Avaliação de redes P2P baseadas no fenômeno *Small World* para o compartilhamento de conteúdos similares gerados por funções LSH

Rodolfo S. Villaça[1], Luciano B. de Paula[2], Maurício F. Magalhães[2], Fábio Luciano Verdi[3]

[1] Departamento de Engenharias e Computação (DECOM/CEUNES)
Universidade Federal do Espírito Santo (UFES) – São Mateus, ES – Brasil

[2]Departamento de Engenharia de Computação e Automação Industrial (DCA/FEEC)
Universidade Estadual de Campinas (UNICAMP) – Campinas, SP – Brasil

[3]Universidade Federal de São Carlos (UFSCAR) – Sorocaba, SP – Brasil

rodolfovillaca@ceunes.ufes.br, {luciano,mauricio}@dca.fee.unicamp.br
verdi@ufscar.br

Abstract. *The organization of content sharing P2P networks greatly influences how to search and discover contents. Among several works in this area, we highlight the use of LSH (Locality Sensitive Hash) functions in the creation of keys used to index semantically similar contents. When it comes to structured P2P networks, such as Chord, the use of these functions creates an undesirable imbalance among peers participating in the network. This paper proposes an Overlay network exploiting the Small World phenomenon to share semantically similar contents. Using Oversim, semantically similar contents are hashed with LSH and MD5 functions and distributed in a Chord network and in a Small World based network, and the results compare the number of hops needed to recover similar contents, showing that the latter network is more appropriate in this scenario.*

Resumo. *A organização de uma rede P2P influencia na busca e recuperação de conteúdos. Dentre muitos trabalhos nesta área destacamos o uso de funções LSH para a criação de chaves utilizadas na indexação de conteúdos similares. Em redes P2P estruturadas, como Chord, o uso destas funções gera um desbalanceamento na distribuição das chaves entre os nós participantes da rede. Este trabalho propõe uma rede Overlay construída segundo o fenômeno Small World para o compartilhamento de conteúdos semanticamente semelhantes. Utilizando o Oversim, conteúdos semanticamente semelhantes são indexados através de funções LSH e MD5 e distribuídos em uma rede Chord e em uma rede Small World, e os resultados comparam o número de saltos (hops) necessários à sua recuperação, mostrando que essa última é mais adequada neste cenário.*

1. Introdução

Dentre as principais causas do declínio na utilização de redes P2P [Labovitz et al. 2009] podemos citar as dificuldades encontradas pelos usuários na busca e recuperação de conteúdo de seu interesse, frequentemente espalhados em grandes áreas e em muitos nós.

As redes P2P organizadas sob a forma de DHTs (*Distributed Hash Tables*) facilitam a recuperação de conteúdo, com a desvantagem de haver um custo associado à manutenção dessa organização. Além disso, geralmente os conteúdos indexados nestas DHTs através de funções de *hash* tradicionais, como o MD5, não mantêm relação de similaridade entre si, dificultando a busca por conteúdos semelhantes.

As funções de *hash* sensíveis à localidade, ou LSH (*Locality Sensitive Hash*), procuram manter a proximidade entre os identificadores gerados para conteúdos similares e, se aplicadas em DHTs, estas funções podem facilitar a busca inexata. Para exemplificar, podemos citar [Zhu 2005] que utiliza funções LSH para indexação de textos e [Haghani et al. 2009] para imagens, ambos aplicadas em DHTs. Nesses dois trabalhos, textos e imagens similares são indexados e mantidos próximos, possibilitando a busca por um grupo de elementos similares. Neste sentido utilizamos uma função LSH com similaridade baseada em ontologias para gerar identificadores de conteúdos semanticamente relacionados. Os resultados mostram que a função LSH utilizada concentra as chaves geradas em uma região do espaço de identificadores comparando os resultados com o espalhamento gerado por uma função de *hash* tradicional, nesse caso o MD5.

Entretanto, essa distribuição não-uniforme das chaves geradas por funções LSH leva a um desbalanceamento do espaço de identificadores. Com o objetivo de propor e avaliar uma forma adequada para a recuperação de um conjunto desbalanceado de identificadores gerados por conteúdos similares através da função LSH mencionada anteriormente, este trabalho propõe a construção de redes P2P baseadas no fenômeno *Small World*.

O fenômeno *Small World* [Kleinberg 2000] aplicado em redes P2P privilegia o estabelecimento de contatos (*fingers*) próximos em relação ao estabelecimento de contatos distantes, onde essa noção de proximidade está relacionada à distância lógica entre dois pontos no espaço de identificadores. Estas redes exibem ao mesmo tempo características de *clusterização* e baixo número de saltos no roteamento entre dois pontos quaisquer da rede. Tais características fazem com que estas redes sejam candidatas naturais para a distribuição de identificadores gerados por funções LSH. Em redes como o *Chord* o estabelecimento de contatos não possui relação com a similaridade entre os conteúdos.

Inspirado em [Girdzijauskas 2009], este trabalho propõe alterações no algoritmo de estabelecimento de contatos do *Chord* com o objetivo de construir uma DHT que possua as características do modelo *Small World*. Através do software de simulação de redes P2P *Oversim*[1] implementamos uma rede com estas características. O resultado mostra que o número de saltos necessários para a recuperação de conteúdos similares em uma rede *Small World* é menor quando comparado com uma DHT *Chord*. Esta redução acontece tanto para conteúdos indexados com uma função LSH quanto para conteúdos indexados com funções de *hash* tradicionais, como o MD5, devido à garantia de existência de poucos saltos entre quaisquer dois pontos de uma rede *Small World*. Entretanto, o melhor resultado e o menor número de saltos acontece com a utilização do conjunto LSH e *Small World*. Tais resultados mostram que é possível facilitar a busca inexata por conteúdos similares em uma rede P2P estruturada através do uso de redes *Small World* e funções LSH, sendo esta a principal contribuição deste trabalho.

