de domínio, é possível identificar todo o conjunto de termos que poderiam ser expandidos de acordo com cada termo da consulta original. Utilizando o coeficiente de *Jaccard*, avaliamos a similaridade de cada reformulação produzida, em relação à RS, nas execuções do roteamento *QRwithout_Pruning* e *QRpruning*. Assim, dado QTermos o conjunto de termos obtidos na reformulação de uma consulta Q e RStermos os conjunto de termos na RS de Q, dizemos que o coefieciente de Jaccard para análise da similaridade entre QTermos e RStermos é definido a partir da seguinte fórmula [Han *et al.* 2012].

$$J = \frac{a}{a+b+c}$$

Onde,

a é o número de termos que existem tanto em QTermos quanto em RStermos;

b é o número de termos que existem apenas em QTermos;

c é o número de termos que existem apenas em RStermos.

A Figura 5.6 mostra o gráfico com os resultados obtidos em cada ponto deste experimento.

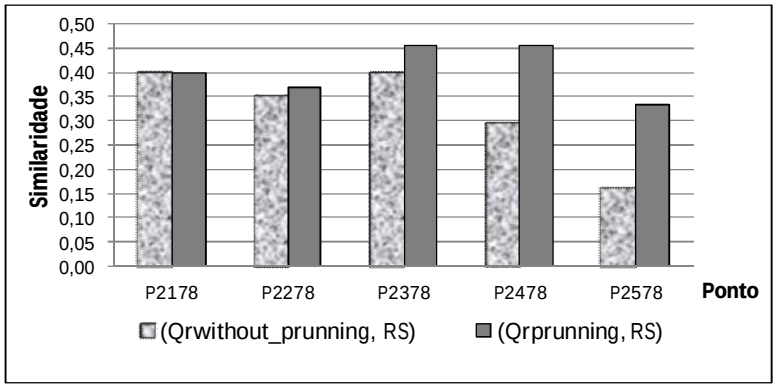


Figura 5.6– Similaridade de Consultas Reformuladas em relação à RS

De acordo com o gráfico, é possível observar que *QRpruning* se mantém em um melhor nível de similaridade com a RS. Além disto, em *QRwithout_Pruning*, devido à heterogeneidade das fontes e ao distanciamento semântico entre o ponto de submissão e os demais pontos, a consulta reformulada tem uma tendência natural de diminuir o seu nível de similaridade em relação a SR.

Outro conjunto de resultados obtido com este experimento está relacionado à avaliação do grau de preservação semântica desta consulta. Para isto foi gerada uma medida de precisão a partir da seguinte fórmula:

$$Precisão_{RQ} = \frac{|T_{rel\_RQ}|}{|T_{rec\_RQ}|}$$

Onde,

RQ é a consulta reformulada

T_{rel_RQ} é o conjunto de termos relevantes gerados em RQ .

T_{rec_RQ} é o conjunto de termos gerados em RQ.

Esta fórmula foi aplicada em cada ponto, após a reformulação da consulta, com o objetivo de produzir o valor de precisão das consultas ao longo do roteamento. A Figura 5.7 mostra os resultados obtidos na comparação entre os cenários de roteamento *QRwithout_Pruning* e *QRpruning*. Os resultados, apontados no gráfico, mostram o ganho de precisão quando o roteamento da consulta considera a consulta original ao longo do roteamento e mantém as reformulações semanticamente próximas da consulta original. Da mesma forma que a análise da similaridade, a precisão diminui à medida em que os pontos se distanciam do ponto de submissão. No caso do *QRprunning*, esta perda de precisão está relacionada ao distanciamento semântico existente entre os dois pontos. Em relação a QRwithout_*prunning*, além do distanciamento semântico, a consulta reformulada a cada ponto tende a incorporar termos que não têm relação semântica direta com os termos da consulta original.

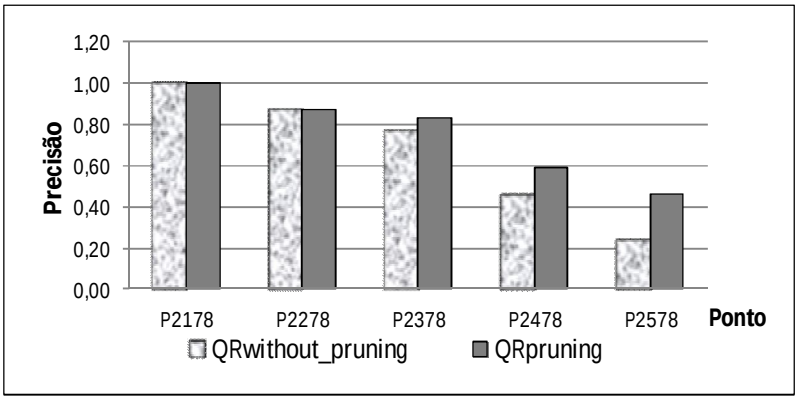


Figura 5.7 - Precisão das Consultas Reformuladas em Relação a Referência Semântica

- **Verificação da Hipótese**

  Hipótese H2. Ao longo do roteamento, a eliminação de termos da consulta reformulada, que não possuem correspondência semântica com os termos da consulta original, permite preservar a semântica da consulta original.

A análise dos resultados mostrou que é preciso considerar a consulta original, ao longo do roteamento, se desejamos produzir reformulações mais próximas ao que o usuário definiu no momento da submissão. Os gráficos representados nas Figuras 5.6 e 5.7 mostram o distanciamento gerado a cada reformulação, ao longo do roteamento, caso os conceitos sem RSD (Relação Semântica Direta) com os conceitos

originais sejam mantidos. Assim, é possível concluir que os resultados apresentados confirmam a hipótese H2.

5.3.2. Experimento 2

Este experimento objetiva avaliar as hipóteses H1 e H3 a partir dos resultados obtidos utilizando todos os elementos que compõem a abordagem *SemRouting*: estratégia de preservação semântica, estratégia para seleção de pontos e mecanismo de parada.

- **Cenário**:

Uma rede formada por um conjunto de quarenta pontos (mostrada no Apêndice B). Neste experimento, o roteamento foi limitado em um número de saltos de valor igual a seis. Para análise do valor mínimo de relevância do ponto, para o qual a consulta poderá ser encaminhada para execução, utilizamos o valor de *threshold* de 0.5 (justificado no Apêndice C). Neste cenário, uma consulta *Q={NetworkManager* ∪ *Project}* foi submetida no ponto P4678 com preferências de enriquecimento estabelecidas para as variáveis (*Generalization, Specialization*). Os experimentos foram realizados considerando duas abordagens: *Flooding,* como baseline, e a *SemRouting*.

- **Descrição do Experimento**:

Na primeira execução por *Flooding*, a consulta é roteada sem considerar qualquer critério de relevância semântica, ou seja, qualquer ponto é consultado desde que exista uma consulta que possa ser encaminhada para este ponto.

Na segunda execução, utilizando a abordagem *SemRouting*, foi necessário estabelecer algumas configurações complementares. Foram atribuídos valores às variáveis de ambiente conforme mostra a Figura 5.8. Na figura, *limitRelevance* define o limite mínimo para seleção dos pontos relevantes. As variáveis, *pc*, *ec*, *responsiveness*, *timeliness* e *executionTime* correspondem aos valores iniciais associados aos critérios de qualidade de Relevância (preservação + enriquecimento), Capacidade de Resposta, Atualidade e Tempo de Execução, respectivamente. Os valores foram definidos considerando que a relevância é o fator determinante para

seleção do melhor conjunto de pontos dentre os pontos relevantes (ou seja, pontos cuja relevância está acima de 0.5).

```
1  package speed2010.routing.quality;
2
3  public class Weights {
4      public static final double limitRelevance = 0.5; // threshold de relevancia
5      public static double pc = 0.4;                   // preservação da consulta
6      public static double ec = 0.2;                   // enriquecimento da consulta
7      public static double responsiveness = 0.16;      // capacidade de resposta
8      public static double timeliness = 0.13;          // atualidade
9      public static double executationTime = 0.11;     // tempo de execução
10
11 }
```

Figura 5.8 – Variáveis de Ambiente

- **Análise dos Resultados**:

Analisamos os resultados diante da seguinte perspectiva: o número de pontos consultados e a similaridade semântica entre as reformulações geradas e a consulta original, a cada salto do roteamento.

A Figura 5.9 mostra o comparativo entre o percentual de pontos consultados ao longo do roteamento. No roteamento por *Flooding*, como as consultas são encaminhadas sem considerar qualquer tipo de critério, existe um aumento no volume dos pontos a cada salto do roteamento. As consultas são reformuladas e encaminhadas sem considerar a consulta original. No caso da abordagem *SemRouting*, o número de pontos consultados decresce em decorrência do processo de perda semântica da consulta e da eliminação de pontos que não atendem aos requisitos de qualidade. Na *SemRouting* pontos com baixa relevância são eliminados da lista de pontos candidatos no roteamento. Em seguida, dentre os relevantes os melhores são selecionados de acordo com os critérios de classificação (relevância, capacidade de resposta, tempo de execução e atualidade).

