# Disseminação de Conteúdo XML Baseada em Ontologias

**Mirella M. Moro[1*], Renata Galante[2*], Deise de B. Saccol[3], Bernadette F. Loscio[4]**

[1]Universidade Federal de Minas Gerais (UFMG), Belo Horizonte – MG – Brazil

[2]Universidade Federal do Rio Grande do Sul (UFRGS), Porto Alegre – RS – Brazil

[3]Universidade Federal do Pampa (UNIPAMPA), Alegrete – RS – Brazil

[4]Universidade Federal do Ceara (UFC), Fortaleza – CE – Brazil

mirella@dcc.ufmg.br, galante@inf.ufrgs.br,
deisesaccol@gmail.com, bernafarias@lia.ufc.

***Abstract.*** *As Internet and distributed systems evolve, a new paradigm aggregates the concept of content dissemination to XML query engines. The query is still evaluated over the stored data, but it is also registered into the system. Then, those queries will also be evaluated over the incoming data such that the documents that satisfy them are disseminated back to the users. In such a context, this paper proposes, that ontologies be applied in order to improve the performance of content-based dissemination systems. Our initial experimental evaluation shows that such solution is viable and exhibits considerable advantage over the state-of-the-art techniques.*

***Resumo.*** *Com a evolução da Internet e de sistemas distribuídos, um novo paradigma agrega o conceito de disseminação de conteúdo aos sistemas de consulta de SGBDs XML. Além de responder às consultas sobre os dados armazenados, as consultas também são mantidas no sistema, e a avaliação é realizada sobre os novos dados que são adicionados ao SGBD. Neste contexto, este artigo propõe, a utilização de ontologias para melhorar o desempenho de sistemas de disseminação de conteúdo. Nossos resultados iniciais mostram que tal solução é viável e apresenta considerável vantagem sobre os mecanismos do estado-da-arte.*

## 1. Introdução

Em sistemas de gerenciamento de bancos de dados (SGBDs) XML, um usuário envia uma consulta ao sistema, que por sua vez a avalia sobre seus dados armazenados e retorna os resultados de tal avaliação ao usuário. Com a evolução da Internet e de sistemas distribuídos, um novo paradigma agrega o conceito de *disseminação de conteúdo* aos sistemas de consulta. Além de responder às consultas considerando os dados armazenados, as consultas também são armazenadas no sistema, e a avaliação continua sendo realizada à medida que novos dados são adicionados ao SGBD. Neste caso, os resultados são disseminados aos usuários *a posteriori*.

Essa forma de consulta é amplamente utilizada nos serviços de disseminação de informação baseados em conteúdo (*content-based information dissemination services*),

* Pesquisa parcialmente financiada pelos projetos INCT Web (CNPq no. 573871/2008-6), CTINFO 550891/2007-2 e Amanajé (CNPq no. 479541/2008-6).

ou simplesmente sistema de disseminação de conteúdo. Este novo paradigma de avaliação de dados tem criado oportunidades para novas aplicações tais como vários serviços de alerta e notificação que informam o usuário interessado em novos produtos no mercado, atualizações da bolsa de valores, variação do valor de moedas estrangeiras, ofertas de negócios, governo eletrônico e assim por diante. Além disso, com a expansão de serviços Web (*web services*), novos sistemas desses são lançados frequentemente.

Com o reconhecimento de XML como o padrão para troca de dados, serviços de disseminação de conteúdo próprios para gerenciar dados XML são necessários [Diao et al. 2004; Snoeren et al. 2001]. Em sistemas de disseminação XML, a transmissão de mensagens é realizada por uma rede sobreposta (*overlay network*) sofisticada formada por roteadores baseados em conteúdo (*content-based routers*), que são chamados de roteadores XML (*message brokers*). Tais roteadores avaliam os perfis do usuário (consultas) sobre as mensagens e encaminham as mensagens (baseado na avaliação) aos seus destinos, ou seja, outros roteadores ou usuários. A tarefa de avaliar perfis e mensagens é chamada de *filtrar mensagens*.

É importante notar que sistemas de disseminação de conteúdo estão em constante atualização. Novas consultas são requisitadas e novos dados são adicionados. Desse modo, a principal característica de um sistema desse porte (do ponto de vista de banco de dados) é a escalabilidade em termos da quantidade de consultas e do volume de dados que podem ser processados ao mesmo tempo. Além disso, outra característica fundamental é a capacidade de processar milhares de consultas de maneira eficiente (em termos de tempo de processamento). Note que a avaliação de consultas é no estilo de filtragem, no qual as consultas que satisfazem um documento são identificadas. Neste contexto, geralmente emprega-se alguma forma de processamento de consultas múltiplas (*multi query processing*) na qual um conjunto de consultas é avaliado ao mesmo tempo (ao invés de processar cada consulta individualmente).

Uma solução clássica para a escalabilidade e a eficiência de sistemas é particionar o seu processamento (no nível da arquitetura, do armazenamento, entre outros). Especificamente, para o processamento de consultas em grandes volumes de dados, pode-se considerar tanto particionar os dados quanto o conjunto de consultas. A grande questão é como realizar esse particionamento de maneira eficiente (ou seja, de modo que melhore tanto a escalabilidade e o tempo de processamento de consultas). Além disso, a escolha da arquitetura de rede utilizada no sistema de disseminação de conteúdo pode influenciar no seu processamento como um todo. Uma opção é utilizar a estrutura fornecida por redes *peer-to-peer* (P2P) [Zhu and Hu 2007]. Dentre as arquiteturas possíveis, a rede P2P tem as vantagens de reduzir o congestionamento da rede, minimizar a profundidade de roteamento e preservar essas características no caso de alterações e falhas na rede.

Entretanto, documentos de um mesmo domínio geralmente estão *espalhados* pela rede P2P, o que pode influenciar o desempenho da filtragem de mensagens. Uma solução é utilizar o conceito de ontologias para agrupar documentos de mesmo domínio em pares específicos [Saccol et al. 2008b]. Esse agrupamento pode ser coordenado pela especificação de *super-peers* que mantêm a informação sobre as ontologias e os documentos de seu domínio.