[1]The Oversim P2P Simulator - http://www.oversim.org/

Este artigo está organizado da seguinte maneira: na Seção 2 há uma apresentação dos conceitos principais e trabalhos relacionados envolvendo redes P2P e do modelo *Small World*. A Seção 4 apresenta um exemplo de função LSH com similaridade baseada em ontologias, responsável pela geração dos identificadores de conteúdo utilizados neste trabalho. Em seguida, na Seção 3, estão detalhadas as modificações no algoritmo de estabelecimento de contatos do *Chord* de tal forma a montar uma rede que se assemelhe ao modelo *Small World*. Na seção seguinte (Seção 5), a metodologia de testes utilizada neste trabalho é descrita, junto com os resultados obtidos da comparação entre o *Chord* e a rede *Small World*. Por fim, na Seção 6, apresentamos as conclusões finais e perspectivas de trabalhos futuros.

2. Redes P2P e o Modelo *Small World*

Esta seção descreve um modelo de referência para redes P2P e o modelo *Small World* aplicado a essas redes. A proposta desse trabalho visa as redes P2P descentralizadas, porém estruturadas. Nesse tipo de rede P2P, o roteamento segue alguma regra baseada na topologia, que pode ser organizada em anel (*Chord*), árvore (*KAD*), *plaxon-style mesh* (*Pastry*) ou em um plano cartesiano de várias dimensões (*CAN*) [Lua et al. 2005].

2.1. Conceitos básicos

Podemos descrever uma rede P2P como sendo um conjunto de nós $\mathcal{P}$ que provê acesso a um conjunto de recursos (conteúdos) $\mathcal{R}$ através de uma função que mapeie $\mathcal{P}$ e $\mathcal{R}$ em um espaço virtual de identificadores $\mathcal{I}$ utilizando-se de duas funções $\mathcal{F}_{\mathcal{P}} : \mathcal{P} \rightarrow \mathcal{I}$ e $\mathcal{F}_{\mathcal{R}} : \mathcal{R} \rightarrow \mathcal{I}$. Estas funções permitem o estabelecimento de relações de proximidade entre nós e recursos através de métricas no espaço virtual de identificadores $\mathcal{I}$. Para permitir o acesso aos nós, e consequentemente aos recursos compartilhados em uma rede P2P, torna-se necessário embutir um grafo no espaço de identificadores proposto $\mathcal{I}$. Este grafo auxilia o roteamento, facilitando a recuperação de conteúdos associados às suas chaves (identificadores) na rede.

Um projeto de redes P2P deve considerar os seguintes aspectos:

- a escolha do espaço de identificadores $\mathcal{I}$;
- a escolha das funções de mapeamento $\mathcal{F}_{\mathcal{P}}$ e $\mathcal{F}_{\mathcal{R}}$;
- associação de pertinência entre os nós e os recursos;
- o grafo embutido no espaço de identificadores;
- a estratégia de roteamento;
- a manutenção da rede em função da entrada e saída de nós.

O atual trabalho não aborda as questões relacionadas à manutenção da rede e utiliza uma organização dos nós em anel, como no *Chord*.

2.1.1. Espaço de Identificadores e Roteamento

Uma característica fundamental na escolha do espaço de identificadores é a existência de alguma métrica de proximidade d: $\mathcal{I}\ x\ \mathcal{I} \rightarrow$ R, onde R denota o conjunto dos números reais. Essa métrica d deve satisfazer as propriedades 1, 2 e 3, apresentadas a seguir, e se possível também as propriedades 4 e 5.

$\forall x,y \in \mathcal{I} : d(x,y) \geq 0$ (Propriedade 1)

$\forall x \in \mathcal{I}$: *d(x,x)* = 0 (Propriedade 2)

$\forall x,y \in \mathcal{I}$: *d(x,y)* = 0 $\rightarrow$ *x=y* (Propriedade 3)

$\forall x,y \in \mathcal{I}$: *d(x,y)* = *d(y,x)* (Propriedade 4)

$\forall x,y,z \in \mathcal{I}$: *d(x,z)* $\leq$ *d(x,y)* + *d(y,z)* (Propriedade 5)

A escolha adequada do espaço de identificadores dos nós e recursos da rede é uma questão relevante pois, a partir desta escolha, dependerão fatores como escalabilidade, separação de identificadores e localizadores, roteamento, *clusterização* e manutenção da semântica dos conteúdos mapeados neste espaço. Estes três últimos são os fatores principais no desenvolvimento da proposta deste trabalho, e todos são diretamente dependentes desta métrica de distância *d*.

2.1.2. Funções de Mapeamento $\mathcal{F}_\mathcal{P}$ e $\mathcal{F}_\mathcal{R}$

A função de mapeamento $\mathcal{F}_\mathcal{P}$ associa um nó a um ponto no espaço $\mathcal{I}$. Na maioria dos casos este ponto é único, tal que $\forall p,q \in \mathcal{R} : p \neq q \rightarrow \mathcal{F}_\mathcal{P}(p) \neq \mathcal{F}_\mathcal{P}(q)$. Entretanto, em casos de replicação de conteúdo nos nós como uma solução de tolerância a falhas, admite-se o relaxamento desta condição. Essa mesma condição pode ser aplicada à função $\mathcal{F}_\mathcal{R}$, associando um recurso na rede a um ponto no espaço de identificadores. Neste caso o relaxamento desta condição representa a probabilidade de conteúdos similares ocuparem o mesmo ponto no espaço de identificadores.

