III Simpósio Doutoral do DI (SDDI'06)

Identificação

Nome: **Nuno Alberto Ferreira Lopes**
Email: nuno.lopes@di.uminho.pt

Título: *Partilha de Dados em Ambientes Par-a-Par*

Orientador: Carlos Baquero-Moreno, Departamento de Informática – Universidade do Minho

Data Início: 27 de Fevereiro de 2002
Data Término: 27 de Fevereiro de 2006 (prolongada por mais um ano)

Resumo

Área de Investigação e Desenvolvimento: Sistemas Distribuídos

Resumo: Esta tese estuda o problema da partilha de informação em sistemas "Peer-to-Peer" de larga escala completamente descentralizados e administrativamente autónomos. Mais concretamente foi focado o problema da escalabilidade num sistema com uma quantidade elevada de informação partilhada e consequentemente com um elevado número de pesquisas para aceder à mesma.

Objectivos:

- Construir sistemas escaláveis capazes de armazenar grandes quantidades de informação e de suportar pesquisas com uma utilização uniforme dos recursos de armazenamento e rede por entre os participantes do sistema.
- Dotar estes sistemas com tolerância a faltas de modo a manter a informação disponível apesar da elevada volatilidade com que participantes entram e saem do sistema. Não só a informação deve estar disponível como também estar actualizada, o que levanta problemas de coerência de dados em redes de larga-escala.

Contribuições

- Definição de um novo algoritmo distribuído que pretende dotar sistemas P2P estruturados (DHTs) com uma implementação básica de estrutura básica de dados (conjunto) ou índice (B-tree). Este algoritmo irá permitir o uso de DHTs como plataforma base para novas aplicações que se baseiam nestas estruturas básicas e que até agora se encontravam limitadas a ambientes centralizados (incluindo clusters), expandindo-as para larga-escala. Após a estabilização da versão inicial do algoritmo foram testadas várias heurísticas para melhorar a performance do mesmo em operações básicas de pesquisa tais como calcular a intersecção de dois conjuntos por exemplo. Um caso de teste é a criação de um índice invertido sobre uma colecção de documentos.
- Para validar o algoritmo foi construído um simulador de eventos discretos baseado no modelo definido pela Scalable Simulation Framework, mas com uma interface mais simples permitindo a implementação de protótipos mais rapidamente.

Publicações:

- *Evaluating distributed Balanced Trees over DHTs for building large-scale index*, relatório técnico em preparação.

  > DHT systems are structured overlay networks capable of using P2P resources in a scalable way for object location by implementing distributed hash-table functionality. However, their efficiency expects a level of uniformity in stored data that is not often present in inverted indexes. Popular contents yield a large set of locations, that would be assigned to a single host and create a local hotspot. We propose a new distributed tree algorithm that will enable full-text indexing over a DHT layer. The approach ensures that the underlying DHT receives a uniform load by assigning fixed sized, or low variance, blocks and applying data caching techniques that avoid communication and content hotspots.

- *Análise da Disponibilização de um Indice Invertido em P2P*, poster em Conferência de Redes e Computadores (CRC2003), Bragança, Setembro 2003.

  > A procura de informação com base em conteúdos é uma funcionalidade fundamental na construção de sistemas de partilha de ficheiros e recursos. Contudo, as soluções eficientes para a partilha peer-to-peer que tem por base DHTs - distributed hash tables - perdem a capacidade de efectuar buscas ao requerer a existência de identificadores unívocos. Tendo em conta que a manutenção de um índice invertido de conteúdos é essencial para

a concretização eficiente de buscas, apresenta-se neste trabalho um modelo de construção deste índice, com base em árvores-B+, que permite continuar a tirar partido das propriedades de escala dos actuais mecanismos de DHT.

- *Towards Peer-to-Peer Content Indexing*, em ACM Operating Systems Review, Vol. 37 No. 4, Outubro 2003.

  Distributed Hash Tables are the core technology on a significant share of system designs for Peer-to-Peer information sharing. Typically, a location mechanism is provided and object identifiers act as keys in the index of object locations. When introducing a search mechanism, where single words are used as keys, the key image cardinality will be driven by the word popularity and most of the present designs will be unable to load balance the index among the nodes. We present two contributions: A design that allows participating nodes to load balance the indexing of popular keys and avoid content hot-spots on single nodes; A distributed mechanism for probabilistic filtering of popular keys (with low search relevance) that paves the way for scalable full content indexing.