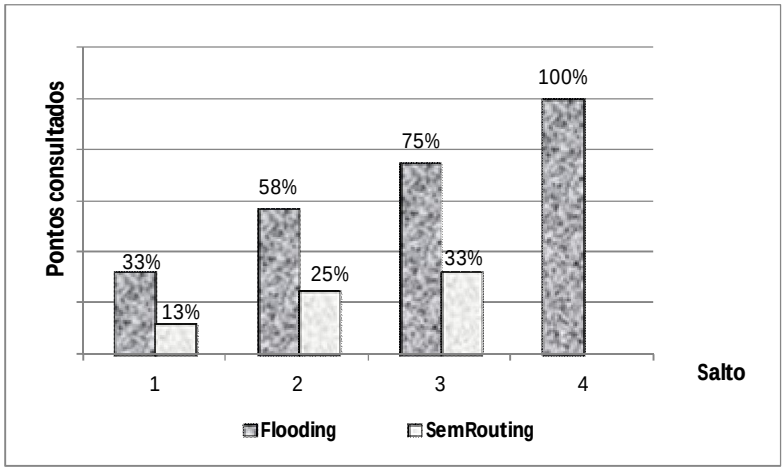


Figura 5.9– Percentual acumulado de pontos consultados

No *SemRouting*, aspectos semânticos e de qualidade foram combinados de forma a obter os melhores pontos, dentre o conjunto de pontos relevantes disponíveis, no momento da atividade de seleção. Para ilustrar, a Figura 5.10 mostra o contexto da atividade para o ponto P3778. Na figura, é possível observar: os pontos disponíveis (coluna Disponibilidade com valor igual a TRUE), os pontos eliminados pela baixa relevância (coluna Eliminado com valor igual a TRUE) e os pontos desclassificados de acordo com os critérios (coluna Classificado com valor null). Para o ponto P3778, 5 pontos foram analisados, 1 foi eliminado pela baixa relevância e 2 desclassificados de acordo com os critérios de qualidade. As colunas V.Preservação e V.Enriquecimento mostram os valores de preservação e enriquecimento utilizados para calcular a Relevância. As demais colunas, Capac.Resposta, Atualidade e Tempo Execução correspondem aos critérios de qualidade que são utilizados, junto com a relevância, para produzir o GlobalIQ.

Ponto	Ponto Vizinho	Relevancia	V.Preservação	V.Enriquecimento	Capac. Resposta	Atualidade	Tempo Execução	GlobalIQ	Disponibilidade	Eliminado	Classificado
3778	3978	.40	.00	.40	.00	.00	.00	0	TRUE	TRUE	null
3778	4078	.90	.50	.40	.00	61.46	.00	.15	TRUE	FALSE	null
3778	4678	1.40	1.00	.40	.64	811.46	1.71	.67	TRUE	FALSE	true
3778	4778	.90	.50	.40	.62	140.50	2.00	.24	TRUE	FALSE	true
3778	4978	.90	.50	.40	.00	81.46	.00	.15	TRUE	FALSE	null

Figura 5.10– Contexto Associado à Atividade de Seleção de Pontos em P3778.

De forma semelhante ao Experimento 1, utilizando o coeficiente de *Jaccard* foi possível estabelecer medidas para avaliação da similaridade entre as reformulações produzidas e a RS. A Figura 5.11 mostra a média deste coeficiente para o conjunto de reformulações produzidas a cada salto. De acordo com o resultado, as consultas

roteadas pelo *SemRouting* apresentam um melhor grau de similaridade em relação à consulta original. No *Flooding*, as consultas sofrem uma maior perda de similaridade se comparada a *SemRouting*. Este aumento é consequência da não preservação semântica da consulta, em relação à consulta original, e do encaminhamento de consultas com baixa relevância semântica, ou seja, sem critério.

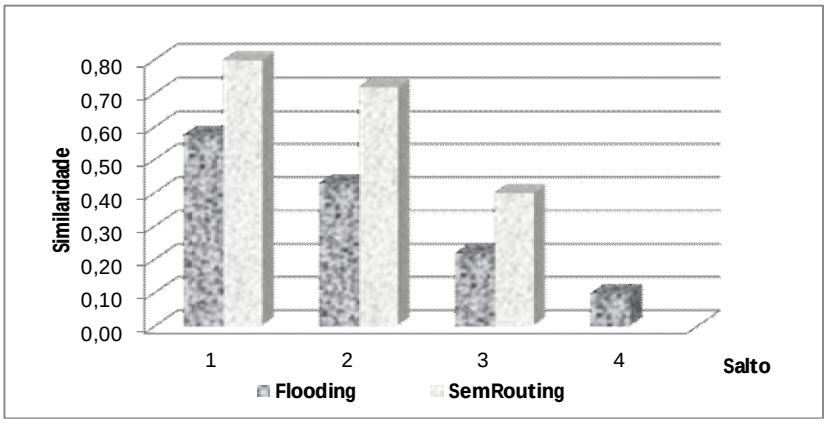


Figura 5.11– Média da Similaridade entre as Consultas Reformuladas e a Referência Semântica

Na Figura 5.12 é possível observar que as reformulações produzidas no *SemRouting* estão preservadas no melhor nível de aproximação semântica com a RS. As variações em número de termos ocorrem a cada nova reformulação mas são preservadas de acordo com a estratégia de preservação semântica da *SemRouting*. Não existem nas reformulações termos não relacionados com a consulta original.

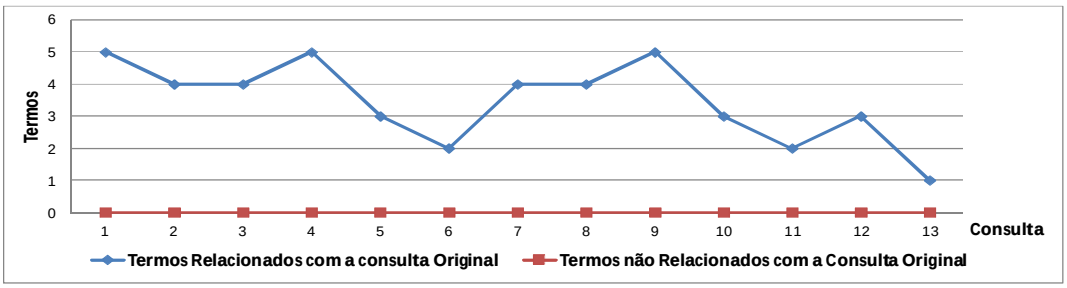


Figura 5.12– Consultas Roteadas pelo SemRouting

Nas consultas por *Flooding* (Figura 5.13), termos sem relação com os termos originais vão sendo adicionados à consulta a partir da segunda reformulação (posição 2 do eixo Consulta). Termos adicionados à consulta sem qualquer critério semântico, em relação à consulta original, comprometem a semântica da consulta a cada nova reformulação. Em algumas consultas no gráfico (por exemplo, consultas de 30 a 34) os

termos obtidos na reformulação não possuem relação semântica com os termos originais. Nessa situação, as consultas reformuladas deverão produzir um conjunto de resultados inadequados à consulta original submetida pelo usuário.

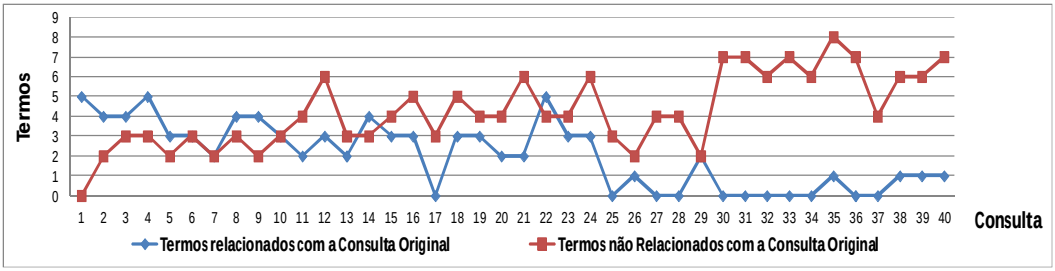


Figura 5.13– Consultas Roteadas por Flooding

- **Verificação da Hipótese**

  Hipótese H1. O uso de informações semânticas combinadas à qualidade da informação produz resultados que permitem ao sistema obter um melhor conjunto de pontos capazes de responder a consulta do usuário.

De acordo com os resultados apresentados, os critérios de qualidade podem ser utilizados para eliminar e classificar o melhor conjunto de pontos. Estes critérios podem ser melhor representados quando definidos dentro de um contexto específico no momento em que um ponto recebe uma consulta. No *SemRouting*, os pontos que não possuem relevância são inicialmente eliminados. Em seguida, utilizando os demais critérios e produzindo uma classificação, é possível fazer a escolha dos melhores pontos dentre aqueles considerados relevantes (Figura 5.10). No roteamento por *Flooding* a consulta é roteada sem qualquer critério, por consequência, nem todos os pontos selecionados são capazes de responder a consulta original. Assim, concluímos que o experimento confirma a hipótese H1.

Hipótese H3. A análise da perda semântica da consulta reformulada, em relação à consulta original, identifica os pontos onde as consultas devem ser executadas durante o roteamento.

Com o *SemRouting* o roteamento é interrompido (Figura 5.11), com limite de salto igual a 3, em função da perda semântica da consulta, ao longo do roteamento. Por outro lado, o *Flooding* continuou propagando a consulta sem considerar a

semântica da consulta original, apenas levando em conta o limite de saltos definido (TTL). Analisar a perda semântica da consulta, ao longo do roteamento, identifica se a consulta pode ser executada no ponto para o qual será encaminhada. Com base nas análises dos resultados confirmamos a hipótese H3.

5.4. Considerações

Este capítulo mostrou aspectos relacionados à implementação e à execução da abordagem *SemRouting* no protótipo do SPEED. Foram descritos os experimentos realizados para verificação das hipóteses levantadas nesta tese em relação à preservação semântica da consulta e à seleção do melhor conjunto de pontos durante o roteamento da consulta. A análise dos resultados produzidos nos experimentos mostrou que as estratégias adotadas na abordagem *SemRouting* confirmam estas hipóteses. Outros cenários de execução do experimento foram realizados e apresentaram resultados similares ao apresentado neste capítulo. O Apêndice D mostra um destes cenários e apresenta a análise dos resultados obtidos.

O próximo capítulo apresenta as conclusões desta tese e indica trabalhos futuros.

Capítulo 6

Conclusão e Trabalhos Futuros

Em um PDMS, devido à ausência de um esquema global, estratégias de roteamento têm se dedicado a fazer o encaminhamento da consulta para somente aqueles pontos que possam melhor contribuir com o resultado final dessa consulta. Cada ponto precisa decidir independentemente dos demais pontos que compõem a rede, para qual dos seus vizinhos a consulta deverá ser encaminhada. No momento desta decisão, pontos podem estar indisponíveis ou até mesmo serem irrelevantes para a consulta submetida. Além destes fatores, devido à heterogeneidade das fontes, uma consulta quando reformulada durante o roteamento pode sofrer alterações em sua semântica.