A probabilidade de um nó ou recurso ser mapeado em um determinado ponto ou região do espaço virtual $\mathcal{I}$ também é uma característica importante na escolha das funções $\mathcal{F}_\mathcal{P}$ e $\mathcal{F}_\mathcal{R}$. Essa distribuição pode ser homogênea (probabilidades iguais) ou não-homogênea (probabilidades diferentes). Funções de *hash* como MD5 e SHA-1 possuem distribuição com probabilidades iguais, enquanto que funções LSH, como a que será apresentada na seção 4, possui distribuição com probabilidades diferentes pois a mesma possui a característica de manter a similaridade dos recursos $\mathcal{R}$ mapeados em $\mathcal{I}$. A função proposta neste artigo deverá satisfazer às relações 6 e 7 apresentadas a seguir, onde *sim* é a similaridade:

$\forall\ s_1,s_2 \in \mathcal{R}$: *d*($\mathcal{F}_\mathcal{R}(s_1)$,$\mathcal{F}_\mathcal{R}(s_2)$) $\propto$ *sim*(s_1,s_2) (6)

$\forall\ s_1,s_2 \in \mathcal{R}$: *P*[$\mathcal{F}_\mathcal{R}(s_1)$ = $\mathcal{F}_\mathcal{R}(s_2)$] = *sim*($s_1$,$s_2$) (7)

Em (6) e (7), definimos uma função de similaridade *sim(a,b)* : $\mathcal{R}$ *x* $\mathcal{R}$ $\rightarrow$ [0..1], onde 0 significa nenhuma similaridade, e 1 significa similaridade total. A relação (6) implica que a distância entre dois conteúdos quaisquer (s_1,s_2) é proporcional à similaridade entre eles, ou seja, quanto mais similares, mais próximos no espaço de endereçamento eles se encontram, e vice-versa. A relação (7) indica que a probabilidade de que ambos tenham um mesmo identificador é igual à similaridade entre eles, ou seja, quanto maior a similaridade, maior a probabilidade de possuírem o mesmo identificador, e vice-versa.

2.1.3. Roteamento em Redes P2P estruturadas

O serviço básico provido por uma rede P2P estruturada é encaminhar a requisição de um identificador *i* ($i \in \mathcal{I}$) a partir de um nó *p* ($p \in \mathcal{P}$). A estratégia de roteamento influencia no número de saltos necessários no encaminhamento da requisição entre *p* e *i* e pode ser definida como uma função não determinística $\mathcal{R} : \mathcal{P}\ x\ \mathcal{I} \rightarrow \mathcal{P}$ que seleciona, no conjunto

$\mathcal{N}$ de nós vizinhos a p ($\mathcal{N}_\mathcal{P}$), um outro nó r ($r \in \mathcal{N}_\mathcal{P}$) para encaminhar a requisição. O nó escolhido deverá satisfazer a seguinte relação: $d(i,\mathcal{F}_\mathcal{P}(r)) < d(i,\mathcal{F}_\mathcal{P}(p))$ (8)

2.2. O fenômeno *Small World*

As primeiras pesquisas envolvendo as características deste fenômeno, popularmente conhecido como "os seis graus de separação", pertencem à década de 60. Desde então, muitos outros trabalhos se dedicaram a explicar e modelar este fenômeno através de grafos e algoritmos. Dentre estes trabalhos destacam-se [Watts and Strogatz 1998] e [Kleinberg 2000].

Podemos afirmar que uma população possui as características do fenômeno *Small World* se, para quaisquer pares de indivíduos pertencentes a esta população, houver um caminho curto que interligue ambos através de uma sequência de outros indivíduos vizinhos que possuam algum conhecimento em comum. Em uma rede, esse "conhecimento em comum" pode ser traduzido em uma agregação, fazendo com que conteúdos que possuam características similares estejam armazenados próximos entre si, sendo esta proximidade medida no espaço virtual de identificadores.

[Watts and Strogatz 1998] propuseram um modelo para a construção de redes baseadas no fenômeno *Small World* onde o grafo de construção destas redes possui vizinhos divididos entre dois tipos: locais e de longa distância. Em seu trabalho, supondo um conjunto $\mathcal{I}$ de pontos pertencentes ao espaço virtual de identificadores, para cada nó $p \in \mathcal{P}$ mapeado no espaço virtual $\mathcal{I}$, interliga-se o mesmo com k vizinhos próximos, aleatoriamente escolhidos segundo alguma métrica de proximidade d. Estes são os vizinhos locais de p. Feito isso, escolhe-se aleatoriamente, um pequeno número de nós distantes. Estes são os vizinhos de longa distância de p.

[Kleinberg 2000] demonstrou que dentre uma grande variedade de redes *Small World* existe uma forma de construção onde o roteamento é mais eficiente e o número de nós intermediários que interligam dois pontos quaisquer da rede é mínimo. A principal contribuição deste trabalho é mostrar que o estabelecimento dos vizinhos não deve ser realizado de forma puramente aleatória, mas sim de forma inversamente proporcional à distância, conforme veremos a seguir.

Na proposta de Watts e Strogatz existe um número fixo k de vizinhos locais, e um pequeno número de vizinhos de longa distância, enquanto Kleinberg relaxa a necessidade desse valor fixo. Kleinberg estabelece que um nó p possui contatos curtos com seus vizinhos adjacentes e contatos de longa distância com nós pertencentes à rede segundo uma probabilidade inversamente proporcional a $d(p,q)^r$, onde r é um parametro estrutural e q é um nó qualquer pertencente à rede. Para pequenos valores de r o grafo formado é puramente aleatório (*Random Networks*), enquanto que a medida que r aumenta, o grafo torna-se cada vez mais clusterizado. Em nenhum destes casos extremos a rede apresentará as características de *Small World* desejadas.

Os resultados apresentados em nosso trabalho baseiam-se na construção de redes *Small World* proposta por [Girdzijauskas 2009], [Girdzijauskas et al. 2010]. Nela, o espaço virtual é organizado em um anel, e cada nó u possui contato direto com seus vizinhos adjacentes, à direita, e à esquerda. A principal diferença entre esta proposta e a proposta de Kleinberg é que cada nó deverá possuir $\log_2 N$ contatos, onde N representa o número de nós na rede. [Girdzijauskas 2009] mostra que esta construção possui um custo

máximo de $O(\log_2(N))$ saltos, além de mostrar um algoritmo de descoberta que permite que seja realizada uma estimativa do valor de *N*.