Este artigo foca nos aspectos de disseminação de documentos, distribuição de consultas e processamento dessas consultas em sistemas de disseminação de conteúdo XML. Essas três tarefas utilizam uma estrutura homogênea baseada em ontologias, chamada Tricot (*TRIple Content-based OnTology*). É importante notar que é a *primeira* vez que tal homogeneidade é proposta para um sistema de tamanha complexidade. Desse modo, as contribuições deste artigo são as seguintes.

- Este artigo considera o contexto de sistemas de disseminação de conteúdo XML e discute vários cenários recentes e relevantes que mostram a versatilidade e a complexidade de tais sistemas (seção 2). Esses cenários podem servir de casos de uso para a Tricot e apresentam importantes requisitos próprios que podem ser futuramente explorados.
- Para melhorar a escalabilidade em relação ao número de documentos avaliados, este artigo propõe que seja usada uma rede P2P como topologia base para os roteadores, distribuindo assim o processamento dos documentos. O sistema de disseminação utiliza as ontologias para agrupar documentos de mesmo domínio no mesmo *peer*, ou conjunto de *peers* (seção 3).
- Para melhorar a escalabilidade em relação ao número de consultas, este artigo propõe a utilização de ontologias para o particionamento, a distribuição e a avaliação das consultas. O objetivo é que um documento seja avaliado considerando apenas as consultas definidas no seu mesmo domínio, em vez de considerar o universo completo de consultas (seção 4).
- Uma avaliação experimental inicial demonstra a viabilidade da Tricot e as vantagens sobre os algoritmos do estado-da-arte. Além disso, os resultados mostram o potencial da nossa solução para o emprego de ontologias em sistemas de disseminação de conteúdo XML sobre redes P2P (seção 5).

O restante do artigo está organizado da seguinte forma. A seção 6 apresenta trabalhos relacionados aos diversos temas abordados no artigo: disseminação de conteúdo, redes P2P e ontologias. A seção 7 encerra o artigo com as considerações finais e trabalhos futuros.

## 2. Disseminação de Conteúdo

Sistemas de disseminação de conteúdo utilizam um modelo de interação baseado em eventos que funciona de maneira oposta ao tradicional modelo de *request/reply*. Quem inicia a comunicação é o provedor dos dados (e não o consumidor), e os papéis de consulta e dados são invertidos. No modelo *request/reply*, uma consulta é processada sobre os dados para identificar qual parte dos dados satisfaz à consulta. Em sistemas de disseminação, um conjunto de consultas é processado sobre um conjunto de dados para identificar quais consultas são satisfeitas pelos dados. Neste contexto diferenciado, os dados são definidos por provedores, são recebidos através de *streams* de mensagens, são processados online (*on the fly*) ou em grupo (*batches*), e enviados aos consumidores. A transmissão de mensagens é realizada por uma rede sobreposta (*overlay network*) formada por roteadores próprios. Tais roteadores avaliam os perfis do usuário (consultas) sobre as mensagens e encaminham as mensagens (baseado na avaliação) aos seus destinos, ou seja, outros roteadores ou clientes.

## 2.1. Aplicações

Entre os cenários para sistemas de disseminação de conteúdo, destacam-se os seguintes.

*Bibliotecas Digitais.* Bibliotecas Digitais focam nos problemas de procurar uma informação, entregá-la ao usuário e preservá-la para o futuro, e podem ser estáticas ou dinâmicas [Meghini and Spyratos 2007]. Em bibliotecas dinâmicas, os usuários estão interessados nos dados armazenados e nos novos dados adicionados. Sistemas de disseminação se encaixam nesse contexto, ou seja, notificação de novas informações.

*Indústria de Seguros.* Geralmente, escritórios de companhias de seguro são conectados através de uma rede de roteadores de conteúdo [Li et al. 2007]. As mensagens publicadas incluem reivindicações de apólices, ofertas e propostas de seguros. Essas mensagens são distribuídas a especialistas (cujos interesses foram definidos através do seu perfil) que irão contactar os clientes conforme necessário.

*Avisos de Segurança.* Técnicas de contenção automática de *worms* analisam o tráfego de rede e definem um pacote classificador que bloqueia ou limita o encaminhamento de pacotes contaminados [Costa et al. 2005]. As mensagens encaminhadas são pacotes na rede, e as consultas são a presença de *worms* conhecidos. O aplicativo funciona como um sistema de disseminação inverso, no qual as mensagens (pacotes) que satisfazem às consultas (características dos *worms*) *não* são encaminhadas (essas mensagens são colocadas em quarentena, por exemplo).

*Mercado de Ações.* Uma das aplicações mais conhecidas é o mercado de ações, que possibilita a negociação e a troca de ações, opções e debêntures. As mensagens contêm informações financeiras, e as consultas refletem os interesses dos investidores.

*Outras Aplicações*. Entre outras aplicações populares, podemos ainda citar *feeds* RSS e disseminação de notícias. Por exemplo, anúncios em páginas da Internet e atualizações de blogs são tipicamente disseminados através de RSS; leitores de jornais eletrônicos querem ser informados de qualquer notícia que seja publicada de acordo com seus interesses, sem importar onde e como a notícia é publicada.

O trabalho de pesquisa e desenvolvimento propostos aqui se caracterizam não somente por sua atualidade científica, mas também pelo fato de refletirem necessidades industriais reais, conforme exemplificado. Essa lista de aplicações não é exaustiva. Existe uma ampla aplicabilidade para sistemas de disseminação nos mais diversos tipos de indústria. Todas essas aplicações têm um requisito comum que é a capacidade e o desempenho do processamento dos dados, bem como das consultas sobre os mesmos.

## 2.2. Tendência e Relevância

Cabe também salientar que as mais importantes conferências internacionais de banco de dados recentemente incluíram assuntos relacionados a disseminação em seus tópicos de interesse (tais como *Information Filtering and Dissemination, Replication, caching, and publish-subscribe systems* e *XML data processing, filtering, routing, and algorithms*). Enquanto as soluções atuais se baseiam apenas nos dados e nas consultas armazenadas, nós propomos que sejam utilizados mecanismos de inteligência em todas as fases do processo. Uma maneira de adicionar inteligência é considerar ontologias tanto no processamento dos dados quanto no das consultas. Esse casamento de técnicas de inteligência artificial com banco de dados não é uma idéia nova, porém, o que

propomos é estender técnicas consagradas para serem aplicadas a este novo domínio de serviços de disseminação de conteúdo. Até onde sabemos, esta é a *primeira* vez que tal casamento (entre inteligência artificial e disseminação de conteúdo) é proposto.