Plano de Trabalhos Resumido

- Estudo do estado da arte nos algoritmos de comunicação para sistemas P2P que incluem não-estruturados (ex: Gnutella) e estruturados (ex: Chord, Pastry, Can).
- Identificação da limitação dos sistemas actuais em fazer um balanceamento uniforme da utilização de recursos por todos os participantes. Nomeadamente em termos de armazenamento de dados e de carga de comunicação.
- Adaptação de um algoritmo baseado em B^+-Trees aos sistemas DHT para uniformizar carga pelos participantes em situações cujos dados apresentam uma assimetria elevada. Aplicação de heurísticas para tornar o algoritmo mais eficiente na utilização do recurso de comunicação.
- Implementação de uma infra-estrutura de simulação para validar o algoritmo e avaliar a sua performance.
- Análise das assumpções e requisitos necessários para replicação de dados no ambiente característico de sistemas P2P: completamente distribuído constituído por um elevado número de máquinas com um comportamento extremamente dinâmico. Estudo de um algoritmo de reconciliação de dados para manter a coerência dos dados neste tipo de ambiente.

Plano de Trabalhos Detalhado

O projecto de doutoramento teve início em Abril de 2002, sendo financiado por uma bolsa de doutoramento do MCT com efeitos a partir de Novembro de 2002. Estando previamente identificada a área de inserção da tese, Sistemas Distribuídos Peer-to-Peer, o período inicial foi dedicado à consolidação da recolha bibliográfica e ao aprofundamento de duas linhas de contribuição: Modelos de *distributed hash tables* com capacidade de adaptação de carga; Mecanismos P2P de catalogação de conteúdos em larga escala.

Estas contribuições levaram à redacção de um breve artigo, com o título *Towards Peer-to-Peer Content Indexing*, que foi publicado na revista *ACM Operating Systems Review, Vol. 37 No. 4, October 2003*. O período final de 2002 foi dedicado à especificação deste novo modelo e à implementação de um simulador para estudar as suas propriedades operacionais.

A partir de Janeiro de 2003 a investigação debruçou-se sobre a possibilidade de reutilizar os sistemas DHT já existentes como uma plataforma base para um sistema P2P capaz de armazenar um índice invertido. O índice invertido é uma ferramenta que pode ser utilizada para a catalogação de conteúdos. Esta funcionalidade é muito importante para a usabilidade dos sistemas DHT, no contexto de partilha de documentos, e ainda não está presente nos sistemas actuais. Nesta fase procurou-se identificar uma estrutura de dados capaz de ser implementada sobre um sistema base DHT, sem afectar no entanto a escalabilidade do mesmo. Foi redigido um artigo descrevendo este novo sistema intitulado *Análise da Disponibilização de um Índice Invertido em P2P*, que foi publicado como poster na conferência de Redes de Computadores (CRC'03) organizada pela FCCN. Simultaneamente com a escrita do artigo, foi implementado um simulador simplificado para validar a eficiência deste sistema.

Entre Agosto e Outubro de 2003 estudou-se a possível extensão do sistema de modo a oferecer a mesma funcionalidade presente nos sistemas de catalogação centralizados. A especificação do simulador foi também redesenhada de modo a permitir analisar com maior pormenor certas propriedades do sistema. Durante o mês de Outubro escreveu-se um pequeno artigo (em inglês) que apresenta sucintamente o sistema atrás referido e que foi apresentado no 1º Simpósio Doutoral do Departamento de Informática da Universidade do Minho.

No final de 2003 foram definidos como objectivos futuros utilizar os resultados do novo simulador para redigir um artigo longo descrevendo em pormenor o novo sistema e estudar uma potencial nova linha de investigação com o intuito de implementar classes genéricas de dados eficientemente sobre sistemas base DHT.

O ano de 2004 focou-se no estudo mais aprofundado do sistema proposto nos artigos anteriores, nomeadamente na implementação de um simulador genérico para validação de sistemas Peer–to–Peer e na definição exaustiva e completa das operações e protocolo de mensagens para o sistema atrás referido. Foi implementado de raiz um simulador genérico para validar o funcionamento de sistemas Peer–to–Peer, mais concretamente para observar as propriedades escaláveis do sistema, nomeadamente o número de mensagens trocadas entre os nodos do

sistema e a correcção das operações.

Paralelamente ao desenvolvimento do simulador, foi definido um protocolo de troca de mensagens para a execução das operações envolvidas neste sistema de procura. No decurso desta especificação foram detectados problemas clássicos da área de Sistemas Distribuídos, que precisam de ser resolvidos para o correcto funcionamento do sistema. Os problemas incluem entre outros a coordenação entre várias réplicas com a mesma informação e a coordenação entre vários nodos para a tomada de uma mesma decisão entre todos, no contexto de sistemas P2P em que cada réplica tem um tempo de vida limitado e a troca de mensagens tem de ser escalável para o número total de nodos. Estes problemas são ainda alvo de estudo presentemente.

O ano de 2005 foi dedicado à especificação mais formal do algoritmo proposto de modo a ser utilizado como uma extensão aos sistemas de DHTs já existentes, juntamente com o desenvolvimento de um novo simulador de eventos discretos agora escalável a um maior número de máquinas virtuais de modo a poder simular com maior fiabilidade a execução do algoritmo a uma escala maior.