3. Construção e simulação de uma Rede P2P *Small World*

Esta seção descreve uma maneira de modificar o algoritmo de aquisição de contatos (*fingers*) do *Chord* de tal forma a construir um grafo que possua as características do modelo *Small World*. A rede P2P gerada a partir destas modificações possui as seguintes características:

- espaço virtual formado por identificadores binários de *k* bits;
- identificadores dispostos em um anel, com roteamento no sentido horário;
- distribuição balanceada dos nós, cujos identificadores são gerados através de funções *hash* comuns;

O algoritmo de roteamento guloso (*greedy routing*) baseado na distância entre identificadores utilizado no *Chord* é mantido, assim como a forma de associação entre um nó e os identificadores pelos quais ele é responsável no anel (em resumo, um nó é responsável por todos os identificadores localizados na faixa compreendida por ele mesmo e seu antecessor no anel). Os testes descritos na Seção 5 foram realizados somente em uma rede estática, após a conclusão do procedimento de entrada de todos os nós. Além disso, fixamos o número de nós em cada teste de tal forma que não houve necessidade de adotar nenhum procedimento para estimar este valor.

3.1. Espaço virtual de identificadores

Conforme dito anteriormente, o espaço de identificadores é binário e apresenta uma organização circular. Considerando-se um espaço de *k* bits, admite-se identificadores na faixa compreendida entre 0 e 2^k no domínio dos números naturais. Considere também que a rede seja composta por *N* nós, e estes sejam dispostos aleatoriamente neste espaço segundo uma função de distribuição de probabilidade uniforme.

Desta forma, dividindo o espaço $\mathcal{I}$ em $\log_2 N$ partições logarítmicas, conforme proposto em [Girdzijauskas 2009], define-se as partições A_1, que contêm aproximadamente a metade da população $\mathcal{P}$ de nós da rede, correspondendo aos nós cujos identificadores pertencem ao intervalo $[2^{k-1}..2^k)$; A_2, que contém aproximadamente $\frac{1}{4}$ da população $\mathcal{P}$ de nós da rede, correspondendo aos nós cujos identificadores pertencem ao intervalo $[2^{k-2}..2^{k-1})$; e assim, sucessivamente.

A Figura 1 mostra um exemplo contendo um espaço de endereçamento de k = 4 bits com 8 nós. É importante ressaltar, observando a mesma figura, que a partição A_3 contém somente um nó, correspondente ao sucessor e ao antecessor de *p* no espaço $\mathcal{I}$, considerando ambos os sentidos no anel circular onde os identificadores foram distribuídos.

Esta divisão é importante pois dela extrairemos a função de probabilidade usada para estabelecimento dos contatos de cada nó na rede proposta. Supondo um nó *u* com identificador $\mathcal{F}_{\mathcal{P}}(u)$, o objetivo é fazer com que contatos pertencentes às partições mais próximas estejam em maior número na tabela de roteamento deste nó. É importante esclarecer que não se trata de privilegiar uma região em detrimento de outra, trata-se apenas de privilegiar os nós que estão mais próximos em $\mathcal{I}$ em relação aos nós mais distantes.

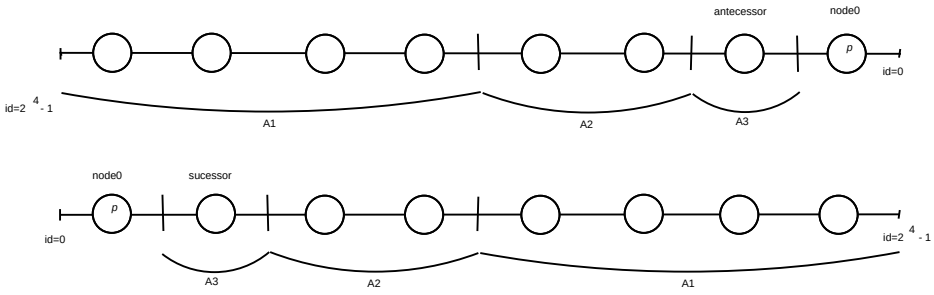

Figura 1. Espaço $\mathcal{I}$ com 8 nós e identificadores de *k*=8 bits

A maneira encontrada para realizar a implementação desse procedimento está diretamente relacionada com a divisão do espaço em partições. Com a mesma probabilidade (P=$1/\log_2 N$), sorteia-se uma dentre as $\log_2 N$ possíveis partições. O resultado deste sorteio determinará qual partição receberá uma solicitação de estabelecimento de contato para inserção na tabela de roteamento de u. Após a determinação de qual partição terá uma entrada na tabela de roteamento de u, sorteia-se novamente, com probabilidade igual (P=$1/\frac{N}{2^n}$), um identificador i pertencente à partição sorteada (relação (8)). Segue-se normalmente o procedimento adotado pelo *Chord* para o estabelecimento da ligação. Já foi dito na Seção 2.2 que este procedimento garante que o custo máximo do roteamento nesta rede, em termos do número de saltos, é $O(\log_2(N))$.

Todo o procedimento descrito no parágrafo anterior está perfeitamente adequado para um nó u com identificador 0, ou seja, localizado no início do espaço de endereçamento. No caso de um nó u qualquer, com identificador $\mathcal{F}_{\mathcal{P}}(u) \neq 0$ basta utilizar o valor sorteado como um deslocamento δ a ser adicionado ou subtraído ao identificador do nó (u), de tal forma que o nó v a ser inserido na tabela de roteamento de u possuirá identificador $\mathcal{F}_{\mathcal{P}}(v) = \mathcal{F}_{\mathcal{P}}(u) \pm \delta$. Isto levará em conta a proximidade à direita e à esquerda no espaço circular com probabilidade igual pois selecionaremos, aleatoriamente e com probabilidades iguais, o sinal do deslocamento δ durante a rotina de preenchimento da tabela de roteamento de cada nó.

3.2. Roteamento dos identificadores

O algoritmo de roteamento usado na rede *Small World* proposto neste trabalho baseia-se no roteamento guloso (*greedy routing*) usado no *Chord*, onde a cada encaminhamento aproxima-se cada vez mais do destino procurado. Essa aproximação é feita sempre levando-se em consideração o sentido horário e a distância em um espaço de identificadores circular. A Figura 2 mostra um exemplo de uma rede com k = 4 bits e 8 nós construída segundo a proposta deste trabalho. Na figura podemos observar os contatos locais de um nó u qualquer, representados pelo seu antecessor e sucessor no espaço $\mathcal{I}$ e três contatos de longa distância, representados por f_1, f_2 e f_3.

