Relações entre a Topologia de rede e a Geografia dos servidores WWW em Portugal

EXPOSTO, José; PINA, António; MACEDO, Joaquim; ALVES, Albano; RUFINO, José;

resumo.

Esta comunicação centra-se no estudo das localizações geográficas e da topologia de rede da WWW portuguesa, baseado em recolhas de rotas de encaminhamento de rede. O objectivo é a identificação das nuvens de densidade entre os servidores WWW, utilizando, para tal: medidas de proximidade temporal baseadas nos tempos médios de ida e volta (RTT), de forma a determinar as localizações mais apropriadas para a instalação de robôs cooperativos que minimizem o tempo de descarga global das fontes de informação. O processo é realizado através do recurso a técnicas vulgares de aglomeração num ambiente experimental, recorrendo a dados recolhidos, activamente, através de sondas colocadas na Internet, para criar uma grafo de distâncias fim-a-fim e calcular a distância entre os arcos dados pelos os tempos médios de ida e volta de uma comunicação de Internet entre quaisquer dois servidores. Um outro objectivo, é a confirmação da existência de relação entre a distância lógica (RTT) e a distância física (Geográfica). Os resultados dos experimentos realizados vêm confirmar as hipóteses iniciais pelo que estão abertos caminhos para novas experiências no âmbito do estudo entre as relações entre as topologias de rede e Internet e a geografia dos servidores.

PALAVRAS-CHAVE: WWW Portuguesa, Robôs Cooperativos, Geografia de Servidores.

INTRODUÇÃO

Actualmente, é incontestável a importância da WEB como fonte incomensurável de informação. Os factos e os números justificam essa importância e alimentam a investigação na área da Recuperação de Informação (RI) [1] A questão é que, face à imensidão dos dados a tratar, o acesso às fontes de informação realizado por robôs da WWW [5],tem vindo a desembocar num gargalo que é a recolha e a indexação de recursos da WWW [6]. Com efeito, as soluções tradicionais, baseadas em sistemas centralizados, não têm sido capazes de responder da forma mais eficaz aos problemas de escala relacionados com: o acesso às fontes de informação, o processamento dos dados e o seu armazenamento e ainda a disponibilidade global do serviço [2, 3].

O estabelecimento de parcerias envolvendo diferentes instituições, dispersas geograficamente, poderia vir a tirar partido das especializações temáticas ao nível do conteúdo para a construção de um sistema informacional integrado, com vista à utilização racional dos recursos de computação e de comunicação disponíveis, em cada local, para a extracção e tratamento de documentos acessíveis na WWW e a sua posterior disponibilização, da forma mais conveniente aos potenciais interessados. No caso específico da extracção de documentos, seria dessa forma possível a utilização orquestrada dos robôs para criar um poder acrescido para a descarga de documentos através da Internet.

Numa outra dimensão, em cada país, ou região, a exploração do acesso à Internet está entregue à responsabilidade de diferentes fornecedores de serviços de acesso à Internet, públicos e privados, cada um dos quais com diferentes capacidades, quer no que diz respeito aos meios disponibilizados quer na variedade e qualidade das ligações que proporcionam às fontes de informação.

É este quadro que alimenta e orienta a investigação dos meios e técnicas que visam a descoberta dos padrões de dispersão e de concentração dos pontos geográficos na origem da informação que a nosso ver podem contribuir para a definição de estratégias integradas concertadas de forma a poder configurar da forma mais apropriada a actividade dos robôs. É neste sentido que apontam, também, vários projectos de investigação relacionados com as tecnologias emergentes de P2P [4].

OBJECTIVOS

Este trabalho enquadra-se, no tema geral da investigação associado à recolha e pesquisa de informação na WWW, visando, em particular, estudar os modelos e as técnicas que visam a optimização do funcionamento de robôs, distribuídos e cooperativos, suportados por máquinas paralelas baseadas em tecnologias de *Cluster*. Neste contexto, são de salientar tópicos já identificados tais como: a proximidade dos robôs às fontes de informação, a comunicação gerada entre os robôs, no encaminhamento de informação e o balanceamento da carga de trabalho entre robôs.

A abordagem, baseia-se na definição de um modelo de partição do espaço WWW baseado na recolha de informação geográfica e nas topologias de rede e de hiper-ligações do espaço WWW, utilizando medidas de similaridade tais como: a distância física, a proximidade temporal e o fluxo de hiper-ligações entre os servidores.

Esta comunicação centra-se no estudo das localizações geográficas e da topologia de rede da WWW portuguesa, baseado em recolhas de rotas de encaminhamento de rede, com vista à descoberta dos tempos médios de ida e volta (RTT) entre quaisquer servidores WWW. O objectivo é a identificação das nuvens de densidade entre os servidores WWW, utilizando, para tal: medidas de proximidade temporal baseadas nos RTT, de forma a determinar as localizações mais apropriadas para a instalação de robôs cooperativos que minimizem o tempo de descarga global das fontes de informação. Para além disso, pretendem-se verificar a existência de alguma correlação entre a distância lógica (RTT) e a distância física (Geográfica). Na eventualidade de se verificar tal correlação, a proximidade lógica poderá ser deduzida proporcionalmente da proximidade física, em situações em que a primeira não esteja disponível.

RECOLHA DE ROTAS

A descoberta de rotas, enquadra-se no problema, mais geral, da descoberta de topologias de rede para o qual existem muitas abordagens, nenhuma das quais, até ao momento, conseguiu capturar totalmente topologia real da Internet [10].