É importante enfatizar que as questões de pesquisa relacionadas aos sistemas de disseminação de conteúdo estão fortemente relacionadas à gestão da informação em grandes volumes de dados distribuídos, o qual constitui o primeiro dos *Grandes Desafios da Pesquisa em Computação no Brasil da SBC*. Especificamente, o foco do primeiro desafio está no tratamento, na recuperação e na disseminação de informação relevante a partir de grandes volumes de dados multimídia. Neste contexto, este artigo especifica melhor uma das contribuições do desafio original. Além disso, este trabalho também estende esse desafio original para a inclusão de aspectos de inteligência artificial no mecanismo de disseminação, nesse caso, o uso de ontologias em vários aspectos do sistema. Dada a ampla aplicabilidade de sistemas de disseminação de conteúdo, espera-se que a melhoria das aplicações existentes tenha impacto social relevante em diversas áreas, como, por exemplo, bibliotecas digitais, serviços de saúde, disseminação de notícias, acesso universal à informação, entre outros.

## 3. Gerenciamento de Domínios de Aplicação com o uso de Ontologias

A escolha da arquitetura de rede utilizada no sistema de disseminação influencia no seu processamento global. Uma opção é utilizar a estrutura fornecida por redes P2P [Zhu and Hu 2007] cuja arquitetura de processamento distribuído reduz o congestionamento da rede, minimiza a profundidade de roteamento e preserva tais características no caso de alterações e falhas de rede. Porém, documentos XML de um mesmo domínio são espalhados pela rede P2P, o que pode influenciar o desempenho como um todo.

Neste trabalho, documentos relacionados a um mesmo domínio são agrupados e o processamento de consultas é realizado somente nos documentos pertencentes a este domínio. O conceito de ontologias é empregado para agrupar documentos de um mesmo domínio em *peers* específicos [Saccol et al. 2008b]. Em resumo, o processo de gerenciamento de ontologias encapsula a heterogeneidade dos recursos, fornecendo informações sobre as ontologias existentes e os documentos. Esse processo também é responsável por gerar novas ontologias quando nenhuma das existentes é apropriada para um certo (conjunto de) documento(s) XML. A geração da ontologia é feita a partir da integração dos esquemas dos documentos envolvidos. O casamento entre documentos e ontologias é responsável pela identificação da ontologia que melhor descreve um documento. Finalmente, o processador de consultas permite também ao usuário escolher uma ontologia e formular consultas com base nesta ontologia.

O uso de domínios de aplicação fornece duas vantagens principais: *(i)* restringe o espaço de busca, pois a busca é realizada considerando apenas os documentos de mesmo domínio; e *(ii)* aumenta a eficiência no processamento de consultas, uma vez que consultas relativas a um determinado domínio de aplicação não são processadas sobre documentos que não possuem conceitos pertencentes a este domínio. Esta seção segue com a apresentação do mecanismo para construção de ontologias a partir de documentos XML. Considerando que a ontologia exista, o próximo passo é realizar o casamento dos documentos XML com as ontologias disponíveis. E, finalmente, documentos de mesmo domínio são encaminhados para os *peers* daquele domínio.

### 3.1. Abordagem para Geração de Ontologias

A abordagem para geração de ontologias visa obter uma ontologia que descreva os conceitos e relacionamentos existentes em um domínio de aplicação. A ontologia é criada a partir da integração de esquemas, nos casos onde não exista uma ontologia de domínio conhecida e que possa ser reutilizada. Primeiramente, gera-se o esquema XML correspondente ao documento compartilhado. O esquema XSD (*XML Schema Definition*) é então traduzido para um formato OWL (*Ontology Web Language*), o qual enfatiza os elementos complexos (i.e., não léxicos) e atômicos (i.e., léxicos). Através da aplicação de várias regras de integração, um esquema integrado é gerado, também representado em OWL. A geração de ontologias está implementada na ferramenta Ontogen, cujos detalhes podem ser encontrados em [Saccol et al. 2008a].

### 3.2. Casamento de Ontologias e Documentos XML

Com o conjunto de ontologias (por exemplo, geradas no passo anterior), é necessário identificar uma ontologia para o documento sendo avaliado no sistema de disseminação. Esse casamento tem como objetivo escolher a ontologia que melhor descreve os conceitos e relacionamentos de um documento XML.

***Definição 1.*** *(Casamento de Ontologia e Documento XML).* Dado um documento XML *d* e um conjunto de ontologias $O = o_1,.., o_n$, o mecanismo de casamento processa um escore de similaridade *sim(d,O)*. A ontologia com o escore mais alto (desde que superior a um limiar *t*) é escolhida para representar o domínio de aplicação de *d*.

A estratégia proposta objetiva encontrar semelhanças entre os elementos do XML e os elementos da ontologia. Duas questões devem ser consideradas: quais pares de elementos são comparados e quais são os critérios para determinar o quão similares são tais elementos. Várias abordagens tratam da análise de similaridade [Madhavan et al. 2001, Maedche et al. 2002]. Especificamente, para avaliar a similaridade entre os documentos, dois tipos de perspectivas são considerados: a perspectiva léxica avalia as relações entre termos, comparando as *strings* e a perspectiva semântica foca no significado e na relação entre os termos. Para a análise de similaridade, este trabalho considera ambas as perspectivas, conforme descrito a seguir.

**Análise de similaridade léxica.** Duas abordagens são utilizadas nesta análise: *(i)* Funções de edição de distância [Levenshtein 1996], que analisam o número mínimo de operações para transformar uma seqüência de caracteres em outra; e *(ii)* Algoritmos de extração de radicais[1], que reduzem a seqüência de caracteres ao radical. Uma vez que a taxonomia dos elementos é definida a *priori* em nossa abordagem, este trabalho utiliza algoritmos de extração de radicais para análise de similaridade léxica.

**Análise de similaridade semântica.** Dois tipos de abordagens são geralmente utilizados nesta análise: *(i) Thesaurus*, para avaliar os relacionamentos terminológicos (por exemplo, *WordNet*[2]); *(ii)* Procedimento de sobreposição de taxonomias [Maedche et al. 2002], que compara a taxonomia entre elementos. Esta comparação não analisa individualmente os elementos, mas o contexto do elemento. Para calcular o nível de

[1] Lancaster Stemming Algorithm – HTTP://www.comp.lancs.ac.iuk/computing/research/stemming

[2] HTTP://wordnet.princeton.edu

similaridade entre dois conjuntos de elementos, pode-se utilizar o coeficiente *Jaccard* [Manning and Schutze 1999]. O mecanismo proposto neste trabalho agrega e estende as vantagens de algumas abordagens existentes, como descrito na próxima subseção.