Motivada por estes problemas, esta tese investigou, levantou hipóteses e propôs uma abordagem que utiliza aspectos semânticos e de qualidade no roteamento de consultas em PDMS. A abordagem, denominada *SemRouting*, compreende uma estratégia para identificação e seleção do melhor conjunto de pontos, um modelo para representação das informações semânticas e de qualidade, e uma estratégia para análise e preservação da semântica da consulta original durante o roteamento.

Este capítulo está organizado da seguinte forma: a Seção 6.1 mostra as contribuições desta tese e a Seção 6.2 relaciona os trabalhos futuros.

6.1. Contribuições da Tese

As contribuições desta tese podem ser descritas como segue:

- Modelo para representação das informações semânticas e de qualidade [Freire *et al.* 2012a; Freire *et al* 2012c]

  Especificamos e implementamos um modelo para representação das informações semânticas e de QI com o objetivo de expandir o conhecimento local do ponto e facilitar o processo de roteamento da consulta em um PDMS. Identificamos os elementos contextuais associados à consulta e ao ponto como forma de compor o contexto para a atividade de seleção de pontos. O armazenamento do contexto permitiu conhecer as interações do ponto em relação aos seus vizinhos a cada consulta recebida para um novo encaminhamento.

- Estratégia para análise e preservação da semântica da consulta original ao longo do roteamento [Freire *et al.* 2014].

  Especificamos e implementamos uma estratégia para análise e preservação da semântica da consulta original durante o roteamento. Formalizamos a definição de perda semântica e definimos uma referência semântica para a consulta original baseada na ontologia de domínio. A partir da referência semântica, as consultas puderam ser preservadas no melhor nível de aproximação semântica da consulta original. Baseado nos pesos associados a cada correspondência semântica, foi possível interpretar de forma correta o tipo de associação semântica existente entre os termos obtidos em cada reformulação e os termos originais. Mostramos os ganhos alcançados em termos de precisão e similaridade em relação a referência semântica com a aplicação da estratégia.

- Estratégia para identificação e seleção do melhor conjunto de pontos [Freire *et al.* 2012b]

  Especificamos e implementamos uma estratégia para seleção do melhor conjunto de pontos aptos a responder uma dada consulta submetida. Consideramos que as decisões do ponto para o roteamento da consulta são baseadas nas informações locais que tem do ambiente. Assim, expandimos o conhecimento do ponto a partir da definição do conjunto de elementos contextuais (disponibilidade do ponto, preferências do usuário e medidas de qualidade) que poderiam tornar mais específica a atividade de seleção de pontos. Para produzir as medidas de

qualidade, definimos um conjunto de critérios de qualidade (Relevância, Capacidade de Resposta, Atualidade e Tempo de Execução) e suas respectivas métricas. Devido ao grande número de pontos em um PDMS e com o objetivo de reduzir o esforço computacional na geração das medidas de qualidade, consideramos utilizar a medida de relevância de forma eliminatória e as demais, junto com relevância, de forma classificatória.

- Uma abordagem baseada em semântica para o roteamento de consultas em PDMS [Freire *et al.* 2012c; Freire *et al.* 2013].

  Especificação e implementação da abordagem SemRouting com o objetivo de fazer o roteamento de consultas em um PDMS, preservando a semântica da consulta original e selecionando o melhor conjunto de pontos capazes de responder a consulta. Para isto, a abordagem integra as estratégias de preservação e de seleção de pontos, adota o modelo para representação das informações de roteamento e define um mecanismo de parada para o roteamento. Este mecanismo, também baseado em semântica, utiliza as informações provenientes da análise da perda semântica da consulta para interromper a rota da consulta evitando que consultas com alta perda semântica sejam encaminhadas para execução em outro ponto.

6.2. Trabalhos Futuros

Considerando a área de aplicação desta tese, especificamente nas questões relacionadas ao roteamento, outras pesquisas poderão ser realizadas:

- Preservação Semântica da Consulta

  Expandir o mecanismo de preservação semântica para tratar com operadores da consulta.

- Qualidade da Informação

  Incorporar a *SemRouting* outros critérios de qualidade tais como: confiança, reputação e completude de resultados (relativo aos dados) como forma de melhor qualificar o ponto.

- Integração dos Resultados

  Esta atividade não foi considerada no escopo desta tese. Assim, existe todo um trabalho a ser realizado em relação à atividade de integração, considerando informações relacionadas à proveniência dos dados e tratamento da duplicidade de informações provenientes de diferentes fontes de dados.

- Decisão de Roteamento considerando o Histórico de Consultas

  Expandir o conhecimento do ponto com informações que possam ser inferidas a partir do contexto armazenado no ponto em atividades anteriores. Para isto o modelo de representação das informações do roteamento, definido nesta tese, poderá ser usado como base deste processo.

Referências

AGHAMAHMOODI, S.; RANKOOHI, S. M. T. R.; AGHAMAHMOODI, F. **A new ant colony optimization-based algorithm for range query answering problem in relational schema-based P2P database systems**. In Int. Journal of Knowledge and Information Systems, p. 1-31, 2014.

AHMED, R.; BOUTABA, R. **A Survey of Distributed Search Techniques in Large Scale Distributed Systems**. In IEEE Communications Surveys & Tutorials, v.13, n.2, p.150-167, 2011

AIELLO, C.; BALDONI, R.; BIANCHINI, D.; BOLCHINI, C.; BONOMI, S.; CASTANO, S.; PERGIOVANNI, S. T. **The ESTEEM Architecture for Emergent Semantics and Cooperation in MultiKnowledge Environments**. Relatório Técnico. MIUR PRIN Esteem Project. Università Degli Studi Di Milano, Italy, 2007.

AKBARINIA R.; PACITTI, E.; VALDURIEZ P. **Query processing in P2P systems**. Relatório Técnico de Pesquisa. n. 6112, INRIA – Institute National de Recherche en Informatique et en Automatique – ISSN 0249-6399, 2007

ALVES, D.; FREITAS, M.; MOURA, T.; SOUZA, D. **Using social network information to identify user contexts for query personalization**. InDBKDA 2013, In Proceedings of the 5th International Conference on Advances in Databases, Knowledge, and Data Applications, Seville, Spain, p. 45-51, 2013.

ARAZY, O.; KOPAK, R. **On the measurability of information quality**. In Int. Journal of the American Society for Information Science and Technology, v. 62, n. 1, p. 89-99, 2011.

AUMUELLER, D.; DO, H.H.; MASSMANN, S.; RAHM, E. **Schema and ontology matching with COMA++**. In Proceedings of SIGMOD, demo paper, p. 906-908, Maryland, USA, 2005.

BANDARA, K. Y.; WANG, M.; PAHL, C. **An extended ontology-based context model and manipulation calculus for dynamic Web service processes**. In Int. Journal of Service Oriented Computing and Applications, p. 1-20, 2013.

BATINI C.; SCANNAPIECO M. **Data Quality: Concepts, Methods, and Techniques**. (Chapther 2). Springer, 2006.

BATISTA, M. C. M. **Information Quality Analysis in a Data Integration System**. Federal University of Pernambuco (UFPE/CIn). Thesis Proposal and Qualification Exam. Recife, PE, Brazil, 2006.

BATISTA, M. C. M. **Schema Quality Analysis in a Data Integration System**. Federal University of Pernambuco (UFPE/CIn). Phd Thesis. Recife, PE, Brazil, 2008.

BELIAN, R.; SALGADO, A. C. **A Context-based Schema Integration Process Applied to Healthcare Data Sources**. In Proceedings of the International Conference On the move to meaningful internet systems, Springer-Verlag, Berlin, Heidelberg, 2010.

BERNSTEIN P.; GIUNCHIGLIA F.; KEMENTSIETSIDIS A.; MYLOPOULOS J.; SERAFINI L.; ZAIHRAYEU I. **Data Management for Peer-to- Peer Computing: A Vision**. In Proceedings of the 5th International Workshop on the Web and Databases, Wisconsin, USA, 2002.

BHOGAL, J.; MACFARLANE, A.; SMITH, P. **A review of ontology based query expansion**. In Int. Journal of Information processing & management, v. 43, n. 4, p. 866-886, 2007.

BIANCHINI D.; DE ANTONELLIS V.; MELCHIORI, M. **P2P-SDSD: On-the-fly service-based collaboration in distributed systems**. In Int. Journal of Metadata, Semantics and Ontologies, v. 5, n. 3, p. 222–237, 2009.

BLOOM B. **Space/time tradeoffs in hash coding with allowable errors**. In Communications of the ACM, v. 13, n. 7, p. 422-426, 1970.

BOLCHINI, C.; CURINO, C. A.; ORSI, G.; QUINTARELLI, E.; ROSSATO, R.; SCHREIBER, F. A.; TANCA, L. **And what can context do for data?**. In Communications of the ACM, v.

52, n. 11, p. 136-140, 2009.

BOUZEGHOUB, M; PERALTA, V. **A framework for analysis of data freshness**. In Proceedings of the international workshop on Information quality in informational systems, ACM Press, p. 59-67, 2004.

BRÉZILLON, P. **Context-based development of experience bases**. In Modeling and Using Context. Springer Berlin Heidelberg, p. 87-100, 2013.

BYRNE, N.; DUNNE, C. R.; HOGAN, J.; MCCANN, F.; MCNAMEE, A.; O'SULLIVAN, M. **Methods, systems and devices for dynamic context-based routing**. U.S. Patent Application 13/309,008, 2011.

CAMPOS, L. M.; FERNÁNDEZ-LUNA, J. M.; HUETE, J. F.; VICENTE-LÓPEZ, E. **XML search personalization strategies using query expansion, reranking and a search engine modification**. In Proceedings of the 28th Annual ACM Symposium on Applied Computing, ACM, Coimbra, Portugal, p. 872-877, 2013.

CAO, G.; NIE, J.-Y.; GAO, J.; ROBERTSON, S. **Selecting good expansion terms for pseudo-relevance feedback**. In SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval, New York, USA, ACM, p. 243–250, 2008.