A partir da Figura 2 é possível explicar as diferenças conceituais existentes entre o roteamento no *Chord* e o roteamento na rede *Small World*. No *Chord*, embora exista a proximidade entre identificadores no sentido anti-horário do espaço virtual, esta proximidade não é importante para a escolha dos nós a serem adicionados na tabela de roteamento.

De acordo com a Figura 2, no *Chord* podemos interpretar os contatos f_1 e f_2 e o antecessor do nó u sendo contatos distantes, enquanto que o sucessor e o contato f_3

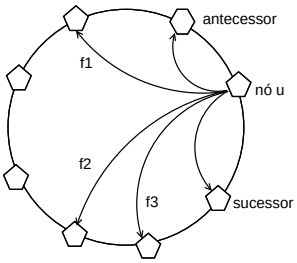

Figura 2. Contatos locais e de longa distância de um nó *u* em $\mathcal{I}$, *k*=4 e *N*=8

como sendo contatos próximos ao nó *u*. No *Chord*, o estabelecimento dos contatos é feito utilizando-se um algoritmo que considera somente o tamanho do anel e a distância no sentido horário, enquanto na rede baseada em *Small World* o estabelecimento de contatos é feito levando-se em consideração a proximidade dos nós no anel virtual em ambos os sentidos. Quanto mais próximos, maior a chance de dois nós estabelecerem contatos entre si.

A próxima seção discute um exemplo de construção de funções LSH cuja similaridade é baseada em ontologias para classificação de conteúdos.

4. Funções LSH com Similaridade Baseada em Ontologias

Neste trabalho estamos interessados não somente em avaliar o comportamento de uma rede P2P estruturada segundo o fenômeno *Small World*, comparando-a com o *Chord*, mas também estamos interessados em avaliar o comportamento destas redes em situações geradas pela agregação dos identificadores segundo a semântica dos conteúdos neles representados.

Neste contexto, uma das formas de se classificar dados, dando-lhes significado, é feita através da utilização de ontologias. Essa prática tem crescido recentemente devido à expansão da Web Semântica e do número de aplicações que consomem dados semanticamente classificados.

No escopo desse trabalho, uma ontologia é um modelo de dados que define um domínio e que pode ser usada para raciocinar sobre conceitos ou termos daquele domínio, estabelecendo relações entre eles [Uschold and Gruninger 2004]. Geralmente são utilizadas pequenas ontologias, com dúzias ou centenas de conceitos, para a definição de um domínio. Estas ontologias possuem menos expressividade mas são de fácil gerência. É possível utilizar essa classificação ontológica como base para uma função de similaridade e, a partir daí, criar uma família de funções LSH, utilizadas na criação de identificadores que mantenham essa relação entre si.

Em [Indyk and Motwani 1998] é definido que uma família de funções de *hash F* é classificada como sensível à localidade (LSH) correspondente a uma função de similaridade *sim(a, b)* se, para toda função $h \in F$, temos: $\mathbf{Pr}_{h \in F}[h(a) = h(b)] = sim(a, b)$, onde **Pr** é a probabilidade e *sim(a, b)* é uma função de similaridade que varia entre 0 e 1, sendo 1 para *a* e *b* completamente relacionados e 0 caso contrário.

Para a criação de uma função LSH baseada na similaridade extraída de uma ontologia, em primeiro lugar é utilizada a técnica *Enhanced Topic-based Vector Space Model* (eTVSM) [Polyvyanyy 2007] para obter a similaridade dos conceitos de uma ontologia

simples (também conhecida como ontologia *lightweight*). Para cada conceito da ontologia, o eTVSM gera um vetor que relaciona o mesmo com os demais termos.

Para todo conceito folha de uma ontologia, um vetor de conceito $\overrightarrow{\tau_i}$ = $\{\tau_{i,1}, \tau_{i,2}, ..., \tau_{i,k}\}$, onde *k* é o número total de conceitos da ontologia, é criado da seguinte maneira:

$$\tau_{i,k} = \begin{cases} 1 & \text{se } \tau_k \in \text{ ao caminho de } \tau_i \text{ até a raiz ou } i = k \\ 0 & \text{caso contrário} \end{cases}$$

Para todo conceito interno temos:

$$\overrightarrow{\tau_i} = |\sum \overrightarrow{\tau_s}|$$

onde $\overrightarrow{\tau_s}$ são todos os vetores de conceitos dos filhos diretos de τ_i. A similaridade entre cada par de conceitos pode ser medida pelo cálculo do cosseno entre os seus respectivos vetores, quanto maior o valor do cosseno, mais similares eles são.

Em seguida aplica-se a função *Random Hyperplane Hash* (RHH) [Charikar 2002] compondo um identificador de *n* bits no espaço $\mathcal{I}$. O RHH, uma família de funções LSH que correspondem à similaridade de cosseno, consiste em criar um vetor $\overrightarrow{r}$, sendo cada coordenada deste vetor gerada a partir de uma distribuição normal e, para cada vetor de conceito, contabiliza-se o produto escalar entre eles. Se o resultado do produto escalar for igual ou maior que 0, o resultado do *hash* é 1, caso contrário, 0. Para uma chave de *n* bits, são criados *n* vetores $\overrightarrow{r}$ e os resultados são concatenados.

A partir deste ponto, o identificador gerado pela aplicação deste método será utilizado como chave semântica daquele conceito da ontologia. Para uma maior ou menor agregação é possível criar *m* grupos de *n* vetores $\overrightarrow{r}$ e, pelo princípio da similaridade de Jaccard, escolhe-se o menor dentre os identificadores gerados como chave do conceito.

5. Testes e Resultados

Nos testes realizados com a função LSH com similaridade baseada em ontologias, comparamos a capacidade de agregação das chaves no espaço de identificadores com o espalhamento ocasionado pela geração de chaves através de uma função de hash tradicional, no caso o MD5. Esta função LSH foi implementada em linguagem Java.

As modificações propostas na Seção 3.2 foram implementadas no simulador *Oversim*, o qual usa a plataforma *Omnet++* e possibilita a simulação de grandes números de nós em uma única estação de trabalho PC usando o sistema operacional Linux. A implementação foi feita através de modificações realizadas na implementação do *Chord* disponível no Oversim, alterando-se as rotinas de aquisição de contatos e roteamento.