### *3.2.1 Abordagem para o Casamento*

Para o cálculo do escore de similaridade *sim(d, o)* entre um documento XML *d* e uma ontologia *o* (apresentado em Definição 1), utiliza-se o seguinte mecanismo. A primeira fase corresponde à normalização (determina quais elementos são semanticamente equivalentes) e categorização (separa os elementos em classes para reduzir o número de comparações). Para cada documento (XML e OWL), mapeia-se todos os elementos componentes e armazena-se o radical (chave) e uma lista com os nomes completos de cada elemento. Se o nome do elemento é composto por n-palavras, então a lista também consiste de cada componente. Este mapeamento inicial corresponde à perspectiva léxica e é exemplificado na Figura 2(a). O próximo passo considera os sinônimos dos elementos (cuja lista é recuperada do *WordNet*), conforme o resultado na Figura 2(b).

(a)

| Key | List | | |
|---|---|---|---|
| skill | skillArea | skill | área |

(b)

| Key | List | | | | | |
|---|---|---|---|---|---|---|
| skill | skillArea | skill | área | accomplishment | acquisition | domain |

**Figura 2. Normalização léxica e normalização semântica de elementos.**

Essa primeira fase ocorre em dois passos. Primeiro, normalizam-se os elementos da ontologia e, em seguida, percorrem-se os elementos XML a fim de verificar em suas listas a existência de qualquer correspondência léxica com outros elementos da lista OWL. Estas correspondências são analisadas observando-se apenas os radicais, utilizando um extrator de radicais. Finalmente, tem-se disponível duas listas categorizadas e normalizadas (XML e OWL). Nesta fase, o sistema está apto a identificar as correspondências existentes entre os elementos XML e OWL, fornecendo o passo inicial para a comparação de elementos.

A segunda fase para o casamento é a comparação, a qual analisa a sobreposição de taxonomias. Não são considerados os elementos individualmente, mas o seu contexto. Para cada elemento que consiste de um conjunto *root-leaf* (isto é, o caminho da raiz aos elementos folha), o sistema conhece a existência de correspondências na ontologia (obtido durante a fase de normalização). Para obter o valor de similaridade, define-se união como o número total de elementos XML do conjunto *root-leaf*. A intersecção é definida como o número de relações com correspondência entre os elementos XML e da ontologia. Por relação, entende-se: uma classe OWL que se relaciona a uma subclasse OWL; uma classe OWL que se relaciona a propriedades *Object* ou *Datatype*; e relações entre classes OWL. Esse método é denominado *Taxonomic Overlap* (MAEDCHE, 2002), que corresponde à comparação taxonômica entre os elementos avaliados.

O método de sobreposição de taxonomias considera duas árvores (chamadas como a da esquerda e a da direita), então percorre a árvore da esquerda e analisa somente os nós que têm correspondência com a árvore da direita, isto é, é feita uma verificação para identificar se os elementos que são lexicamente similares na ontologia possuem equivalência semântica. Este procedimento é repetido para todos os conjuntos

*root-leaf*. Desta forma, obtém-se uma lista com todos os valores de similaridade. O valor final é calculado como a média aritmética desta lista. Em resumo, documentos relacionados a um mesmo domínio são agrupados (Definição 2) para que o processamento de consultas possa ser é realizado somente nos documentos pertencentes a cada domínio específico.

***Definição 2.*** *(Agrupamento de Documentos).* Seja $O$ o conjunto de ontologias $o_i$ pertencentes ao super *peer SP* e $d$ um documento XML. Seja $T$ (em $SP$) a tabela de distribuição dos domínios dos *peers* na qual cada linha $i$ é composta por tuplas [$o_i$, $dom_i$], onde *dom* é uma lista de *peers* para o respectivo domínio. Para cada ontologia $o_i$, esta operação identifica a ontologia $o'$ que apresenta o maior grau de similaridade com $d$, conforme Definição 1. Então, $SP$ consulta $T$ e encaminha $d$ para os *peers* $p$ em $dom_i$.

### 3.3. Evolução de Ontologias

Este trabalho propõe o uso de ontologias para representação semântica dos esquemas XML que definem a estrutura dos documentos que estão sendo compartilhados através do sistema de disseminação de conteúdo. É importante ressaltar que estes sistemas estão em constante atualização, ou seja, novas consultas são requisitadas e novos dados são adicionados. Estas atualizações devem ser propagadas para garantir a consistência entre os documentos compartilhados e as ontologias de domínio utilizadas pelo sistema.

A adição de novos documentos ou alterações nos esquemas XML já existentes podem ocasionar atualizações nas ontologias de domínio associadas, uma vez que estas ontologias são geradas a partir da integração das ontologias extraídas dos esquemas XML. Torna-se necessário, portanto, investigar o problema de *manutenção do sistema* em função de atualizações ocorridas nos esquemas dos documentos compartilhados. Atualmente, estamos trabalhando para definir uma solução viável para esta questão.

## 4. Tricot – TRIple Content-based OnTology

Esta seção apresenta a `Tricot` (*TRIple Content-based OnTology*), uma proposta baseada em ontologias para disseminação de documentos, distribuição de consultas, e processamento dessas consultas em sistemas de disseminação de conteúdo XML.

### 4.1. Processamento de Consulta XML com Ontologia

O princípio básico da `Tricot` é agrupar as consultas (e os documentos) de acordo com um mesmo domínio. Este agrupamento é realizado através do casamento entre a consulta e as ontologias. O processo de casamento é similar à Definição 1. Uma consulta XML é uma expressão de caminho composta por restrições estruturais, ou seja, nomes dos elementos XML (ou atributos) e seus relacionamentos (pai/filho e ancestral// descendente[3]). Adicionalmente, esta expressão de caminho pode conter restrições de valor ou predicados, ou seja, restrições sobre os valores dos elementos e atributos XML.