Tipicamente, a descoberta de topologias visa encontrar as rotas para o maior número de destinos possível. O nosso caso restringe os destinos a um número predeterminado servidores WWW, podendo considerar-se que a detecção de rotas pode ser realizada através de um simples mecanismo de sonda, abdicando, assim, de mecanismos mais complexos que procuram novos destinos. É ainda de salientar a possibilidade de abdicar, também, da descoberta de alguns pontos intermédios nas rotas de grão mais fino, tais como os *swicthes* e concentradores, por não ser desejado tal nível de detalhe. Razão pela qual, as ferramentas que capturam pontos intermédios, abaixo da camada 3 da pilha OSI, são contra-indicadas para as experiências realizadas, sendo apenas recolhidos os pontos intermédios correspondentes a encaminhadores de rede de nível 3.

A ferramenta usada para a recolha de rotas é o *traceroute* [14] , utilitário que permite conhecer o caminho desde uma origem até cada um dos destinos, através da enumeração dos sucessivos RTT entre pontos intermédios e o destino final. Apesar de utilização imediata e prática, medidas de segurança tomadas pelos administradores de redes levam, muitas vezes, à sua desactivação em alguns dos encaminhadores, impedindo a detecção de algumas rotas. A partir das rotas descobertas, pode ser gerado um grafo e posteriormente calculadas, aproximadamente, as distâncias (RTT) entre quaisquer dois servidores WWW, utilizando algoritmos de cálculo de caminhos mais curtos em dois vértices num grafo, que designamos por grafo de *distâncias fim-a-fim*.

A comunicação na Internet é feita através do envio de mensagens tipicamente fragmentadas em pacotes, nos quais está indicado (entre outros) o que se designa por o tempo de vida TTL. A inclusão do TTL nos pacotes de rede é um mecanismo de protecção que evita a sua circulação por períodos excessivos, evitando congestionamento. O *traceroute* tira partido desta funcionalidade enviando pacotes com TTL sucessivamente crescente iniciado em 1 até o destino ser alçado. Quando o TTL atinge o valor 0, os sucessivos encaminhadores devolvem uma mensagem à origem, permitindo desta forma identificar o caminho percorrido pelos pacotes até ao seu destino.

Em cada salto, o *traceroute* devolve a indicação do RTT desde a origem até esse ponto. Calculando a diferença dos RTT em cada salto, ou seja, o ponto seguinte e anterior, obtém-se uma aproximação do RTT entre dois pontos intermédios. É, assim, possível obter um grafo ponderado, em que os vértices terminais são os destinos (servidores WWW) e os nós intermédios são os encaminhadores. Cada arco do grafo corresponde a um salto detectado pela sonda, sendo-lhe atribuído um peso correspondente à diferença entre o RTT do vértice mais afastado e o vértice mais próximo.

A recolha de rotas utilizando uma única, ou poucas, origens mostra-se incapaz de detectar ligações cruzadas entre nodos destino. Este problema pode ser minimizados aumentando o número de origens das sondas. Devido a questões operacionais, durante este trabalho, apenas foram usadas duas sondas. Esta limitação tem como consequência a variação da quantidade de rotas com o aumento do número de sondas facto que será objecto de estudo na secção de análise de dados.

LOCALIZAÇÃO GEOGRÁFICA DOS SERVIDORES WWW

A Internet tal como foi, inicialmente, projectada não permite a determinação da posição geográfica das máquinas que usam os seus serviços. Para colmatar esta deficiência, em 1996, ao já estabilizado Serviço de Resolução de Nomes (DNS) [16] foi adicionada uma nova funcionalidade que permite o registo da localização geográfica das máquinas da Internet, através da uma nova facilidade que permite expressar essa informação (ver RR LOC [17]). Na prática a sua utilização é, ainda hoje, escassa, motivo pelo qual a sua utilização neste trabalho foi limitada a situações em que não é possível determinar a localização de uma máquina através de outros que a seguir descrevemos, embora com uma baixa taxa de sucesso.

Existe um outro serviço de directório chamado *whois* [15] suportado, entre outras, por entidades responsáveis pelo registo e gestão dos endereços IP, como, por exemplo, a RIPE NCC (http://www.ripe.net/) e a ARIN (http://www.arin.net/). No entanto, apesar da grande quantidade informação disponibilizada por este serviço a localização geográfica dos IPs também não está contemplada. A primeira disponibiliza no seu serviço informação regional europeia e também espelha a informação das entidades de outras regiões do globo. Por esta razão, no que segue a sua base de dados *whois*, continua a

ser o recurso de informação mais completo que é do nosso conhecimento motivo pelo qual veio a tornar-se em lugar privilegiado de consulta durante todo o desenvolvimento deste trabalho.

Partindo do endereço IP de uma máquina a interrogação da base de dados *whois* pode gerar a seguinte informação:

- O bloco de endereços a que o endereço IP pertence. Tipicamente, os blocos de endereços são atribuídos a instituições geridas pelos fornecedores de serviço (entidade INETNUM, na Figura 1);
- A rota (agregado de endereços) anunciada por um Sistema Autónomo (AS) para os outros e que inclui um conjunto de blocos de endereços do mesmo AS (entidade ADDRESS_AGR, na Figura 1);
- O Sistema Autónomo que anuncia o agregado de endereços. Um AS é uma entidade lógica constituída por um conjunto bem definido de endereços de máquina e que é da responsabilidade de um fornecedor de serviço ou instituição (entidade AS, na Figura 1).

Para assistir na construção das rotas desenvolvemos uma base de dados a que corresponde o diagrama relacional apresentado na Figura 1 a que, doravante, sempre que necessário recorreremos para ilustrar a sua utilização. Desta forma, é possível por consulta àquela base de dados, para cada novo salto obtido da execução do *traceroute*, obter informação sobre a localidade geográfica do, respectivo, endereço IP. Em caso de insucesso da interrogação, a base dados *whois* é consultada e as respostas tratadas de forma a determinar a informação geográfica pretendida. Seguidamente, aquela informação é usada para actualizar a base de dados local.