Também foram realizadas modificações na aplicação de busca e recuperação de chaves disponibilizadas no próprio *Oversim*, denominada *KBRTestApp*. As modificações feitas permitem realizar buscas por chaves específicas, determinadas pelo usuário.

Os testes feitos possuem dois objetivos principais:

1. Apresentar o comportamento da função LSH com similaridade baseada em ontologias, verificando a agregação dos conteúdos similares em uma determinada área do espaço de identificadores. A análise será feita em função do tamanho desta área, comparada à área utilizada por uma função de *hash* MD5, aplicada sobre os mesmos conteúdos;

2. Avaliar a implementação da rede *Small World* proposta neste artigo através da realização de buscas por chaves criadas por uma função LSH (agregadas no espaço dos identificadores) e por chaves criadas por um *hash* tradicional (MD5), medindo-se o número de saltos necessários para que estas buscas sejam concluídas. Essas buscas também são realizadas no *Chord* e os resultados são comparados.

O resultado esperado no primeiro teste é que, com o uso de uma função LSH, os identificadores sejam confinados em uma mesma região do espaço, enquanto que usando-se uma função *hash* MD5 estes mesmos identificadores encontrem-se distribuídos uniformemente no espaço.

Para o segundo teste espera-se que o custo, medido no número de saltos, para recuperação dos conteúdos gerados pela função LSH seja menor em uma DHT baseada no modelo *Small World* do que no *Chord*. Isto deve-se ao fato de que no *Small World* estes conteúdos estarão confinados em uma mesma região, e os nós pertencentes a esta região terão contatos diretos entre si com uma maior probabilidade. Para as buscas por conteúdos gerados pelo MD5, é desejado que os resultados obtidos com a rede *Small World* sejam, no mínimo, próximos àqueles obtidos utilizando o *Chord*. Isto deve-se à característica de baixo número de saltos presente no fenômeno *Small World*.

A seguir descrevemos os cenários de teste e as ontologias utilizadas para classificação do conteúdo. Na realização dos testes o seguinte cenário foi padronizado:

- Uma ontologia *lightweight* organizada sob a forma de árvore binária de 127 tópicos *t*, representando a classificação de algum conteúdo não especificado;
- Adoção de um espaço de identificadores de 128 bits e *N* variando em 1000, 2000, 5000 e 10000 nós, com uma distribuição aleatória e uniforme neste espaço;
- Comportamento estático das DHTs, de tal forma que durante as simulações não haveria entrada ou saída de nós.

A ontologia usada é representada por uma árvore binária de 7 níveis e 127 tópicos. A classificação segundo uma ontologia possui a granularidade de um conceito, ou seja, um conjunto de dados que possam ser classificados como similares e deveriam estar localizados próximos no espaço de identificadores. Assume-se, também, que todos os dados classificados segundo uma mesma ontologia, mas em conceitos distintos, possuem alguma relação entre si e, portanto, estarão próximos no espaço de identificadores.

5.1. Agregação de conteúdo usando LSH e ontologias para classificação de conteúdos

Utilizando a ontologia descrita na seção anterior são gerados 127 identificadores, cada um deles representando um tópico t_x através da função LSH descrita na Seção 4. Outros 127 identificadores foram gerados utilizando-se uma função *hash* MD5 no nome de cada tópico. De forma aleatória e uniforme distribuimos 1000, 2000, 5000 e 10000 nós no espaço de identificadores.

O resultado dos testes mostra o intervalo de armazenamento dos conceitos classificados segundo a ontologia para os dois conjuntos (LSH e MD5) de 127 identificadores. Esse intervalo é calculado pelo número de nós existentes entre o primeiro que tenha

chaves armazenadas e o último, seguindo o sentido horário no anel. Quanto menor o intervalo, mais próximos os conceitos estarão uns dos outros.

A Figura 3 mostra que o conteúdo indexado com a função LSH foi agregado no espaço de identificadores. Em uma rede com 1000 nós, a função de *hash* MD5 utilizou aproximadamente todos os nós na distribuição dos identificadores de conteúdo (989 nós), enquanto que a função LSH concentrou os identificadores em pouco mais de 25% dos nós (261). Resultados semelhantes podem ser observados na mesma figura em redes com 2000, 5000 e 10000 nós.

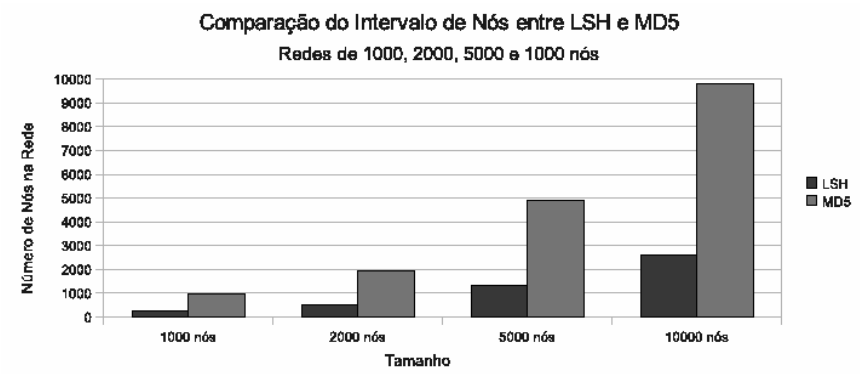

Figura 3. Intervalo de agregação para 1000, 2000, 5000 e 10000 nós

Como esperado, os resultados nos permitem afirmar que a função LSH utilizada mantêm a similaridade entre os identificadores gerados para conteúdos classificados segundo mesma ontologia, enquanto os identificadores criados pelo *hash* MD5 não possuem nenhuma agregação por similaridade, espalhando-se uniformemente no espaço.

Essa métrica de distância obedece a todas as propriedades de (1) a (5), pois mede o comprimento de uma região, o qual não pode ser negativo (1), pode ser nula se e somente se todo o conteúdo for agregado em um único identificador (2) e (3), é uma medida de distância obtida através da diferença entre o maior e o menor identificador (4) e, por ser unidimensional, necessariamente obedece à propriedade (5).