CARPINETO, C.; ROMANO, G. **A Survey of Automatic Query Expansion in Information Retrieval**. In ACM Computing Surveys, v.44, n.1, p. 1-50, 2012.

CASTANO S.; MONTANELLI, S. **Semantically routing queries in peer-based systems: the H-Link approach**. In Int. Jounal The Knowledge Engineering Review, v. 23, p. 51-72, 2008.

CASTANO S.; FERRARA, A.; MONTANELLI, S. **Matching ontologies 2006 in open networked systems: Techniques and applications**. In Journal on Data Semantics, p. 25-63, 2006.

CASTANO, S.; FERRARA, A.; MONTANELLI, S. **P2P routing-by-content on a lightweight community basis**. In Proceedings of On the Move to Meaningful Internet Systems

2007: OTM 2007 Workshops, p. 477-486, Springer Berlin Heidelberg, 2007.

CHEN, G.; HU, T.; JIANG, D.; LU, P.; TAN, K. L.; VO, H. T.; WU, S. **BestPeer++: A Peer-to-Peer Based Large-Scale Data Processing Platform**. In Proceedings of 28th International Conference on. Data Engineering (ICDE), p. 582-593, 2012.

CRUZ, I.F.; ANTONELLI, F.P.; STROE, C. **AgreementMaker: Efficient Matching for Large Real-World Schemas and Ontologies.** PVLDB 2(2), demo paper, p. 1586-1589, 2009.

DEDZOE, W.; K., LAMARRE, P.; AKBARINIA, R. **As-Soon-As-Possible Top- k Query Processing in P2P Systems**. In Transactions on Large-Scale Data- and Knowledge-Centered Systems IX. Lecture Notes in Computer Science, v. 7980, p. 1-27, 2013.

DEL PILAR ANGELES, M.; MACKINNON, L. M. **Assessing Data Quality of Integrated Data by Quality Aggregation of its Ancestors**. Computación y Sistemas, v. 13, n. 3, p. 331-344, 2010.

DELVEROUDIS Y.; LEKEAS, P. V. **Managing Semantic Loss during Query Reformulation in Peer Data Management Systems**. In Proceedings of the Databases for Next Generation Researchers, 2007. SWOD 2007. IEEE International Workshop on. IEEE, 2007Washington, USA, p. 61-66, 2007.

DELVEROUDIS, Y.; LEKEAS, P. V.; SOULIOU, D. **On Estimating Semantic Loss in Peer Data Management Systems**. In Proceedings of First International Conference on Advances in P2P Systems, Sliema, Malta, p. 51-53, 2009.

DEY A. K.; SALBER D.; ABOWD G. D. **A Conceptual Framework and a Toolkit for Supporting the Rapid Prototyping of Context-Aware Applications**, In Int. Journal of Human Computer Interaction, v. 16, n. 2, p. 97-166, 2001.

DOAN, A.; HALEVY, A.; IVES, Z. **Principles of data integration**. 1st Edition, Elsevier, 2012.

DONG, X. L.; SAHA, B.; SRIVASTAVA, D. **Less is More: Selecting Sources Wisely for Integration**. In Proceedings of VLDB, v.6, n.2, p. 37-48, 2012.

DOVAL D.; O'MAHONY D. **Overlay Networks: A Scalable Alternative for P2P**. In Int.

Journal of IEEE Internet Computing, v. 7, n. 4, p. 79-82, 2003.

EUZENAT, J.; SHVAIKO, P. **Ontology Matching**. 2nd Edition, Springer, ISBN: 978-3-642-38720-3, 2013.

FAYE, D.; NACHOUKI, G.; VALDURIEZ, P. **Semantic query routing in senpeer, a p2p data management system**. In Proceedings of Network-Based Information Systems, p. 365-374. Springer Berlin Heidelberg, 2007.

FIORANO. **Super-Peer Architectures for Distributed Computing**. White Paper, Fiorano Software, Inc. Disponivel em http://www.fiorano.com/whitepapers/superpeer.pdf. Acessado em 20 de janeiro de 2014.

FREIRE, C. A. **Roteamento Semântico de Consultas em Sistemas Gerenciadores de Dados P2P**. Qualificação e Proposta de Tese. Centro de Informática, Universidade Federal de Pernambuco, 2012.

FREIRE C.; CYSNEIROS N.; SOUZA D.; SALGADO A. C. **Preserving the Original Query Semantics in Routing Processes**. In Proceedings of 16th International Conference on Enterprise Information Systems (ICEIS), Lisboa, Portugal, 2014.

FREIRE C.; SOUZA B. F. F.; SOUZA D.; BATISTA M. C. M.; SALGADO A. C. **Semantic Measures as Information Quality Criteria for Query Routing Processes**. In Int. Journal of Business Intelligence and Data Mining (IJBIDM), v. 8, n. 2, p. 167 – 183, 2013.

FREIRE C.; SOUZA D.; LÓSCIO, B. F.; SALGADO A. C. **Enriching Query Routing Processes in PDMS with Semantic Information and Information Quality**. In Proceedings of the 6º Alberto Mendelzon International Workshop on Foundations of Data Management, 2012, Ouro Preto- MG, Brasil, 2012a.

FREIRE C.; SOUZA B. F. F.; SOUZA D.; BATISTA M. C. M.; SALGADO A. C. **Towards an Information Quality Approach to Enhance Query Routing Processes**. In Proceedings of the 14th International Conference on Information Integration and Web-based Applications & Services (IIWAS 2012), Bali, Indonesia, 2012b.

FREIRE C.; SOUZA D.; SALGADO A. C. **Semantic-based Query Routing for PDMS**. In the Simpósio Brasileiro de Banco de Dados - SBBD, 2012, São Paulo, Brasil, 2012c.

GENNARO, C.; LENZI, R.; MANDREOLI, F.; MARTOGLIA, R.; MORDACCHINI, M.; PENZO, W.; SASSATELLI, S. **A unified multimedia and semantic perspective for data retrieval in the semantic web**. In Int. Journal of Information Systems, v. 36, n. 2, p. 174-191, 2011.

GENNARO, C.; MORDACCHINI, M.; ORLANDO, S.; RABITTI, F. **MRoute: A peer-to-peer routing index for similarity search in metric spaces**. In Proceedings of the 5th International Workshop on Databases, Information Systems and Peer-to-Peer Computing (DBISP2P 2007), p. 1-12, 2007.

GNUTELLA. **The Gnutella System** web site http://www.gnutella2.com/. Último acesso em março de 2014

GRUBER T. **A Translation Approach to Portable Ontology specification**. In Int. Journal of Knowledge Acquisition, v. 5, n. 2, p. 199-220, 1993.

HAASE, P.; SIEBES, R.; VAN HARMELEN, F. **Expertise-based peer selection in Peer-to-Peer networks**. In Int. Journal of Knowledge and Information Systems, v. 15, n. 1, p. 75-107, 2008.

HALEVY, A.; RAJARAMAN, A.; ORDILLE, J. **Data integration: the teenage years**. In. Proceedings of the 32nd international conference on Very large data bases, Seoul, Korea, p. 9-16, 2006.

HALEVY, A.; IVES, Z.; MADHAVAN, J.; MORK, P.; SUCIU, D.; TATARINOV, I. **The Piazza Peer Data Management System**. In Journal of IEEE Transactions on Knowledge and Data Engineering TKDE, v. 16, n. 7, p. 787-798, 2004.

HAN, J. W.; KAMBER, M.; PEI, J. **Data Mining Concepts and Techniques**: Elsevier Inc, 3nd edition, 2012.

HATEM, A. M.; ASHRAF, A. **Schema clustering and retrieval for multi-domain pay-as-you-go data integration systems**. In Proceedings of SIGMOD Conference, Indian, USA, p. 411–422, 2010.

HERSCHEL, S.; HEESE, R. **Humboldt discoverer: A semantic p2p index for pdms**. In

Proceedings of the International Workshop Data Integration and the Semantic Web, Porto, Portugal, 2005.

HOSE, K.; ROTH, A.; ZEITZ, A.; SATTLER, K. U.; NAUMANN, F. **A research agenda for query processing in large-scale peer data management system**s. In Int. Journal of Information Systems, v. 33, n. 7, p. 597-610, 2008.

HWANG, C.-L.; YOON, K. **Multiple Attribute Decision Making: methods and applications. A State-of-the-art Survey**" Berlin ; Springer-Verlag, 1981.

ISMAIL, A.; QUAFAFOU, M.; DURAND, N.; NACHOUKI, G.; HAJJAR, M. **Queries Routing In Super-Peer-Based System: Simulation and Evaluation**. In Int. Journal of Emerging Technologies in Web Intelligence, v. 3, n. 3, 2011.

ISMAIL, A.; HAJJAR, M.; QUAFAFOU, M.; DURAND, N.; EL SAYED, M. **Clustering using Hypergraph for P2P Query Routing**. In Proceedings of ICEIS, Angers, France, p. 277, 2013.

JOUNG, Y.; CHUANG, F. **OntoZilla: An ontology-based, semi-structured, and evolutionary peer-to-peer network for information systems and services**. In Int. Journal of Future Generation Computer Systems, v. 25, n. 1, p. 53-63, 2009.

KANTERE, V; BOUSOUNIS, D.; SELLIS, T. K. **GrouPeer: A System for Clustering PDMSs**. In Proceedings of the VLDB Endowment, Washington, USA, v. 4, p. 1371-1374, 2011.

KARVOUNARAKIS, G.; GREEN, T. J.; IVES, Z. G.; TANNEN, V. **Collaborative data sharing via update exchange and provenance**. In ACM Transactions on Database Systems, v.38, n.3, p. 1-42, 2013.

LI, J.; VUONG, S. **OntSum : A Semantic Query Routing Scheme in P2P Networks Based on Concise Ontology Indexing**. In Proceedings of 21st International Conference on Advanced Networking and Applications (AINA'07), Ontario, Canada, p. 94-101, 2007.