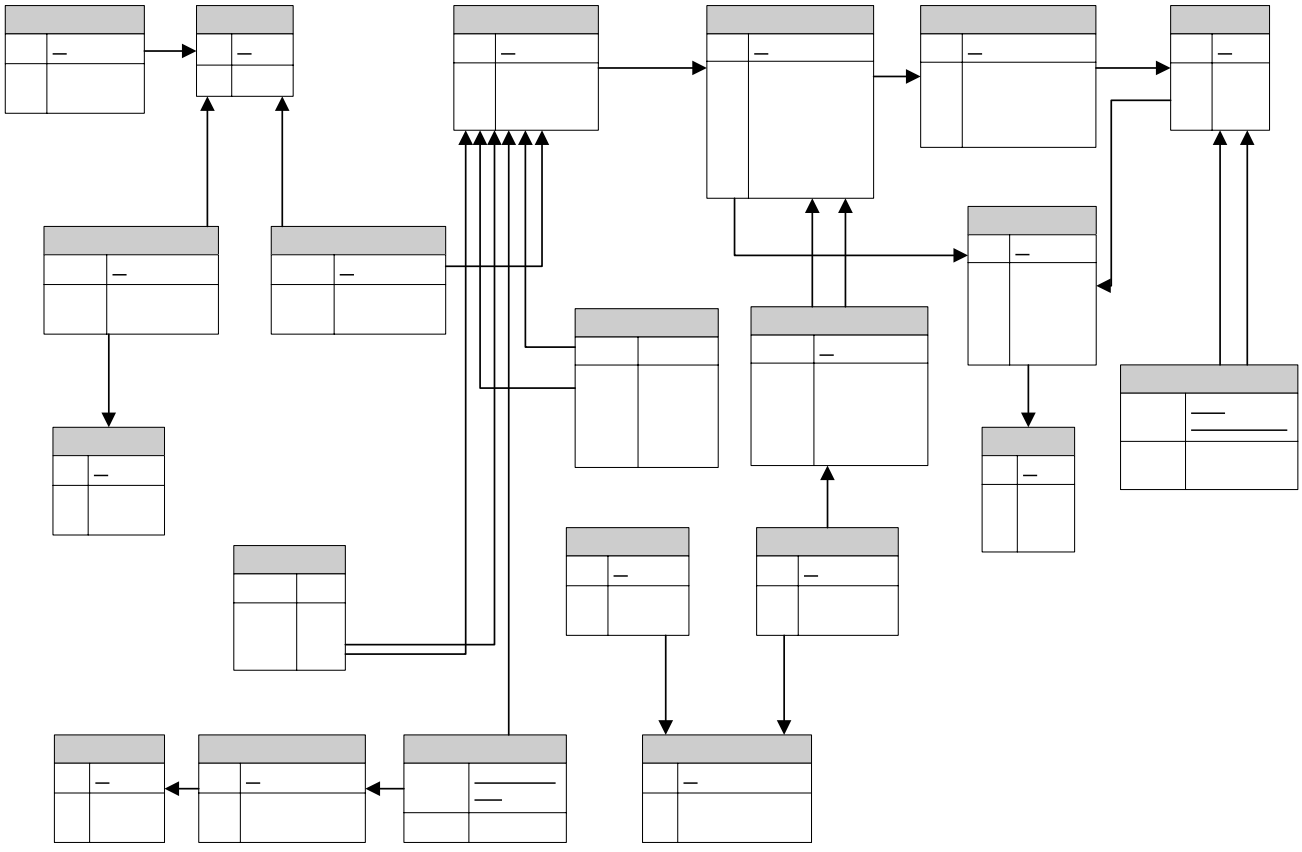

Figura 1 – Diagrama relacional da base dados utilizada para armazenar a informação recolhida

A informação obtida a partir das moradas administrativas introduz, em muitos casos erros assinaláveis. A exemplo, cita-se o caso da situação, detectada, das máquinas situadas no Porto no sistema autónomo AS da FCCN (responsável pela gestão da rede nacional universitária), terem registadas a morada dos administradores, que por sinal são de Lisboa. Não nos foi, ainda, possível estimar a dimensão dos desvios introduzidos por causas de situações erróneas como a que apresentámos acima.

Existe, neste momento, um projecto, designado por *NetGeo* [12] cujo objectivo é precisamente registar a localização exacta geográfica de um endereço IP ou de um sistema autónomo AS - nome da cidade do IP, nome do país e coordenadas geográficas - através da análise das moradas dos respectivos responsáveis, administrativos, esses sim registados no *whois*..

Porém, em várias situações, o *NetGeo* não conseguiu identificar parte ou nenhuma informação relativa a localização geográfica e nalguns casos o grão de precisão das coordenadas é passado para o nível do país.

De forma a ultrapassar as dificuldades decorrentes das imprecisões geográficas obtidas através dos elementos citados acima recorremos, ainda, ao *GeoNames*[11], designação usada para uma base de dados que possui as coordenadas geográficas de cidades de diversos países entre os quais, Portugal, Bélgica, Brasil, Canadá, Dinamarca, França, Alemanha, Holanda, Espanha, Suécia, Suiça e Reino Unido, num total de 659.048 pontos.

A determinação da localização geográfica de um IP (ou AS) quando a informação geográfica não se encontra completa passa pelas três fases de consulta que se apresentam a seguir. Na impossibilidade de encontrar a localização exacta, a cidade é dada como desconhecida.

1. Consulta ao *NetGeo* por IP ou AS;
2. Consulta ao *GeoNames*, quando são desconhecidas a cidade ou as coordenadas geográficas;
3. Consulta ao DNS LOC, quando são desconhecidas a cidade ou as coordenadas geográficas;

Na figura 2 podem ver-se a distribuição geográfica das cidades em que existem IPs (pontos no mapa e respectivo nome) e os saltos existentes entre duas cidades (segmentos rectos). Os mapas contemplam ainda, em mais pormenor, a América do Norte e a Europa tendo sido gerados pela aplicação *GeoMedia Professional* [18].