5.2. Comparações entre *Chord* e *Small World* quanto a busca de conteúdos similares

O segundo conjunto de testes realizados neste trabalho tem como objetivo verificar se houve redução no custo de recuperação de conteúdos similares na rede P2P *Small World*, comparando-se o custo de recuperação dos mesmos conteúdos em uma DHT *Chord*.

Nestes testes utilizamos o mesmo conjunto de identificadores gerados no teste anterior, criados pela função LSH e pelo *hash* MD5. O procedimento empregado na realização dos testes é descrito a seguir:

- Para cada tópico t_x ($x \in [1..127]$) da ontologia foram realizadas 127 buscas por cada um dos demais tópicos t_y ($y \in [1..127]$), incluindo o próprio t_x. Estas buscas eram originadas a partir do nó responsável pelo armazenamento de t_x;
- Ao alcançar o tópico de destino t_y, pertencente a algum nó da rede P2P, o número de saltos desde t_x até t_y foi armazenado em um arquivo de *log*;

Este procedimento foi repetido variando-se a função de *hash* (e consequentemente o conjunto de 127 identificadores), o modelo de rede utilizado (*Chord* e *Small World*) e o número de nós na rede P2P (1000, 2000, 5000 e 10000 nós). Em cada caso, calculamos a média e o seu respectivo intervalo de confiança, mediana, desvio padrão, máximo, mínimo. Os resultados obtidos para a média de saltos estão apresentados nas Figuras 4(a), 4(b), 4(c), 4(d).

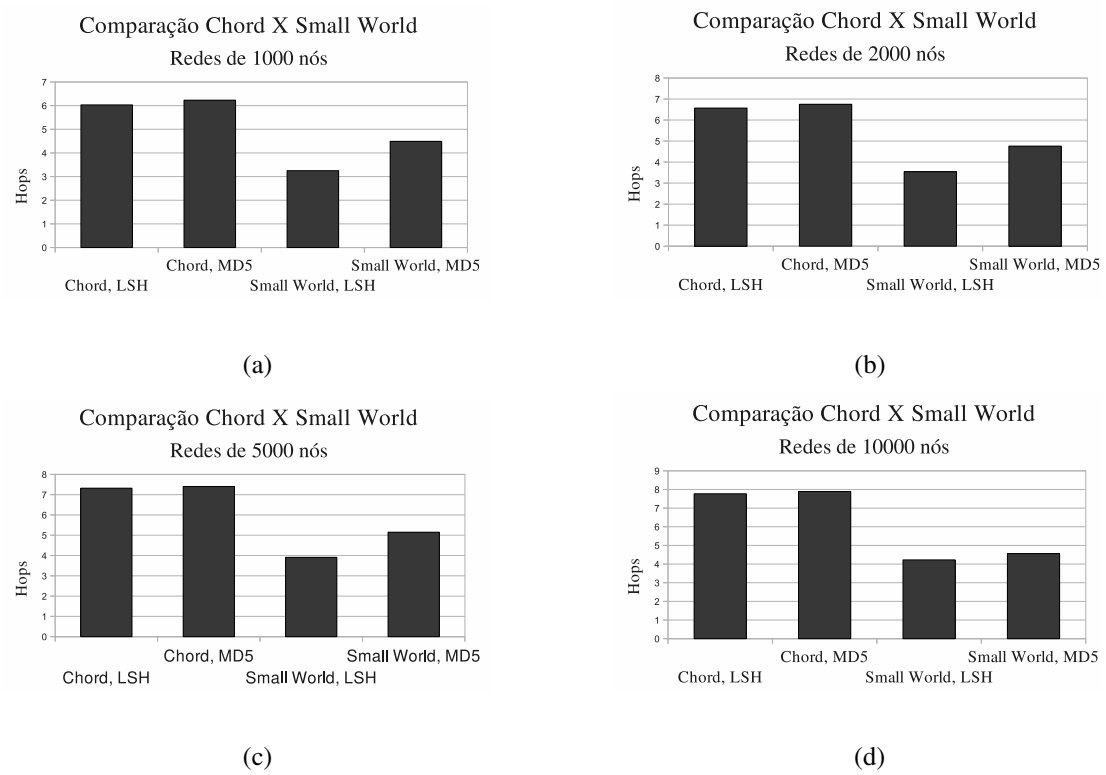

Figura 4. Comparação do número de saltos para redes de 1000(a), 2000(b), 5000(c) e 10000(d) nós

Em todos os testes, o número mínimo de saltos obtido foi 0, correspondendo às situações nas quais tópicos distintos estavam armazenados em um mesmo nó na rede. Em redes *Chord* de 1000 nós o número máximo obtido para o *hash* MD5 foi de 10 saltos, mesmo valor obtido com a função LSH. Usando *Small World* estes números caem para 8 e 6, respectivamente. Em redes *Chord* de 2000 nós o número máximo obtido para o *hash* MD5 foi de 11 saltos, mesmo valor obtido com a função LSH. Usando *Small World* estes números caem para 7 e 6, respectivamente. Em redes *Chord* de 5000 nós o número máximo obtido para o *hash* MD5 foi de 12 saltos, mesmo valor obtido com a função LSH. Usando *Small World* estes números caem para 8 e 7, respectivamente. Em redes *Chord* de 10000 nós o número máximo obtido para o *hash* MD5 foi de 13 saltos, mesmo valor obtido com a função LSH. Usando *Small World* estes números caem para 9 e 8, respectivamente.

Estes resultados mostram que o *Chord*, pela característica de estabelecimento de contatos somente no sentido horário, não aproveita adequadamente, na média, o ganho oferecido pela agregação dos identificadores gerados por uma função LSH. Isso acontece pois a agregação permite que dois identificadores estejam próximos segundo a métrica de distância utilizada porém, no sentido anti-horário do anel. Neste caso a agregação não terá

efeito na redução do número de saltos entre estes dois identificadores e a média tende a se aproximar do resultado obtido no MD5. Para comprovar esta hipótese, realizamos testes de tal forma que as buscas pelos identificadores dos conceitos na DHT *Chord* fossem realizadas sempre em ordem crescente, obedecendo o sentido horário no anel. Nesta situação específica comprovamos que houve uma redução no número médio de saltos no *Chord* usando-se uma função LSH, comparando com o resultado do MD5 nas mesmas condições. Estes resultados não são mostrados aqui por falta de espaço.