As restrições estruturais de uma consulta XML (com a *XPath* por exemplo) podem ser representadas por uma árvore de consulta, formando uma representação

[3] Outros relacionamentos, como "próximo filho", podem ser reduzidos a combinações de pai/filho e ancestral//descendente, conforme a especificação da linguagem XPath [Berglund et al. 2007].

*parcial* de um documento XML. Neste caso, os relacionamentos *ancestral//descendente* são considerados um caso especial de *pai/filho* (ou seja, o requisito de que o elemento filho tem de estar no nível seguinte ao do elemento pai é aliviado para considerar qualquer nível abaixo na sub-árvore do elemento pai). Desse modo, a operação de casamento entre a consulta e a ontologia pode ser considerada como um caso especial do casamento de um documento XML e uma ontologia, conforme a definição a seguir.

***Definição 3.*** *(Casamento de Ontologia e Consulta).* Seja *q* uma consulta XML e *o* uma ontologia. A operação de casamento entre *q* e *o* retorna o percentual de similaridade entre *q* e *o*, no qual os relacionamentos de *ancestral//descendente* são considerados um tipo especial do relacionamento *pai/filho*.

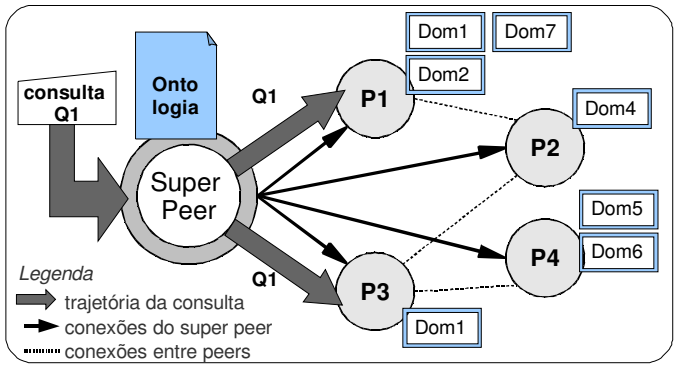



**Figura 3. Elementos da rede *peer*-to-*peer* – super-*peer*, *peer*s comuns ($P_1$ a $P_4$ com seus domínios ($Dom_1$ a $Dom_7$) -- e trajetória da consulta $Q_1$de acordo com os domínios e a ontologia.**

A Figura 3 mostra como consultas são distribuídas. A consulta é recebida pelo super *peer* que realiza o casamento com o seu conjunto de ontologias. Com o domínio da consulta definido, o super *peer* encaminha a consulta para todos os *peer*s de mesmo domínio. Aqui, a consulta $Q_1$ pertence ao domínio $Dom_1$ e é enviada aos *peer*s $P_1$ e $P_3$.

O *peer* contém apenas os conjuntos de consultas que se referem aos mesmos domínios dos seus documentos, então o número de consultas a serem avaliadas é consideravelmente reduzido. Em vez de o *peer* processar todo o seu conjunto de consultas em cada documento, o *peer* pode agora processar apenas o sub-conjunto de consultas do mesmo domínio do documento. Qualquer algoritmo de avaliação de consultas em documentos XML pode ser utilizado, conforme discutido na Seção 5.

### 4.2. Disseminação de Conteúdo com Tricot

Em sistemas de disseminação de conteúdo, a avaliação de consultas é no estilo de *filtragem*, no qual as consultas que satisfazem um documento são identificadas. É importante notar que, geralmente, emprega-se alguma forma de processamento de consultas múltiplas (*multi query processing*) na qual um conjunto de consultas é avaliado ao mesmo tempo (ao invés de processar cada consulta individualmente). A Definição 4 apresenta a operação de filtro de mensagens considerado neste trabalho.

***Definição 4.*** *(Filtro de Mensagens).* Seja *D* um conjunto de documentos XML recebidos através de um *stream* (infinito) de documentos $d_i$, $i=1,...$ Seja *Q* um conjunto de consultas $q_i$, $i=1..n$. A operação de filtro de mensagens identifica o sub-conjunto de *Q*, denominado *Q'*, que contém as consultas satisfeitas por um documento $d_i$.

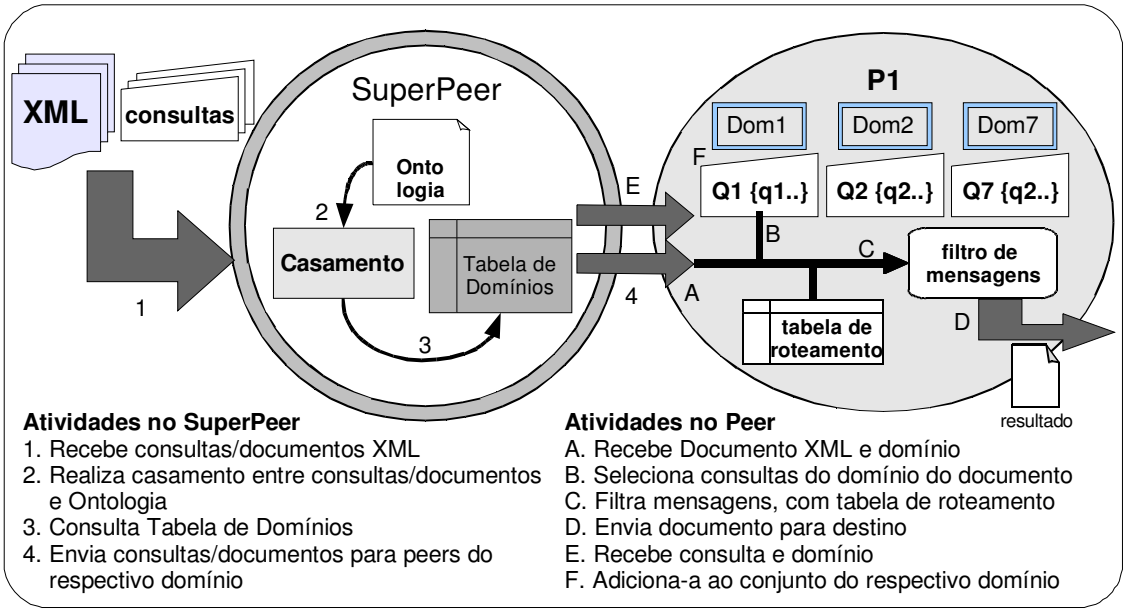



**Figura 4. Disseminação de conteúdo com a Tricot.**

Sendo a filtragem de mensagens a operação central do sistema de disseminação, este artigo foca nos aspectos de distribuição de documentos e de consultas bem como no processamento dessas consultas de modo a melhorar o desempenho do sistema como um todo. Essas três tarefas utilizam a Tricot. Definição 5 resume as principais operações envolvidas.