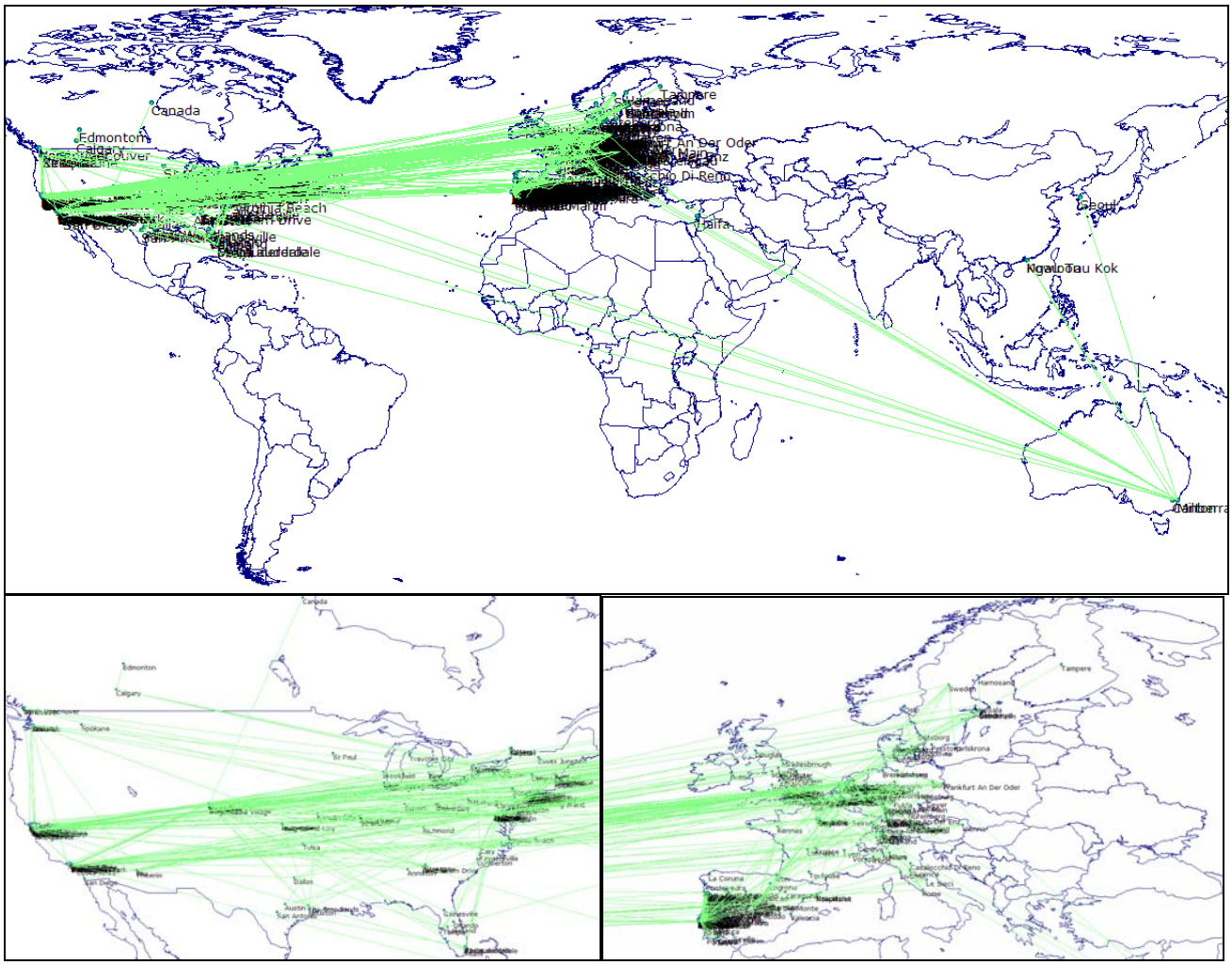

Figura 2 – Localização geográfica das cidades com IP identificados e caminhos das rotas descobertas

APRESENTAÇÃO DOS DADOS RECOLHIDOS

Os dados que fundamentam o trabalho desenvolvido resultam das rotas obtidas, a partir de um conjunto inicial de 33413 servidores WWW, dos quais apenas 8859 não estavam em linha, identificados pelo NetCensus [9]. Este projecto tem como objectivo a descoberta de servidores WWW, através de um mecanismo de robô, a fim de caracterizar a WWW portuguesa e estudar a sua evolução. No que segue, fazemos uma descrição pormenorizada dos dados recolhidos.

A figura 3 apresenta os valores totais e as médias de IPS e número de nomes de máquina por IP. Foram encontrados 3998

nomes de encaminhadores situados em 3994 máquinas, o que significa que, no caso dos encaminhadores, existe em média um nome por cada IP. Os 24554 nomes diferentes de servidores estão distribuídos por 9292 IPs, o que dá como resultado uma média de cerca de 3 nomes diferentes de servidores por cada IP.

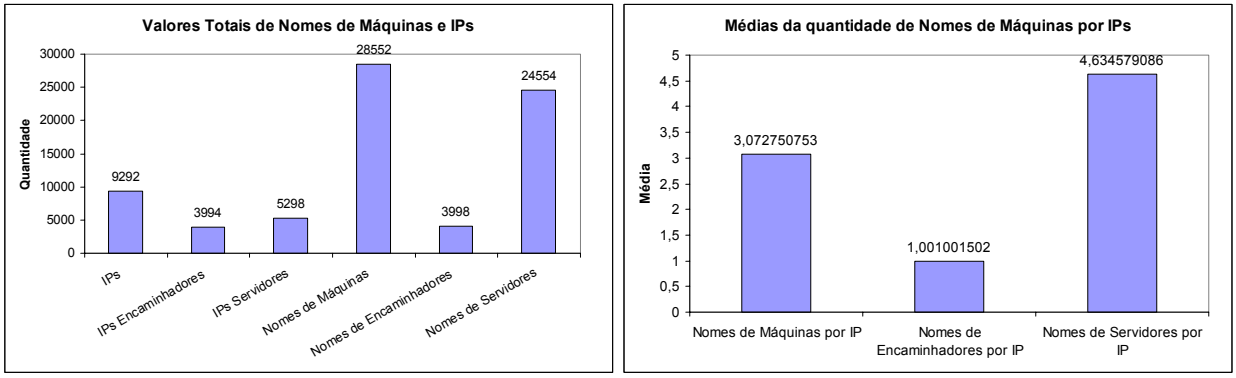

Figura 3 – Valores totais da informação recolhida e da quantidade de nomes de máquinas por IP