LÓSCIO, B. **Managing the Evolution of XML-based Mediation Queries**. PHD Thesis, Federal University of Pernambuco, Brazil, 2003.

LV, Q.; CAO, P.; COHEN, E.; LI, K.; SHENKER, S. **Search and Replication in Unstructured**

Peer-to-Peer Networks. In Proceedings of the 16th ACM International Conference on Supercomputing (ICS'02), New York, USA, p. 84-95, 2002.

MALAKOOTI, B.; THOMAS, I.; TANGUTURI, S. K.; GAJUREL, S.; KIM, H. E.; BHASIN, K. **Multiple criteria network routing with simulation results**. In Proceedings of the 15th Industrial Engineering Research Conference (IERC), London, England, p. 1-5, 2006.

MANDREOLI, F.; MARTOGLIA, R.; PENZO, W.; SASSATELLI, S. SRI @ work : Efficient and **Effective Routing Strategies in a PDMS**. In Proceedings of Web Information Systems Engineering – WISE 2007 8th International Conference on Web Information Systems Engineering, , Nancy, France, p. 285-297, 2007.

MANDREOLI, F.; MARTOGLIA, R.; PENZO, W.; SASSATELLI, S.; VILLANI, G. SUNRISE : **Exploring PDMS Networks with Semantic Routing Indexes**. In Proceedings of 4th European Semantic Web Conference, Innsbruck, Austria, 2007a.

MANDREOLI, F.; MARTOGLIA, R.; PENZO, W.; SASSATELLI, S.; VILLANI, G. Leveraging **Semantic Approximations in Heterogeneous XML Data Sharing Networks: The SUNRISE Approach**. Soft Computing in XML Data Management. Springer, p. 315-350. ISBN: 978-3-642-14009-9, 2010

MÁRQUEZ-BARJA, J.; CALAFATE, C. T.; CANO, J. C.; MANZONI, P. **An overview of vertical handover techniques: Algorithms, protocols and tools**. Computer Communications, v. 34, n. 8, p. 985-997, 2011.

MAZAK, A.; SCHANDL, B.; LANZENBERGER, M. A. **Heuristic-based Method for Approximating the Mismatch-at-Risk in Schema-based Ontology Alignment**. In Proceedings of International Conference on Knowledge Engineering and Ontology Development (KEOD 2010), p. 17-26, 2010.

MONTANELLI, S.; CASTANO, S. **Semantically routing queries in peer-based systems: the H-Link approach**. The Knowledge Engineering Review, Cambridge University Press, v. 23, n.1, p. 51-72, 2008.

MONTANELLI, S.; BIANCHINI, D.; AIELLO, C.; BALDONI, R.; BOLCHINI, C.; BONOMI, S.; TANCA, L. **The ESTEEM platform: enabling P2P semantic collaboration through**

emerging collective knowledge. In Journal of Intelligent Information Systems, v. 36, n. 2, p. 167-195, 2011.

NAUMANN F. **From databases to information systems - information quality makes the difference**. In Proceedings of the 6th International Conference on Information Quality (IQ), Cambridge (MA US), p. 244–260, 2001.

NAUMANN, F.; FREYTAG, J. C. **Quality-driven Source Selection using Data Envelopment Analysis**. In Proceedings of the 3rd Conference on Information Quality (IQ' 98). Cambridge, MA, 1998.

NAUMANN, F.; ROLKER, C. **Assessment Methods for Information Quality Criteria**. In Proceedings of the Conference on International Quality (IQ00) Boston, 2000.

OMHENI, N.; ZARAI, F.; OBAIDAT, M. S.; SMAOUI, I.; KAMOUN, L. **A MIH-based approach for best network selection in heterogeneous wireless networks**. In Journal of Systems and Software. p.1-14, 2014.

ORSI, G.; TANCA, L. **On-the-Fly and Context-Aware Integration of Heterogeneous Data Sources**. In Methodologies and Technologies for Networked Enterprises Springer Berlin Heidelberg, p. 259-276, 2012.

PETRAKIS, Y.; KOLONIARI, G.; PITOURA, E. **On using histograms as routing indexes in peer-to-peer systems**. In Databases, Information Systems, and Peer-to-Peer Computing, Springer Berlin Heidelberg, p. 16-30, 2005.

PILLAC, V.; GENDREAU, M.; GUÉRET, C.; MEDAGLIA, A. L. **A review of dynamic vehicle routing problems**. In European Journal of Operational Research, v. 225, n. 1, p. 1-11, 2013.

PIPINO, L. L.; LEE, Y.; WANG, R. **Data Quality Assessment**. In the Communications of the ACM, v. 45, n. 4, p. 211, 2002

PIRES C. **Um Sistema P2P de Gerenciamento de Dados com Conectividade Baseada em Semantica**. Qualificação e Proposta de Tese. Universidade Federal de Pernambuco, Recife, Brasil, 2007.

PIRES C. **Ontology-based Clustering in a peer Data Management System**. Tese de Doutorado. Universidade Federal de Pernambuco. Recife, Brasil, 2009.

PIRES, C. E.; SOUZA, D.; KEDAD, Z.; BOUZEGHOUB, M.; SALGADO, A. C. **Using Semantics in Peer Data Management Systems**. In Knowledge Creation Diffusion Utilization, p. 2579-2582, 2009.

PIRES, C. E.; SOUZA, D.; LÓSCIO, B.; BELIAN, R.; TEDESCO, P.; SALGADO, A. C. **Using Ontologies to Enhance Data Management in Distributed Environments**. In Ibero American Meeting on Ontological Research, Gramado, Brazil, 2011.

PIRES, C. E.; SOUZA, D.; PACHÊCO, T.; SALGADO, A. C. **SemMatcher: A Tool for Matching Ontology-based Schemas**. In Proceeding of the 24th Brazilian Symposium on Data Bases (SBBD'09), Fortaleza, Brazil, 2009a.

PIRES, C. E.; SANTIAGO, R. M. L.; SALGADO, A. C.; KEDAD, Z.; BOUZEGHOUB, M. **Ontology-Based Clustering in a Peer Data Management System**. In Int. Journal of Distributed Systems and Technologies, v. 3, n. 2, p. 1-21, 2012.

QIN, L.; PEI C.; EDITH C.; KAI L.; SCOTT, S. **Search and replication in unstructured peer-to-peer networks**. In Proceedings of the 16th international conference on Supercomputing (ICS '02). ACM, New York, NY, USA, 2002.

RIJSBERGEN, C. J. **Information Retrieval**, 2nd Ed. Stoneham, MA: Butterworths, 1979.

ROTH, A.; NAUMANN, F. **Benefit and Cost of Query Answering in PDMS**. In Proceedings of the Int. Workshop on Databases, Information Systems and Peer-to-Peer Computing (DBISP2P), 2005.

ROTH, A.; NAUMANN, F. **System P : Completeness-driven Query Answering in Peer Data Management Systems**. In Proceedings of Business, Technologie and web (BTW'07), Aachen, Germany, p. 1-4, 2007.

ROTH, A.; SKRITEK, S. **Peer Data Management**. In Int. Journal of Data Exchange, Information, and Streams, v. 5, p. 185-215, 2013.

ROTH, A. **Efficient Query Answering in Peer Data Management Systems**. PhD thesis,

Humboldt Universität zu Berlin, Germany.

SAATY, T. L. **What is the analytic hierarchy process?** Springer Berlin Heidelberg, p. 109-121, 2012.

SALGADO, A. C.; SANTORO, F. M.; BORGES, M. R. S.; ARAUJO, R. M.; VIEIRA, V. **Gerência de Contexto e suas aplicações**. In Colibri Colloquium, 2009, Bento Gonçalves. XXIX Congresso da Sociedade Brasileira de Computação, Porto Alegre, 2009.

SALLES, M. A. V.; DITTRICH, J.-P.; KARAKASHIAN, S. K.; GIRARD O. R.; BLUNSCHI L. **iTrails: Pay-as-you-go information integration in dataspaces**. In Proceedings of VLDB, Vienna, Austria, p. 663-674, 2007.

SARMA, A. D.; DONG, X.; HALEVY, A. Y. **Bootstrapping pay-as-you-go data integration systems**. In Proceedings of the SIGMOD, Vancouver, Canada, p. 861-874, 2008.

SASSATELLI S. **Query Processing in a P**DMS. PhD. Thesis. Universit` a degli Studi di MODENA e REGGIO EMILIA, 2009.

SASSI, M.; MERIEM, A.; SAMI, T. **Perceived quality of service and context awareness strategy for heterogeneous wireless connectivity management**. In Proceedings of the Tenth International Symposium on Wireless Communication Systems (ISWCS), p. 1–5, 2013.

SCHLOSSER M.; SINTEK M.; DECKER S.; NEJDL W. **HyperCuP - Hypercubes, Ontologies and Efficient Search on P2P Networks**. In Proceedings of the 1st Workshop on Agents and P2P Computing, Bologna, Italy, 2003.

SCHOLLMEIER, R. **A Definition of Peer-to-Peer Networking for the Classification of Peer-to-Peer Architectures and Applications**. In Proceedings of the 1st International Conference on Peer-to-Peer Computing (P2P '01), Lingkoping, Sweden, p.101- 102, 2001.

SELIGMAN, L.; MORK, P.; HALEVY, A.Y.; SMITH, K.; CAREY, J. M.; CHEN, K.; WOLF, C.; MADHAVAN, J.; KANNAN, A.; BURDICK, D. **OpenII: An Open Source Information**

Integration Toolkit, In Proceedings of SIGMOD, p. 1057-1060, 2010.

SHARMA, S.; GUPTA, P. **A Review on Methods for Query Personalization**. In Proceedings of Intelligent Computing, Networking, and Informatics. Springer India, p. 1099-1106, 2014.

SHVAIKO, P.; GIUNCHIGLIA, F.; BUNDY, A.; BESANA, P.; SIERRA, C.; VAN HARMELEN, F.; ZAIHRAYEU, I. **Benchmarking methodology for good enough answers**. Technical report, DISI-08-003, Informatica e Telecomunicazioni, University of Trento, 2008.