***Definição 5.*** *(Operações da Tricot).* Seja um sistema de disseminação *SD* funcionando sobre uma rede P2P composta por *peer*s $p_i$, $i=1..m$, e um super-*peer SP*. São elementos de *SD:D* o conjunto de documentos XML recebidos através de um *stream* (infinito) de documentos $d_i$, $i=1.$; $Q$ o conjunto de consultas $q_i$, $i=1..n$ armazenado em $p_i$; e $O$ o conjunto de ontologias $o_i$, $i=1..r$, armazenado em *SP*. A Tricot abrange as seguinte operações:

- Distribuição de documentos de $D$ segundo $O$ entre os *peer*s $p_i$. Para cada documento $d_i$, o seu domínio é identificado através do casamento com $O$, o documento é enviado a todos os *peer*s $p_i$ que contêm outros arquivos de mesmo domínio (segundo a Definição 2).
- Agrupamento de consultas em $Q$ segundo $O$, conforme a Definição 1.
- Filtro de mensagens, conforme Definição 4, realizado em cada *peer*.

Para clarificar essas operações, a Figura 4 ilustra as atividades da Tricot. O super *peer* é encarregado de receber os documentos XML, identificar o domínio de cada um (através do casamento com as ontologias) e enviá-los para os respectivos *peer*s de acordo com a tabela de domínios. Os *peer*s recebem cada documento XML, identificado com o respectivo domínio, selecionam o conjunto de consultas daquele domínio, avaliam as consultas no documento (através do filtro de mensagens) e enviam os resultados para os usuários ou demais *peer*s de acordo com a tabela de roteamento.

## 5. Avaliação da Viabilidade da Tricot

Esta seção apresenta uma avaliação *inicial* da Tricot, cujas operações foram implementadas em um simulador construído especificamente para tal fim. O objetivo é simular o comportamento do particionamento de documentos e consultas de acordo com as ontologias. Esta avaliação experimental foca nos aspectos de viabilidade da Tricot, em vez de seu desempenho (cuja avaliação será publicada em trabalho futuro).

### 5.1. Descrição Geral

Para avaliar o impacto da utilização de ontologias no processamento como um todo, foi implementado um simulador do sistema de disseminação de conteúdo a fim de estudar empiricamente a viabilidade das idéias propostas. Este simulador reflete o comportamento do super *peer* e dos *peer*s do sistema. É importante notar que não são considerados os aspectos da rede per se (disponibilidade, confiabilidade, tempo de transmissão de documentos e consultas, e assim por diante), pois o foco deste trabalho está na avaliação dos dados e das consultas. Os algoritmos foram implementados em Java, usando Sun JDK versão 1.4.0. Os experimentos foram realizados em uma máquina com processador Intel Pentium IV, 2.6GHz, com 1GB de memória.

### 5.2. Qualidade do Casamento Ontologia-XML

Para avaliar a distribuição dos documentos de acordo com as ontologias, optou-se por realizar um estudo de caso sobre a qualidade do casamento entre ontologias e documentos XML. Os parâmetros de entrada são os documentos XML, a ontologia definida em OWL e um limiar de similaridade mínimo. Neste estudo consideramos o limiar de 70%, o qual foi ajustado manualmente com base nas ontologias e documentos utilizados nos experimentos, com o objetivo de deixar a função flexível a ajustes. Porém, cabe ressaltar que definições adequadas de valores de limiar são discutidas na literatura [Stasiu et. al, 2005] e não fazem parte do escopo do trabalho apresentado neste artigo.. A informação de saída deste experimento é o grau de similaridade entre dois documentos[4]. As ontologias utilizadas definem domínios distintos, incluindo ontologias para pacotes de turismo, *currículo/carreira* profissional e vinhos. Para a avaliação, foi considerado um conjunto de dados com 10 mil documentos XML, sendo mil destes gerados a partir de diferentes taxonomias para descrição de atividades de turismo e cem gerados a partir de currículos da plataforma Lattes (*http://lattes.cnpq.br*).

Os resultados obtidos mostram um grau de similaridade de 96,1% para os mil documentos sobre turismo e a sua ontologia. Em outras palavras, os mil documentos gerados a partir de diferentes taxonomias de turismo são lexicamente similares. Os demais documentos XML (9000 documentos) apresentaram um grau de similaridade praticamente nulo em relação à ontologia de turismo.

Além de comparar à ontologia de turismo, também comparamos com uma ontologia sobre currículos. Neste estudo, o grau de similaridade foi praticamente zero para todos os documentos. Este resultado deve-se principalmente à diferença de linguagem adotada entre a plataforma Lattes (para currículos acadêmicos) e as taxonomias de

[4] A instrumentação das ontologias foi obtida a partir de HTTP://protegewiki.stanford.edu/index.php/Protege_Ontology_Library, maio 2008

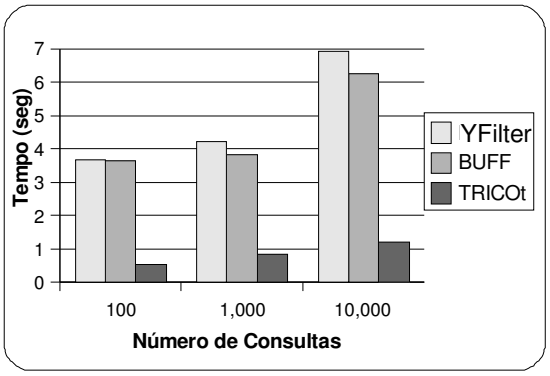
currículos empresariais. Como esperado, os documentos que seguem as taxonomias de vinhos foram detectados como não similares às ontologias de viagem e de currículos.