A Figura 4. mostra um histograma da distribuição da quantidade de nomes de máquinas por IP. Verifica-se, o que já era esperado, a existência de um número considerável de IPs com um número elevado de nomes, o valor máximo corresponde a um só IP que regista 3670 nomes diferentes de máquinas; tipicamente, estes tipos de máquina são servidores para alojamento de páginas. Quanto aos encaminhadores, a quase totalidade dos IPs possui, apenas, um nome de máquina.

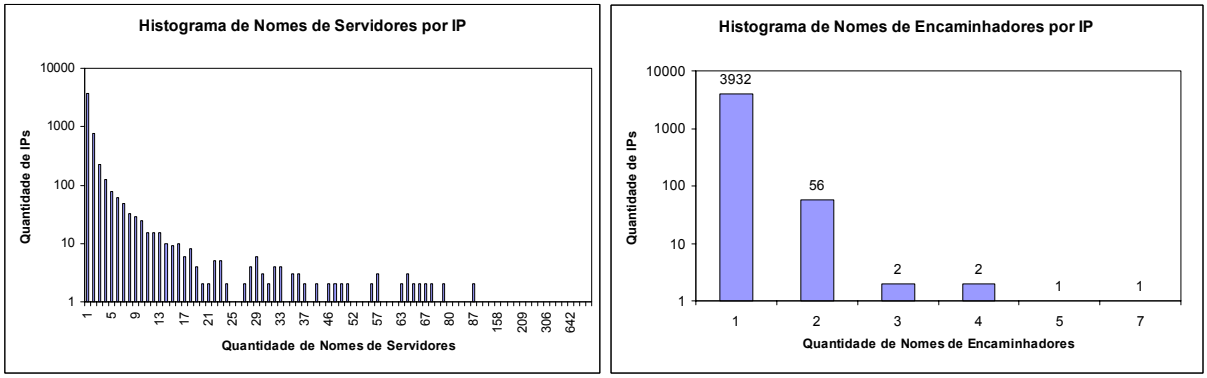

Figura 4 – Histogramas da quantidade de nomes de máquinas por IP

Em termos de quantidades de IPs por nomes de máquinas (ver Figura 5) a situação inverte-se em relação à figura anterior. Aqui, embora a média total seja próxima do valor 1, chegou a encontrar-se um nome de encaminhador com 18 IPs, enquanto que para os servidores o número máximo de IPs por nome de servidor foi 2.

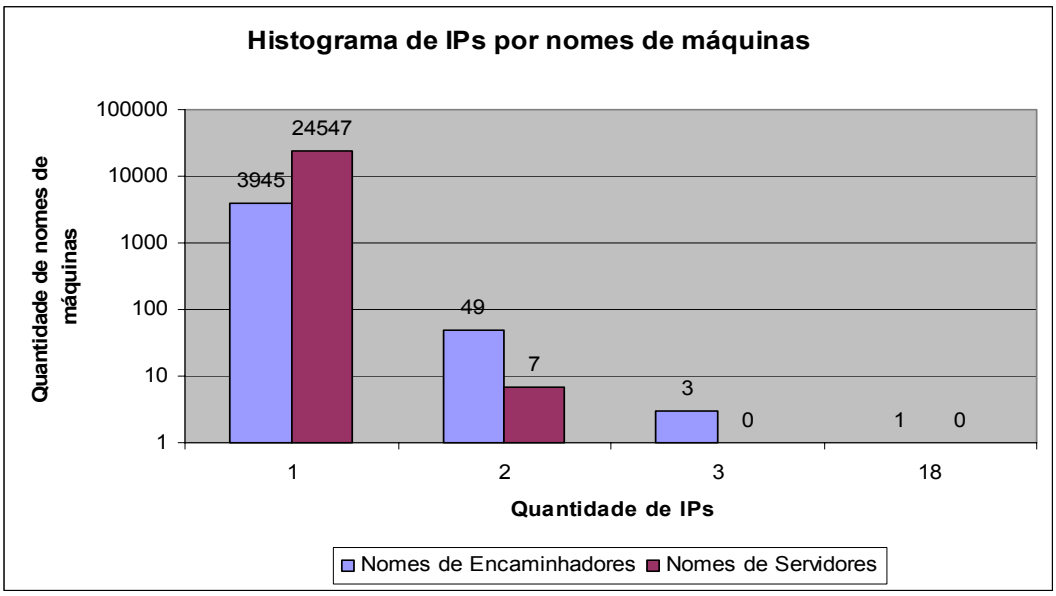

Figura 5 – Histogramas da quantidade de IPs por nomes de máquinas.

Na Figura 6 apresentam-se as distribuições dos IPs por país. Saliente-se que, embora o conjunto inicial de nomes de servidores utilizadas como ponto de partida para descobrir as rotas fossem todos do domínio **'pt'**, a realidade veio a verificar-se que apenas 66% dos IPs daqueles servidores têm localização geográfica em Portugal.