SILVA, E. R.; SOUZA, G. B.; SALGADO, A. C. **Load Balance for Semantic Cluster-based Data Integration Systems**. In Proceedings of the 17th International Database Engineering & Applications Symposium (IDEAS'13). Barcelona, Spain, 2013.

SOUZA D. **Using Semantics to Enhance Query Reformulation in Dynamic Distributed Environments**. PhD Thesis, Federal University of Pernambuco (UFPE), Recife, PE, Brasil, 2009.

SOUZA, D.; ARRUDA, T.; SALGADO, A. C.; TEDESCO, P.; KEDAD, Z. **Using Semantics to Enhance Query Reformulation in Dynamic Environments**. In Proceedings of the 13th East European Conference on Advances in Databases and Information Systems (ADBIS'09), Riga, Latvia, 2009.

SOUZA, D.; LÓSCIO, B. F.; SALGADO, A. C.. **Combining Semantic Information and Information Quality on the Enrichment of web Data**. In Proceedings of the 8th WEBIST, Porto, Portugal, 2012.

SOUZA, D.; PIRES, C. E.; KEDAD, Z.; TEDESCO, P.; SALGADO, A. C. **A semantic-based approach for data management in a P2P system**. In Proceedings of Transactions on large-scale data-and knowledge-centered systems III, Springer Berlin Heidelberg, p. 56-86, 2011.

STOICA I.; MORRIS R.; KARGER D.R.; KAASHOEK M.F.; BALAKRISHNAN H. **Chord: a scalable peer-to-peer lookup service for internet applications**. In ACM Conf. on Applications, Technologies, Architectures, and Protocols for Computer Communications (SIGCOMM), p. 149-160, 2001.

TAHIR, H.; BRÉZILLON, P. **Shared Context for Improving Collaboration in Database Administration**. In Int. Journal of Database Management Systems, v. 5, n. 2, 2013.

TANCA, L.; BOLCHINI, C.; QUINTARELLI, E.; SCHREIBER, F. A.; ORSI, G. **Problems and Opportunities in Context-Based Personalization**. In Proceedings of the VLDB Endowment, v. 4, n. 11, p. 1-4, 2011.

THEOTOKIS, S. A.; SPINELLIS, D. **A survey of peer-to-peer content distribution technologies**. In ACM Computing Survey, v. 36, n.4, p. 335–371, 2004.

VIEIRA, V.; TEDESCO, P.; SALGADO, A. C. **Designing context-sensitive systems: An integrated approach**. In Int. Journal of Expert Systems with Applications, v.38, n. 2, p. 1119-1138, 2011.

WANG, R.Y.; STRONG, D.M. **Beyond Accuracy: What Data Quality Means to Data Consumers**. In Int. Journal of Management Information Systems, v.. 12, n. 4, p. 5-33, 1996.

WENNING, B. **Context-Based Routing in Dynamic Networks**. Tese de Doutorado. Vieweg+ Teubner, 2010.

XIAO H. **Query processing for heterogeneous data integration using ontologies**. PhD Thesis in Computer Science. University of Illinois at Chicago, 2006.

YANG, B.; GARCIA-MOLINA, H. **Improving search in peer-to-peer networks**. Distributed Computing Systems, 2002. In Proceedings of 22nd International Conference on IEEE, p. 5- 14, 2002

ZAIHRAYEU I. **Towards Peer-to-Peer Information Management Systems**. PhD thesis, University of Trento, 2006.

ZHAO, B. Y.; HUANG, L.; STRIBLING, J.; RHEA, S.C.; JOSEPH, A. D.; KUBIA-TOWICZ, J.D. **Tapestry: a resilient global-scale overlay for service deployment**. In IEEE Journal on Selected Areas in Communications (JSAC), v. 22, n. 1, p. 41-53, 2004.

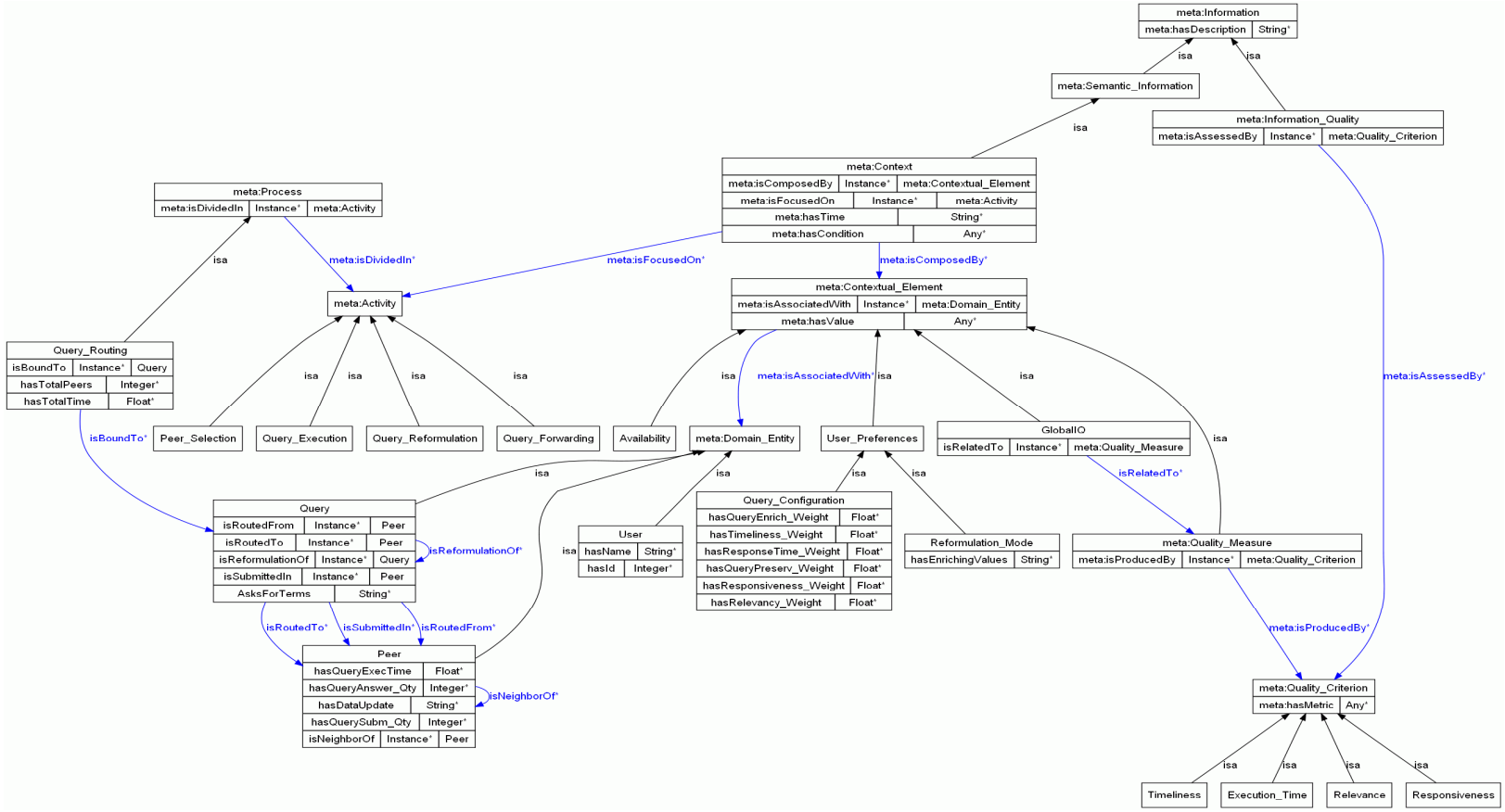
ZHU, Y.; BUCHMANN, A. **Evaluating and Selecting Web Sources as External Information Resources of a Data Warehouse**. In Proceedings of the 3rd International

Conference on Web Information Systems Engineering (WISE' 2002), Singapore, 2002.

ZHUGE, H.; LIU, J.; FENG, L.; SUN, X.; HE, C. **Query Routing in a Peer-To-Peer Semantic Link Network**. In IEEE Computational Intelligence, v. 21, n. 2, p. 197- 216, 2005.

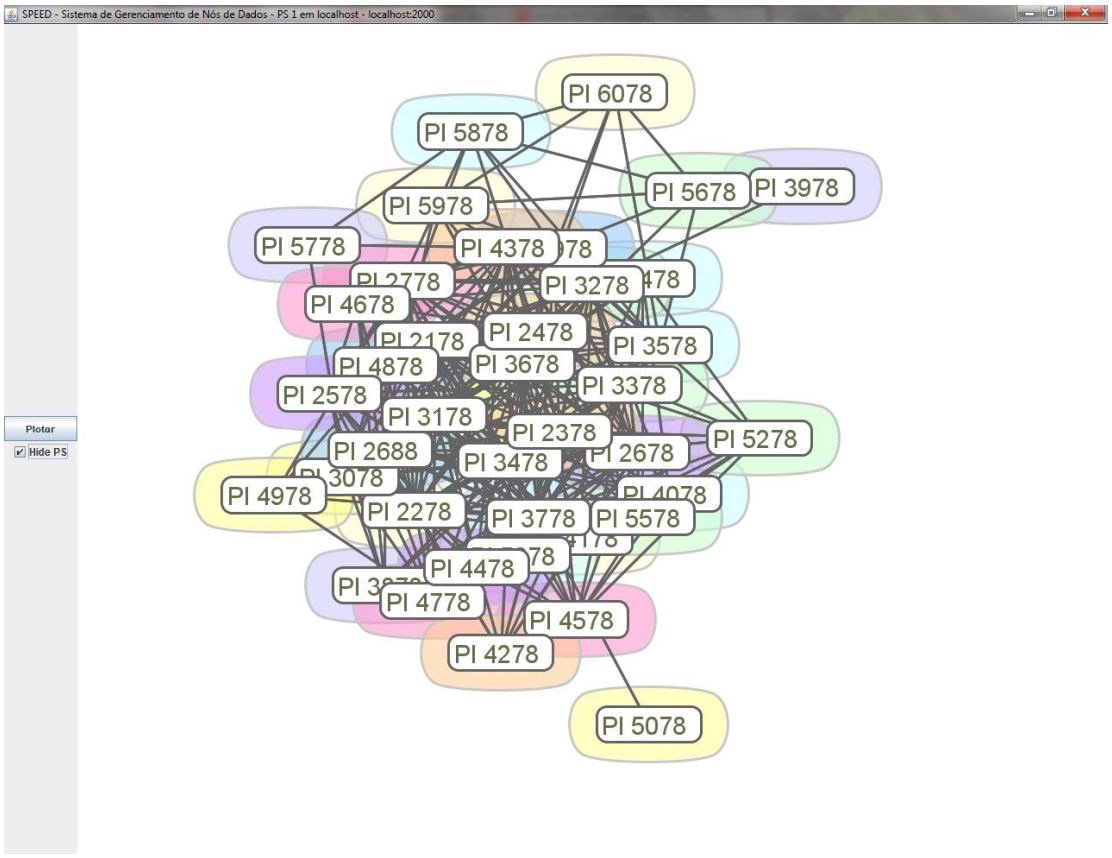
Apêndice A

Modelo para Representação das Informações Semânticas e de Qualidade da Informação (com detalhes das propriedades e relacionamentos)



Apêndice B

Cenário da rede para o Experimento 2.