Observa-se também que a média é sempre menor utilizando-se o conjunto LSH e *Small World*, independente do número de nós da rede P2P e que, à medida em que a rede cresce, a média na rede *Chord* se distancia cada vez mais da média na rede *Small World*, tanto para a função de *hash* LSH quanto para o *hash* MD5, comprovando a característica de baixo número de saltos desta rede, independente do seu tamanho.

A medida do número de saltos também é válida e obedece às propriedades (1), (2), (3) e (5): independente do sentido, o número de saltos é sempre positivo (1) e possui valor 0 (nenhum salto) caso o identificador de destino esteja armazenado no próprio nó que origina a busca (2) e (3); em se tratando de um algoritmo de roteamento guloso, deve-se necessariamente obedecer a regra do triângulo em um espaço geométrico (5). Entretanto, pela assimetria no estabelecimento dos contatos, a regra (4) não é obedecida, comprovado principalmente nos resultados obtidos pelo *Chord*.

Como conclusão destes testes pode-se afirmar que houve uma redução no número de saltos necessários à recuperação de conteúdos similares utilizando-se uma rede *Small World* e funções LSH. A redução da média do número de saltos tende a ser maior com o aumento do número de nós na rede.

6. Conclusão

A primeira parte dos resultados obtidos neste trabalho mostra que a geração de identificadores utilizando uma função LSH, mantém a similaridade dos conteúdos classificados segundo uma ontologia, o que implica em um armazenamento mais próximo no espaço de endereçamento. Essa proximidade propicia uma redução no custo necessário para a recuperação de conteúdos similares, facilitando-se assim as buscas em redes P2P.

Para confirmar essa afirmação, o artigo propôs a criação de uma rede P2P que exiba as duas principais características do modelo *Small World*: 1) baixo número de saltos entre quaisquer dois pontos; 2) um pequeno grau de clusterização, sem que isso transforme a rede em um grafo regular. A rede proposta foi baseada em [Girdzijauskas 2009]. Os resultados obtidos compararam o número médio de saltos necessário para recuperação de conteúdos similares em uma rede P2P *Chord* e na rede P2P *Small World* proposta neste trabalho, confirmando-se a hipótese inicial de redução no custo de recuperação de conteúdos similares através do conjunto LSH e redes *Small World*.

Para realização de trabalhos futuros, pretendemos reconstruir a rede *Small World* segundo uma função de distribuição de probabilidade obtida através de dados extraídos em redes sociais utilizadas como referência. Isso nos permitirá que os contatos sejam estabelecidos não somente em função da distância, mas também em função da distribuição do conteúdo nesta rede, de tal forma que exista uma maior probabilidade de estabelecimento de contatos em regiões do espaço de identificadores que possuam conteúdos similares.

Além disso, pretendemos avaliar os resultados em outras formas de organização de redes estruturadas.

Nas simulações realizadas neste trabalho não foi considerada nenhuma dinâmica para representar a entrada e saída de nós na rede P2P. Sabemos que esta é uma limitação deste trabalho e uma característica importante presente nestas redes. Entretanto, o objetivo principal deste trabalho era a confirmação das hipóteses de agregação de identificadores de conteúdo gerados por uma função LSH com similaridade baseada em ontologias e a redução no custo de recuperação de conteúdos similares utilizado-se uma rede P2P *Small World*.

Referências

Charikar, M. S. (2002). Similarity estimation techniques from rounding algorithms. In *STOC '02: Proceedings of the thiry-fourth annual ACM symposium on Theory of computing*, pages 380–388, New York, NY, USA. ACM.

Girdzijauskas, S. (2009). *Designing Peer-to-Peer Overlays: a Small-World Perspective*. PhD thesis, Ecole Polytechnique Federale de Lausanne (EPFL), Lausane, CH.

Girdzijauskas, v., Datta, A., and Aberer, K. (2010). Structured overlay for heterogeneous environments: Design and evaluation of oscar. *ACM Transactions on Autonomous and Adaptative Systems*, 5(1):1–25.

Haghani, P., Michel, S., and Aberer, K. (2009). Distributed similarity search in high dimensions using locality sensitive hashing. In *Proceedings of 12th International Conference on Extending Database Technology*.

Indyk, P. and Motwani, R. (1998). Approximate nearest neighbors: towards removing the curse of dimensionality. In *STOC '98: Proceedings of the thirtieth annual ACM symposium on Theory of computing*, pages 604–613, New York, NY, USA. ACM.

Kleinberg, J. (2000). The small-world phenomenon: an algorithm perspective. In *STOC '00: Proceedings of the thirty-second annual ACM symposium on Theory of computing*, pages 163–170, New York, NY, USA. ACM.

Labovitz, C., Iekel-Johnson, S., McPherson, D., Oberheide, J., Jahanian, F., and Karir, M. (2009). Internet Observatory 2009 Annual Report. Technical report, 47th North American Network Operators' Group (NANOG47), Dearborn, MI.

Lua, E. K., Crowcroft, J., Pias, M., Sharma, R., and Lim, S. (2005). A survey and comparison of peer-to-peer overlay network schemes. *IEEE Communications Surveys and Tutorials*, 7:72–93.

Polyvyanyy, A. (2007). Evaluation of a Novel Information Retrieval Model: eTVSM. Master's thesis, Universitat of Potsdam, Hasso Plattner Institut, Potsdam, Deutschland.

Uschold, M. and Gruninger, M. (2004). Ontologies and semantics for seamless connectivity. *SIGMOD Rec.*, 33(4):58–64.

Watts, D. J. and Strogatz, S. H. (1998). Collective dynamics of 'small-world' networks. *Nature*, 393(6684):440–442.

Zhu, Y. (2005). *Enhancing Search Performance in Peer-to-Peer Networks*. PhD thesis, Computer Science and Engineering, University of Cincinnati.