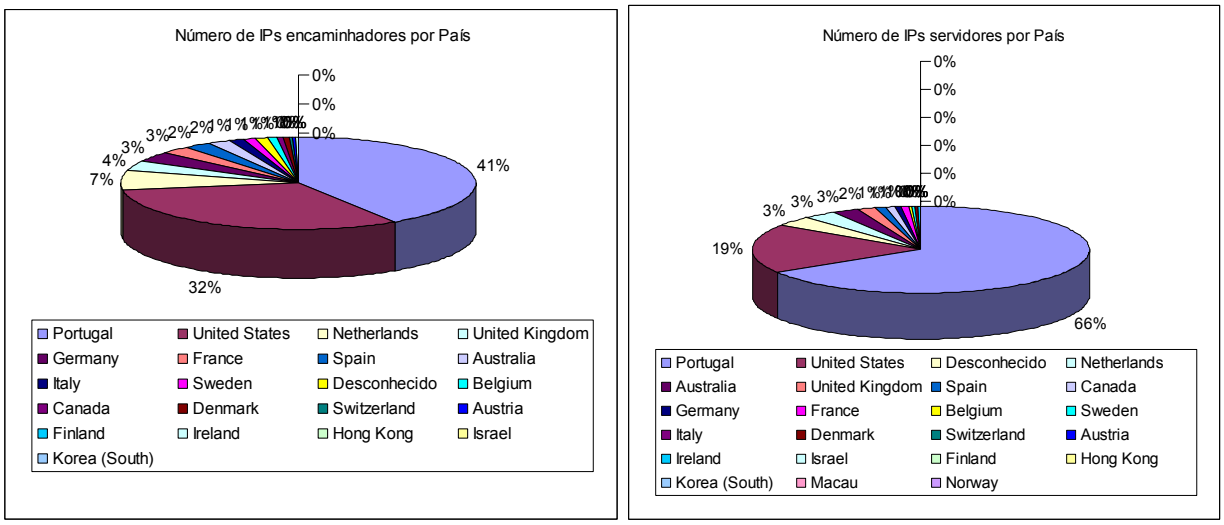

Figura 6 – Distribuição dos IPs por País

Finalmente, a Figura 7 mostra a distribuição dos servidores por sub-domínios de raiz **pt**: sendo que, o nível 1 corresponde a um domínio **x.pt**, o nível 2 a **y.x.pt** e assim sucessivamente. Verifica-se que 71% dos servidores são do tipo **xxx.abc.pt**.

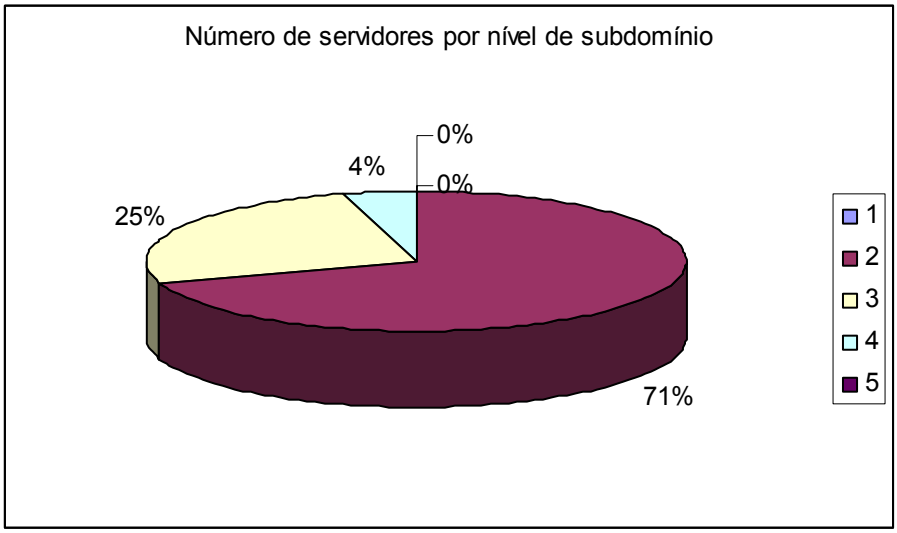

Figura 7 – Número de servidores por nível de sub-domínio

VALIDAÇÃO E ANÁLISE DOS DADOS

A recolha de rotas é um processo moroso que para o universo inicial 33413 nomes de servidores iniciais, pode atingir os 11 dias. Por sua vez, o tempo de cálculo para calcular as distâncias fim-a-fim entre todos os servidores foi estimado em 5 dias. Uma vez que uma única corrida nunca é suficiente para garantir a coerência dos dados recolhidos, tivemos que, naturalmente, de proceder a várias iterações. Por esta razão, para se poderem obter resultados em tempo útil tivemos de escolher e validar amostras reduzidas do universo inicial.

Obtenção de amostras

Uma vez que o que está em estudo é a localização geográfica dos endereços IP, utilizamos a técnica de amostragem aleatória estratificada com repartição proporcional, em que os estratos correspondem às cidades, para obter uma amostra com uma dispersão razoável pelas diferentes cidades. Apesar de não se pretender estimar nenhum parâmetro da população alvo a partir desta amostra, determinámos o tamanho da amostra, com base na estimação da média dos tempos entre dois IPs, utilizando a fórmula: $n = \frac{Z^2 S_0^2}{(\overline{X} - \mu)^2}$, em que Z é o valor crítico da distribuição normal padronizada associado à metade do complemento do nível de confiança $\frac{(1-\alpha)}{2}$, S_0 é a estimativa inicial do desvio padrão e $\overline{X} - \mu$ é a diferença de erro admitida entre a média da amostra (μ) e a média da população $\overline{X}$.

Embora não se possa confirmar que o parâmetro RTT siga uma distribuição normal, o teorema do limite central [13] prova que, independentemente, da população seguir uma distribuição normal, as médias amostrais são normalmente distribuídas quando o tamanho da amostra é suficientemente grande (n>30).

Utilizando um factor de confiança de 90% ($Z = 1{,}645$), podemos igualar $S_0 = 76{,}435$ que foi o valor obtido para o desvio padrão dos RTT entre dois IPs, no experimento inicial.