Apêndice C

Este apêndice apresenta o experimento e a análise realizada para definição do ***threshold* de Relevância** utilizado na abordagem *SemRouting*.

Descrição:

A definição do *threshold* de relevância em 0.5 foi baseada nos valores estabelecidos atualmente paras as correspondências semânticas. Foi observado que, de acordo com os pesos atribuídos às correspondências, as distâncias semânticas para os termos *isSubConceptOf*, *isSuperConceptOf* e *isClosenessTo* em relação ao peso da correspondência de equivalência são de 0.2, 0.2 e 0.3, respectivamente, enquanto que para as correspondências *isPartOf* e *isWholeOf* é de 0.7 conforme mostra a Tabela 1.

Tabela 1 – Distância Semântica das Correspondências em Relação à Equivalência

Tipo de Correspondência	Peso	Distância Semântica em Relação à equivalência
isEquivalentTo	1	0
isSubConceptOf	0,8	0,2
isSuperConceptOf	0,8	0,2
isCloseTo	0,7	0,3
isPartOf	0,3	0,7
isWholeOf	0,3	0,7

Na reformulação *(SemRef)* cada termo original é expandido de acordo com as correspondências semânticas. Assim, é preciso considerar que o valor de relevância da consulta pode ser influenciado:

1. Pela quantidade de termos expandidos por termo original;
2. Pelo peso associado às correspondências que produziram os termos expandidos. Como mostra a Tabela 1, alguns tipos de correspondências possuem um maior distanciamento em relação à correspondência de equivalência.

Quando uma consulta reformulada tem todos os termos originais preservados, dizemos que o valor de relevância calculado para esta consulta tem peso

1. Por outro lado, quando uma consulta reformulada perde todos os termos originais, mas no enriquecimento obtém novos termos, dizemos que a relevância calculada para esta consulta depende do número de termos expandidos por termo original. Considerando os valores associados a cada correspondência e a distância produzida em relação à equivalência, definimos que a relevância mínima deve considerar a metade do valor de relevância da consulta, ou seja, 0.5.

Para verificar a aplicação deste *threshold*, consideramos uma rede formada por 6 pontos, organizados de forma sequencial (o roteamento será nesta ordem: P1, P2, P3, P4, P5) Neste ambiente foram analisados dois cenários de reformulação. Para cada cenário consideramos as reformulações produzidas e os valores de **preservação, enriquecimento e relevância** calculados individualmente para cada consulta reformulada. Para facilitar a explicação de como foi feita a análise das reformulações, as linhas da Tabela 2 foram numeradas (coluna linha). Cada grupo mostra a reformulação de um ponto origem para um ponto destino. Por exemplo, na linha 1, a consulta Q_1 submetida no ponto P1 foi reformulada em Q_2 de acordo com a ontologia do ponto P2. As colunas, preservação, enriquecimento e relevância contêm os valores calculados para cada coluna em relação a Q_2.

Em relação a Q_2 é possível observar que, em termos de preservação dos termos originais, apenas 0,33 do valor semântico da consulta foi preservado. Este valor é decorrente da perda dos demais termos originais e representa menos da metade do valor semântico da consulta para todos os termos preservados, que seria igual a 1. Em relação ao enriquecimento, Q_2 teve um enriquecimento de 0,36 que de forma isolada representaria, novamente, menos da metade do valor que corresponderia à consulta preservada. Entretanto, a relevância, isto é a soma das colunas preservação e enriquecimento, tem valor de 0,69 e mostra uma consulta semanticamente mais próxima da consulta original. Após análise dos cenários, **v**erificamos que as consultas com relevância (preservação+enriquecimento) com valor maior ou igual a 0.5 têm melhor aproximação semântica quando comparadas com a consulta original (Tabela 2 e Tabela 3).

Cenário 1

Qoriginal= ResearchGroup, Journal, Meeting **submetida em P1**

Referência Semântica de Qoriginal

$Q_{ref\_semântica}$ =**ResearchGroup** (1,0), Organization(0,8), Enterprise(0,3), Association(0,3), Program(0,3), Department(0,3), University(0,3), College(0,3), School(0,3), EduacationOrganization(0,3), Institute(0,3), **Journal**, Periodical(0,8), Article(0,1), Magazine(0,3), **Meeting** (1,0), Event(0,8), ProjectMeeting(0,8), Exhibition(0,3), Conference(0,3), Workshop(0,3), Lecture(0,3)

Tabela 2 – Reformulações do Cenário 1

Linhas	Consultas	Pontos	Preservação	Enriquecimento	Relevância
1	Q_1 = ResearchGroup, Journal, Meeting	**P1**			
2	Q_2= ResearchGroup (1,0), Organization(0,8), University(0,3), Event(0,8), ProjectMeeting(0,8), Conference(0,3), Lecture(0,3),	**P2**	**0,33**	**0,36**	**0,69**
3	Q_2= ResearchGroup (1,0), Organization(0,8), University(0,3), Event(0,8), ProjectMeeting(0,8), Conference(0,3), Lecture(0,3),	**P2**			
4	Q_3= Organization(0,8), Program(0,8), University(0,3), Event(0,8), ProjectMeeting(0,8), Conference(0,3), Lecture(0,3),	**P3**	**0**	**0,39**	**0,39**
5	Q_3= ResearchGroup (1,0), Organization(0,8), Program(0,8), University(0,3), Meeting(0,8), Event(0,8), ProjectMeeting(0,8), Conference(0,3), Lecture(0,3),	**P3**	**0,33**	**0,29**	**0,62**

6	Q_4 = Organization(0,8), College(0,3), University(0,3), School(0,3), Meeting(1,0), Event(0,8), Conference(0,3), Workshop(0,3),	**P4**			
7	Q_4 = Organization(0,8), College(0,3), University(0,3), School(0,3), Meeting(1,0), Event(0,8), Conference(0,3), Workshop(0,3),	**P4**	**0**	**0,14**	**0,14**
8	Q_5 = Organization(0,8), College(0,3), University(0,3), Institute(0,3),	**P5**			

Cenário 2

Qoriginal= Professor, Manual **submetida em P1**

Referência Semântica de Qoriginal

$Q_{ref\_semântica}$ =Professor, Faculty(0,8), AssociateProfessor(0.8), VisitingProfesor(0,8), AssistantProfessor(0.8), FullProfessor(0.8), Lecturer(0.7), PostDoc(0.7), Manual, Publication(0,8), Periodical(0,7), Article(0,7), UnofficialPublication(0,7), Proceedings(0,7), Software(0,7), Specification(0,7), Book(0,7), Thesis(0,7),

Tabela 3 – Reformulações do Cenário 2

Linhas	Consultas	Pontos	Preservação	Enriquecimento	Relevância
1	Q_1 = Professor, Manual	**P1**			
2	Q_2= Professor(1,0), VisitingProfesor(0,8), AssociateProfessor(0,8), AssistantProfessor(0,8), Faculty(0,8), Manual(1,0), Publication(0,8), Specification(0,7), Software(0,7)	**P2**	**1,0**	**0,76**	**1,76**

3	Q_2= Professor(1,0), VisitingProfesor(0,8), AssociateProfessor(0,8), AssistantProfessor(0,8), Faculty(0,8), Manual(1,0), Publication(0,8), Specification(0,7), Software(0,7)	**P2**			
4	Q_3= Professor(1,0), VisitingProfesor(0,8), AssociateProfessor(0,8), AssistantProfessor(0,8), Faculty(0,8), Manual(1,0), Publication(0,8), Specification(0,7), Software(0,7)	**P3**	**1,0**	**0,76**	**1,76**
5	Q_3= Professor(1,0), VisitingProfesor(0,8), AssociateProfessor(0,8), AssistantProfessor(0,8), Faculty(0,8), Manual(1,0), Publication(0,8), Specification(0,7), Software(0,7)	**P3**			
6	Q_4 = Professor, VisitingProfessor(0,8), AssistantProfessor(0,8), Faculty(0,8), Software(0,7)	**P4**	**0,5**	**0,75**	**1,25**
7	Q_4 = Professor, VisitingProfessor(0,8), AssistantProfessor(0,8), Faculty(0,8), Software(0,7)	**P4**			
8	Q_5 = Professor, Lecturer(0,7), PostDoc(0,7), Faculty(0,8), Publication(0,8), Article(0,7), UnofficialPublication(0,7), Book(0,7),	**P5**	**0,5**	**0,72**	**1,22**
9	Q_5 = Professor, Lecturer(0,7), PostDoc(0,7), Faculty(0,8), Publication(0,8), Article(0,7), UnofficialPublication(0,7), Book(0,7),	**P5**			
10	Q_6 = Faculty(0,8), AssociateProfessor(0.8), VisitingProfesor(0,8), FullProfessor(0.8), PostDoc(0.7), Publication(0,8), Article(0,7), UnofficialPublication(0,7), Software(0,7), Specification(0,7), Book(0,7),	**P6**	**0**	**0,74**	**1,74**

Apêndice D

Este apêndice apresenta outro experimento realizado e a análise dos resultados obtidos utilizando as abordagens: *Flooding*, como baseline, e a *SemRouting*.