Para o sistema de disseminação de conteúdo, este resultado de identificar 96,14% dos documentos corretamente significa que esses documentos serão encaminhados para os *peer*s que contêm consultas sobre o domínio turismo. Os 4% restantes apresentaram limiar inferior a 70%, o que indica que novas ontologias precisam ser definidas. Os resultados do casamento entre consultas e a ontologia são similares. A diferença é que o limiar teve de ser reduzido para considerar a estrutura pequena das consultas em relação às ontologias. Um estudo mais aprofundado de como chegar ao limiar ideal para o casamento de consultas está sendo realizado e os resultados serão discutidos em trabalhos futuros.

### 5.3. Disseminação de Conteúdo com Tricot

Além dessa avaliação da qualidade do casamento, também realizamos uma avaliação da viabilidade de utilizar a Tricot num sistema de disseminação de conteúdo. O filtro de mensagens é a função principal do sistema e a sua avaliação experimental indica o potencial para a utilização das técnicas apresentadas em sistemas reais. Com o casamento entre documentos, consultas e ontologias, espera-se que cada *peer* da rede contenha apenas documentos e consultas do mesmo conjunto de domínios. Neste caso, em vez de avaliar todas as consultas sobre cada documento que é recebido no *peer*, apenas as consultas de mesmo domínio do documento serão avaliadas.

É importante notar que o casamento entre documentos e consultas acontece em paralelo no super *peer*. Existe um *pipeline* formado entre este casamento e o processamento das consultas dentro dos *peers*. Especificamente, primeiro o super *peer* define o domínio do documento e das consultas para depois encaminhá-los aos *peers* que efetivamente irão processar as consultas sobre os documentos. Então, o tempo de casamento não precisa ser considerado como parte do tempo de processamento de consulta (pois acontece em paralelo, numa primeira fase do *pipeline* do sistema distribuído).




*A Tricot afeta o número de documentos e o número de consultas que são avaliadas sobre tais documentos. Neste caso, os algoritmos de avaliação de consultas conseguem operar mais rapidamente, devido ao menor número de consultas sendo processadas. Este resultado (menor tempo para a avaliação de menos consultas em menos documentos) é o esperado. Porém, este experimento serve para demonstrar a viabilidade da Tricot bem como o seu potencial para melhorar o desempenho do sistema como um todo.*

**Figura 5. Viabilidade da Tricot.**

Para avaliar a viabilidade da Tricot, além dos algoritmos de casamento, nós também implementamos alguns algoritmos para avaliar as consultas sobre os documentos de mesmo domínio, especificamente o algoritmo utilizado pelo YFilter [Diao et al. 2004] (avalia o documento em pré-ordem) e o BUFF [Moro et al. 2007] (avalia o documento em pos-ordem). A Figura 5 apresenta e discute os resultados

variando o número de consultas processadas sobre os 10000 documentos da avaliação anterior. Apenas 50% das consultas são relacionadas aos domínios utilizados. Atualmente, estamos trabalhando para estender essa avaliação experimental considerando um sistema real, em vez do simulador.

## 6. Trabalhos Relacionados

Esta seção descreve trabalhos relacionados aos diferentes aspectos considerados: disseminação de conteúdo XML, redes *peer-to-peer* e ontologias. É apresentada também uma discussão mais detalhada de como este trabalho contribui para avançar o estado-da-arte, ressaltando as contribuições científicas e as limitações de nossa solução.

**Disseminação de Conteúdo XML.** Pesquisas recentes sobre disseminação de conteúdo XML têm investigado problemas relacionados a diferentes partes da arquitetura do sistema. Os aspectos mais relevantes incluem: a construção da rede sobreposta [Diao et al. 2004], a indexação e a agregação de perfis (consultas) dentro de um roteador [Li et al. 2007, MBT07], a distribuição de consultas [Diao et al. 2004, LHJ07], a codificação das mensagens na rede [Vagena et al. 2007] e a tarefa de filtrar mensagens [Li et al. 2007, Moro et al. 2007, Vagena et al. 2007]. Entre todos esses aspectos, a tarefa de filtrar mensagens tem recebido maior parte da atenção por dois motivos principais. Primeiro, é a tarefa mais crítica para o desempenho do sistema como um todo. Ao mesmo tempo, é a tarefa mais complexa devido à estrutura em árvore dos dados XML. Para essa função, algoritmos baseados em autômatos têm estado entre as soluções mais populares [Diao et al. 2004, Vagena et al. 2007]. Soluções alternativas têm também sido pesquisadas, como por exemplo, usando *subsequence matching* [Moro et al. 2007].

**Redes *Peer-to-Peer* e Ontologias.** A arquitetura baseada em *super-peers*, no qual o gerenciamento é realizado somente nos *super-peers* (através da monitoração dos *peers* agregados) mostra-se muito vantajosa, uma vez que a comunicação é estabelecida somente entre os *super-peers* e não entre todos os *peers* da rede. A pesquisa é geralmente mais rápida, porém o problema da agregação de *peers* em *super-peers* é uma questão importante. Esta tarefa pode ser realizada de acordo com algumas características, como, por exemplo, assunto e localidade. Nossa proposta utiliza uma ontologia como critério de agrupamento de *peers*. Alguns trabalhos apresentam técnicas para realização de casamento entre esquemas e ontologias, baseado em técnicas de similaridade. O sistema COMA++[6], por exemplo, é uma ferramenta de casamento entre esquemas e ontologias. O sistema suporta os modelos XSD, OWL, XDR (XML *Data Reduced*) e esquemas relacionais [Aumueller et al. 2005]. A medida de similaridade entre dois modelos é determinada por um função de similaridade entre dois elementos pertencentes a taxonomia (descritos pela ontologia). No entanto, não é detalhado como é feito o cálculo da similaridade entre os modelos e a taxonomia. Em [Broekstra et al. 2003], um modelo de dados é proposto para representar informação semântica que combina características de ontologias (hierarquia de conceitos) com uma descrição e um modelo que possibilita manipular visões heterogêneas sobre um determinado domínio. [Boyd et al. 2004] descreve o repositório AutoMed, que fornece a implementação para a abordagem *both-asview* de integração de dados. [Xu and Embley 2006] propõem TIQS,

[6] HTTP://dbs.uni-leipzig.de/Research/coma.html

uma abordagem para integração de dados que usa casamento semi-automático de esquemas para mapeamentos, com base no esquema conceitual global pré-definido.