A existência de desvios padrão tão elevados, causados pelas diferenças significativas de qualidade nas linhas de comunicações existentes, não permite assegurar uma diferença de erro suficientemente reduzida. Utilizaremos, por isso, $\overline{X} - \mu = 4$, resultando o tamanho da nossa amostra em $n = 988$ destinos a considerar. Como a nossa população de estudo ($N = 5298$, correspondente ao número total de IPs) é finita e $N \leq 20n$ ($5298 \leq 19760$), deve utilizar-se o factor de correcção para populações finitas (CPF), em que $n_{CPF} = \frac{nN}{n + (N-1)}$, resultando o tamanho da amostra em $n_{CPF} = 833$.

Utilizando a repartição proporcional, cada estrato ficará com $n_i = N_i \frac{n}{N}$ elementos, em que n é o número total da amostra (agora n_{CPF}), N o tamanho da população e N_i, o número de elementos da população em cada estrato. Após o,

necessário, arredondamento de números a amostra fixou-se no valor final de 813 servidores.

Variação do número de sondas

A fim de compreender o efeito da utilização de pontos de origem distintos, para a obtenção das rotas foram efectuadas duas recolhas a partir de duas origens distintas. O objectivo é determinar o benefício na utilização de um número superior de sondas.

Os resultados foram observados utilizando duas medidas: a sobreposição entre vértices (IPs) (S_v) e a sobreposição de arcos (saltos) (S_a). Uma vez que apenas se pretende considerar a intercepção dos arcos, sem considerar os respectivos pesos (RTT), utilizou-se a medida de similaridade binária de Dice [1], dada pela fórmula: $S = 2\frac{|A \cap B|}{|A| + |B|}$, em que A e B são os conjuntos de vértices e arcos de cada uma das recolhas. Os resultados forma os seguintes: para os vértices obteve-se $S_v = 0{,}647$, sem considerar os vértices de destino, uma vez que são iguais, enquanto que para os arcos $S_a = 0{,}544$. Pode concluir-se que mais de metade dos vértices e arcos é comum nas duas recolhas.

O acréscimo de benefício na utilização de ambas as origens foi calculado através da noção de cobertura, dos vértices de uma das origens em relação à reunião das duas origens através da fórmula, $C = \frac{|A|}{|A \cup B|}$. Os resultados foram os seguintes: para os vértices o valor foi $C_v = 0{,}736$, o que significa que com uma origem apenas obtém-se uma cobertura próxima de 74% para os vértices quando comparada com o dobro das origens. Para os arcos a cobertura é ligeiramente inferior, $C_a = 0{,}69$ o que corresponde a uma cobertura de arcos próxima de 70%.

Conclui-se, como se previa, que o aumento do número de sondas contribui para o acréscimo na quantidade de rotas e encaminhadores, possibilitando um cálculo mais aproximado das distâncias fim-a-fim. Convém ter em atenção que o número de sondas não deve ser substancialmente elevado, por motivos de ordem prática.

Correlação dos RTT vs. Geográfica

Antes de procedermos ao cálculo das localizações que optimizam a descarga de páginas pelos robôs quisemos verificar de que modo a distância lógica acompanha a distância geográfica, através da medida padronizada de relação de duas variáveis, dada pela correlação.

Sendo X a variável associada à distância entre duas cidades e Y a variável associada à média dos RTT entre os servidores da amostra contidos entre duas cidades, a correlação é dada por $\rho_{xy} = \frac{\sigma_{xy}}{\sqrt{\sigma_x^2 \cdot \sigma_y^2}}$, em que σ_{xy} é a co-variância entre X e Y, e σ_x^2 e σ_y^2 as variâncias respectivas.

Para verificar a correlação entre os RTTs e a distância geográfica utilizamos a amostra com 813 IPs, medindo o RTT desde a origem até cada um dos destinos. O resultado é apresentado no gráfico de dispersão, da Figura 8, que representa uma linha de regressão que estabelece uma correlação de 0,473.