- **Cenário**:

Uma rede formada por um conjunto de quarenta pontos (Figura D-1). Neste experimento, a rede foi gerada com um *threshold* de vizinhança de 0.74. Assim, a vizinhança semântica entre os pontos foi reduzida.

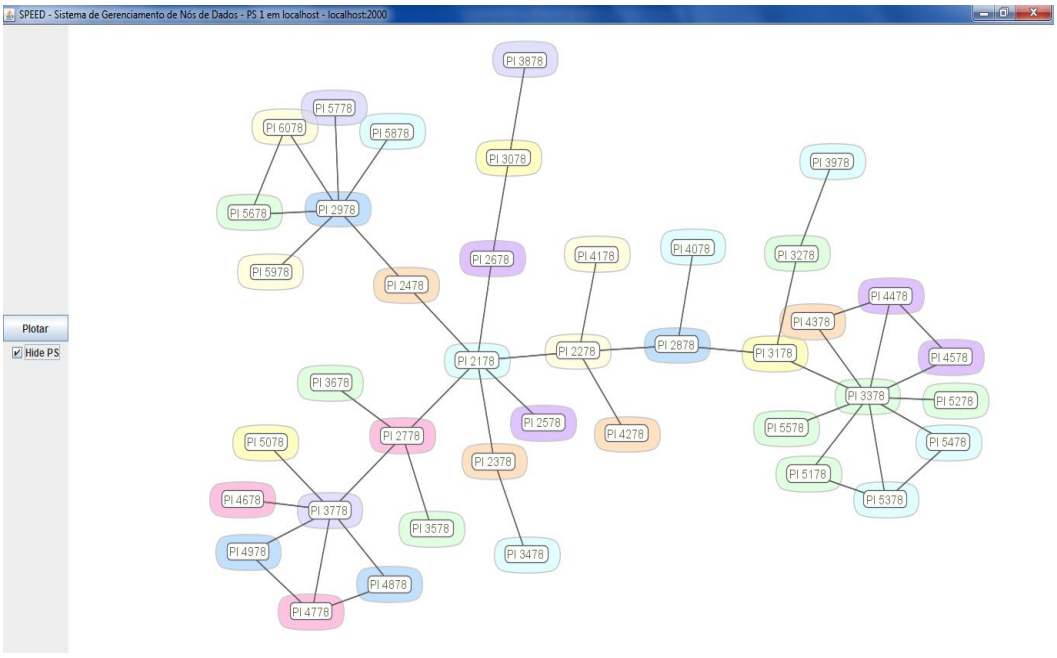


Figura D.1– Distribuição dos Pontos na Rede

O roteamento foi limitado em um número de saltos de valor igual a sete. O valor do *threshold* de relevância foi de 0.5 (justificado no Apêndice C). Neste cenário, uma consulta *Q={Professor* ∪ *Conference* ∪ *Article}* foi submetida no ponto P2178 com preferências de enriquecimento estabelecidas para as variáveis (*Generalization, Specialization, Approximate, Compose*).

Neste experimento observa-se que a rede foi percorrida em um maior número de saltos na *SemRouting* (Figura D.2). Este comportamento é decorrente da eliminação de pontos que não atendem ao *threshold* semântico e de qualidade, e de caminhos alternativos para o encaminhamento da consulta. No caso do *Flooding*, os

pontos são percorridos em menor salto devido a ausência de critérios que eliminem caminhos durante o roteamento.

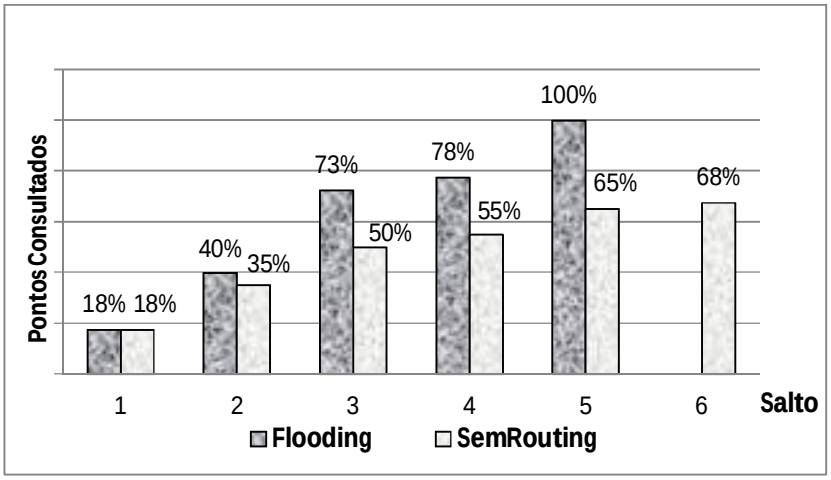


Figura D.2– Percentual acumulado de pontos consultados

No gráfico da Figura D.3, observamos que embora a estratégia por *Flooding* tenha retornado as consultas em um menor número de saltos, a média da similaridade semântica decresce a cada salto do roteamento. Como todas as variáveis de enriquecimento foram solicitadas, a ausência de uma estratégia de preservação permite a adição de um maior número de termos que não possuam correspondência semântica com os termos originais. No caso da *SemRouting*, as consultas são preservadas e a média da similaridade é mantida em um melhor nível mesmo no último salto.

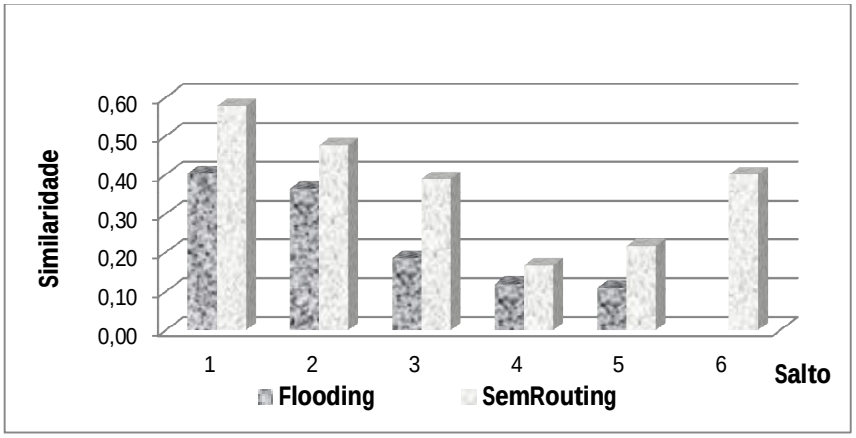


Figura D.3– Média da Similaridade entre as Consultas Reformuladas e a Referência Semântica

As Figuras D.4 e D.5 mostram gráficos comparativos em relação aos termos a cada reformulação nas duas abordagens (*SemRouting* e *Flooding*). Na *SemRouting*, enquanto a estratégia de preservação elimina termos que não possuam relação

semântica direta com os termos originais da consulta, no roteamento por *Flooding* a adição de termos interfere na semântica da consulta.

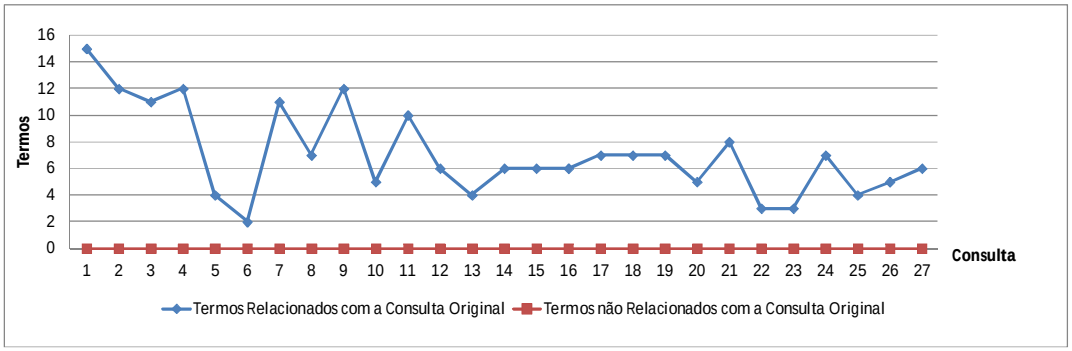


Figura D.4– Consultas Roteadas pelo SemRouting

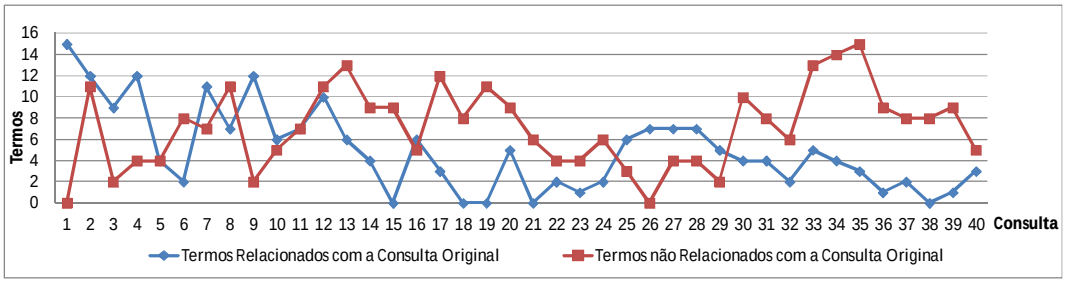


Figura D.5– Consultas Roteadas pelo SemRouting

O resultado deste experimento mostra um aspecto importante em relação ao roteamento da consulta. A forma como os pontos estão organizados e tem sua vizinhança semântica definida tem impacto direto na estratégia utilizada. Influenciando na seleção e no número de pontos que possam ser consultados ao longo do roteamento.