**Discussão.** Todos esses trabalhos relacionados focam em diferentes aspectos do funcionamento de sistemas de disseminação. As pesquisas em sistemas de disseminação para dados XML estão apenas começando. O conceito de roteadores XML foi definido em 2001 [Snoeren et al. 2001], porém, trabalhos de pesquisa relevantes sobre sistemas de *publish/subscribe* com dados XML começaram a ser publicados somente em 2004 (na comunidade de banco de dados). Este artigo procurou focar nos aspectos de distribuição de documentos, distribuição de consultas e processamento dessas consultas. Essas três tarefas utilizam um sistema homogêneo baseado em ontologias. É importante notar que é a primeira vez que tal homogeneidade é definida para um sistema de tamanha complexidade. Enquanto os trabalhos de estado-da-arte visam aprimorar um ou outro ponto, este trabalho considera a melhoria global do sistema. Outra questão fundamental é que, apesar da ontologia ser utilizada no particionamento do conjunto de consultas, o algoritmo que realiza o processamento das consultas nos documentos não utiliza as informações semânticas da ontologia. Esta parte do processamento está sendo atualmente investigado e deverá ser apresentado em trabalhos futuros.

## 7. Considerações Finais

Sistemas de disseminação de conteúdo estão em constante atualização. Novas consultas são requisitadas e novos dados são adicionados. Desse modo, a principal preocupação deste artigo (do ponto de vista de banco de dados) é a escalabilidade em termos da quantidade de consultas e do volume de dados que podem ser processados ao mesmo tempo. Este artigo descreve Tricot, uma proposta baseada em ontologias para o processamento de consultas em sistemas de disseminação de conteúdo sobre redes P2P. A grande contribuição é fornecer uma estrutura homogênea para a distribuição de documentos, a disseminação de consultas e o processamento dessas consultas em sistemas baseados em documentos XML. A Tricot pode ser empregada em uma variedade de aplicações, como, por exemplo, bibliotecas digitais, RSS *feeds*, mercado de ações, governo eletrônico e disseminação de notícias. Finalmente, as questões de pesquisa relacionadas aos sistemas de disseminação de conteúdo estão relacionadas à gestão da informação em grandes volumes de dados distribuídos, o qual constitui o primeiro dos *Grandes Desafios da Pesquisa em Computação no Brasil da SBC*.

Como trabalhos futuros, cita-se o estudo detalhado dos casos de uso mencionados para a adequação da Tricot às peculiaridades e aos requisitos de cada um deles. Nossa pesquisa em relação a esses cenários está apenas começando. É importante notar que os resultados iniciais são suficientes para definir o potencial de aplicação da Tricot nestes e em outros cenários que venham a ser identificados. Estudos mais aprofundados de como chegar ao limiar ideal para o casamento de consultas e como considerar a informação da ontologia no processamento efetivo da consulta dentro de um *peer* estão em andamento e seus resultados serão discutidos em trabalhos futuros.

## Bibliografia

Aumueller, D., Do, H. H., Massmann, S., and Rahm, E. (2005). Schema and ontology matching with coma++. In Proc. of SIGMOD Conference, pages 906–908.

Berglund, A., et. al (2007). XML Path Language (XPath) 2.0. In W3C Recommendation, http://www.w3.org/TR/xpath20.

Boyd, M., et. al (2004). Automed: A bav data integration system for heterogeneous data sources. In Proc. of CAiSE, pages 82–97.

Broekstra, J., Ehrig, M., and Haase, P. (2003). A metadata model for semantics-based peer-to-peer systems. In Work. on Semantics in Peer-to-Peer and Grid Computing.

Costa, M., et. al. (2005). Vigilante: End-to-End Containment of Internet Worms. In Proc. of SOSP.

Diao, Y., Rizvi S. and Franklin, M. J. (2004). "Towards an Internet-Scale XML Dissemination Service", In: Proc. of VLDB, p. 612-623.

Levenshtein, V. (1996). Binary codes capable of correcting deletions, insertions, and reversals. Cybernetics and Control Theory, 10(8):707 – 710.

Li, G., Hou, S., and Jacobsen, H.-A. (2007). Routing of XML and XPath Queries in Data Dissemination Networks. In Proc. of ICDE, pages 1400–1404.

Madhavan, J., Bernstein, P. A., and Rahm, E. (2001). Generic schema matching with cupid. In Proc. of VLDB, pages 49–58.

Maedche, A. and Staab, S. (2002) Measuring similarity between ontologies. In Proc. of European Conference on Knowledge Acquisition and Management, 2002. p.251-263.

Maedche, A., et. al (2002). Mafra - a mapping framework for distributed ontologies. In Proc. of EKAW, pages 235–250.

C. D. Manning and Schutze, H. (1999). Foundations of Statistical Natural Language Processing. Cambridge, MA: MIT Press.

Meghini, C. and Spyratos, N. (2007). Computing Intensions of Digital Library Collections. In Proc. of ICFCA, pages 66–81.

Moro, M.M., Bakalov, P., and Tsotras, V. J. (2007). Early Profile Pruning on XML-aware Publish/Subscribe Systems. In Proc. of VLDB, pages 866–877.

Saccol, D. B., Edelweiss, N., Galante, R. M., Mello, M. R. (2008a). Managing Application Domains in P2P Systems. In: IRI 2008 - IEEE International Conference on Information Reuse and Integration 2008, July 13-15, 2008, Las Vegas, USA.

Saccol, D. B., Noll, R. P., Edelweiss, N., and Galante, R.M. (2008b). An Ontology-based Approach for Semantic Interoperability in P2P Systems. In Proc. of ICEIS.

Silva, R. da, Stasiu, R., Orengo, V. M., Heuser, C. A. (2007). Measuring quality of similarity functions in approximate data matching. Journal of Informetrics, 1(1): 35–46.

Snoeren, A. C., Conley, K., and Gifford, D. K. (2001). Mesh-Based Content Routing using XML. In Proc. of SOSP, pages 160–173.

Vagena, Z., Moro, M. M., and Tsotras, V. J. (2007). RoXSum: Leveraging Data Aggregation and Batch Processing for XML Routing. In Proc. of ICDE, pages 1466–1470.

Xu, L. and Embley, D.W. (2006). A composite approach to automating direct and indirect schema mappings. Inf. Syst., 31(8):697–732.

Zhu, Y. and Hu, Y. (2007). Ferry: A P2P-Based Architecture for Content-Based Publish/Subscribe Services. IEEE Trans. Parallel Distrib. Syst., 18(5):672–685.