O algoritmo inicialmente descrito no ano anterior foi modificado de modo a corrigir erros detectados através da sua simulação. Para no futuro poder apresentar provas teóricas da correcção do algoritmo em conjunção com a sua simulação, foi utilizada uma linguagem formal para a descrição/especificação do mesmo. Esta especificação utilizou o modelo de Autómatos de Entrada/Saída descrito por Lynch et al. A utilização deste modelo deveu-se à sua fácil integração com um simulador de eventos discretos.

Após o algoritmo mostrar estabilidade no seu funcionamento, deu-se início à escrita de um relatório técnico, *Evaluating distributed Balanced Trees over DHTs for building large-scale index*, descrevendo o algoritmo (ainda de uma maneira informal) e à sua aplicação prática a sistemas reais, nomeadamente à construção de índices invertidos em larga-escala. Este relatório técnico não contém ainda resultados de simulação devido à limitação de escalabilidade detectada no simulador, tendo levado a uma nova re-implementação agora seguindo o modelo de *Scalable Simulation Framework* mas com um interface mais simples de modo a acelerar o desenvolvimento de protótipos.

Apesar da existência de vários simuladores disponíveis para a comunidade científica em geral, nenhum deles é especialmente adequado à simulação de sistemas Peer-to-Peer. Daqui resultou cada grupo de investigação na área de Peer-to-Peer desenvolver um simulador específico para cada caso. O simulador implementado no âmbito desta tese segue a mesma linha de raciocínio e destina-se a colmatar a ausência de uma solução especialmente adaptada à simulação de algoritmos distribuídos em larga-escala, que é o objectivo desta tese. Os resultados observados na simulação do algoritmo atrás referido irão demonstrar a escalabilidade e eficiência deste quando aplicado a um sistema de larga-escala (simulado).

No final de 2005 foram obtidos resultados de simulação que demonstraram a adequação do algoritmo para fazer um melhor balanceamento de carga de armazenamento pelos participantes de um sistema de larga-escala. No entanto, os primeiros resultados também mostraram que a versão básica do algoritmo

não reduz o custo de comunicação entre os participantes. Neste momento estão a ser desenvolvidas e testadas heurísticas para melhorar o custo de comunicação no algoritmo de modo a torná-lo ainda mais eficiente.

A investigação actual nos sistemas de DHTs está a propor a extensão a novas estruturas de dados de modo a possibilitar o seu uso por parte de aplicações distribuídas com maiores requisitos de funcionalidade e que não eram suportados inicialmente. Esta tendência vai de acordo com a linha de investigação desta tese de construir novos algoritmos para a implementação de classes de dados genéricas, de modo a permitir o uso de DHTs como uma plataforma distribuída para desenvolver novas aplicações de larga-escala. Espera-se conseguir mostrar através de simulação que o algoritmo desenvolvido nesta tese consegue obter melhores valores do que os propostos no estado da arte em sistemas DHTs funcionalmente equivalentes. Os resultados obtidos até agora são promissores.

Ainda resta uma linha de investigação por concluir que está relacionada com o tema da coerência de informação, no contexto de sistemas P2P cujas réplicas têm tempos de vida muito reduzidos e cujo conjunto de réplicas é extremamente dinâmico ao ponto de não ser possível conhecer num determinado instante todos os participantes desse conjunto. Este problema foi apresentado no Simpósio Doutoral do Departamento de Informática em Janeiro de 2005, no qual se focou a limitação (ou mesmo impossibilidade, quando considerado o caso real de um sistema P2P) das soluções baseadas em acordo definidas nos sistemas distribuídos clássicos quando aplicadas a réplicas em ambientes P2P de modo a evitar a divergência de informação. Uma possível solução, ainda em estudo, é a definição de algoritmos reconciliadores de réplicas baseados na semântica dos dados, dada a impossibilidade de garantir a não-divergência das réplicas. A implementação de tais algoritmos vai também implicar a escolha de algoritmos de multicast adequados aos requisitos de escalabilidade e performance presentes nos sistemas P2P.

Finalmente, quando estes últimos pontos estiverem concluídos dar-se-á ao início da escrita da tese, cujo propósito será a compilação de todo o material criado e recolhido descrito anteriormente.

Recursos Necessários

A simulação deste tipo de sistemas, em que se pretende abstrair detalhes específicos da rede mas validar a execução concorrente de um algoritmo e analisar a sua performance, necessita de máquinas com uma elevada capacidade processamento que não se encontra disponível nos comuns "desktops". Agradece-se os recursos tornados disponíveis pelo grupo de Sistemas Distribuídos do Departamento de Informática para realizar estas simulações.