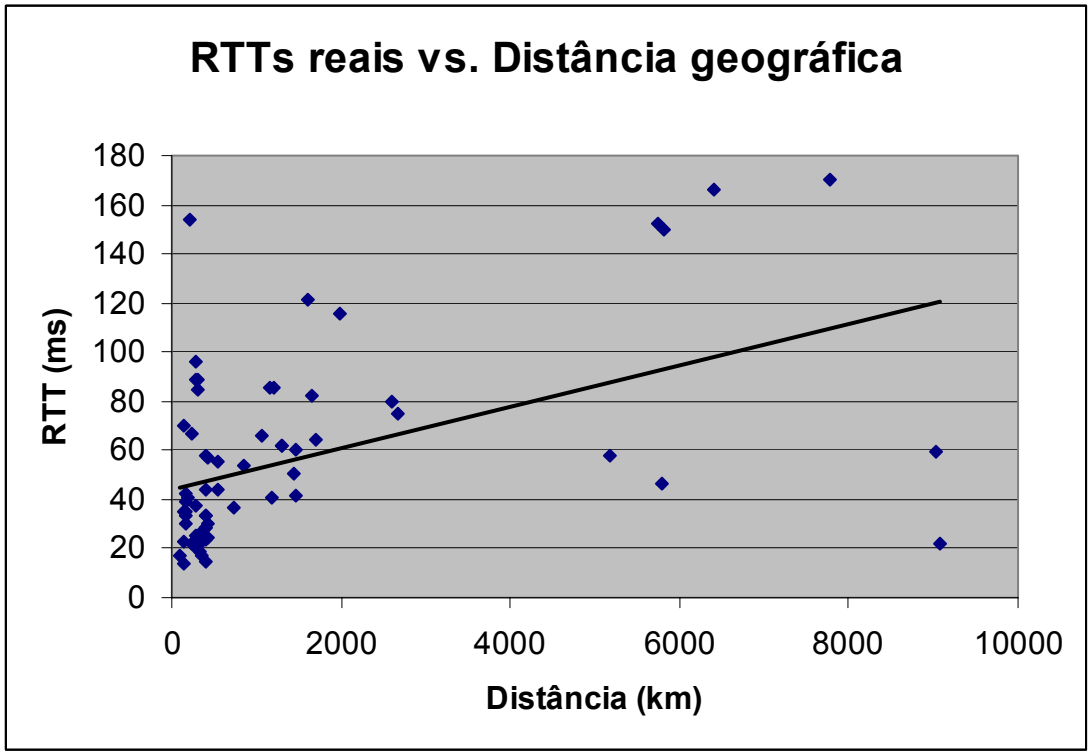

Figura 8 - Gráfico de dispersão entre os RTT reais e a distância geográfica

Pode afirmar-se que, efectivamente, existe um aumento dos RTTs com o aumento da distância geográfica, não sendo, contudo um valor demasiado pronunciado. Podem apontar-se vários factores para que a correlação não seja mais evidente, tais como:

- A instabilidade das comunicações. Verificou–se um desvio padrão de RTT para o mesmo destino bastante elevado.
- Organização do encaminhamento de rede. Por exemplo, foram encontradas situações em que caminhos para a mesma cidade são sujeitos a longos desvios por outras cidades, como é o caso dos acessos residenciais.
- Cidades mal representadas geograficamente. Problema que nunca será completamente resolvido enquanto não for alterada a forma como é feita a análise das localizações levada a cabo pelo *Netgeo*.

Validação do grafo de distâncias fim-a-fim

Depois de estabelecido o patamar máximo para a correlação utilizando os RTT desde a origem até ao destino, estamos em condições de poder validar o grafo de distâncias fim-a-fim. Como já foi referido, este grafo resulta do cálculo das distâncias mais curtas entre todos os servidores WWW, gerado a partir da recolha das rotas para esses servidores. Utilizamos mais uma vez a amostra referida, tendo-se obtido o gráfico da Figura 9, com uma correlação de 0,264.

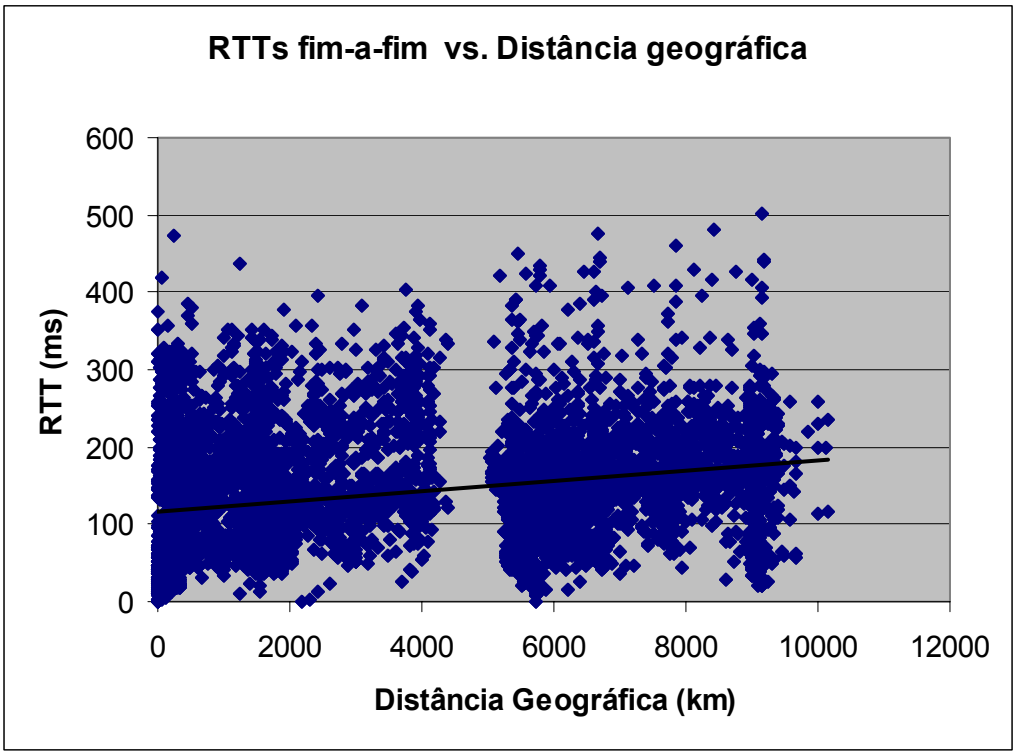

Figura 9 - Gráfico de dispersão entre os RTT do grafo de distâncias fim-a-fim e a distância geográfica

Este valor de correlação, como já era previsto é inferior ao dos RTTs reais, devido, precisamente, ao número de sondas utilizadas e as consequências que daí advêm, tal como já foi referido. Este argumento é reforçado, quando se analisa a mesma correlação de um grafo de distâncias fim-a-fim em que é utilizada a mesma origem dos RTTs reais, tendo-se obtido o valor de 0,414. Um valor muito próximo da correlação dos RTTs reais.

Nuvens de densidade dos servidores WWW

No quadro dos nossos objectivos, pretendemos optimizar a tarefa dos robôs no seu processo de descarga de páginas no contexto de um sistema de Recuperação de Informação distribuído. A descoberta de nuvens de densidade ou aglomerados de servidores WWW, pode contribuir para este objectivo se forem encontrados locais que aproximam um robô de um conjunto de servidores.

Os mecanismos de aglomeração procuram agrupar os objectos em que a similaridade entre eles é maximizada, com base na descrição das suas características e numa medida de similaridade calculada a partir destas, de tal forma que, os elementos pertencentes a um grupo apresentam uma forte semelhança entre os elementos do mesmo grupo e fraca entre os elementos dos restantes grupos. Existem diversas técnicas de aglomeração [7], no entanto, para o nosso caso, as técnicas baseadas no corte mínimo-máximo de grafos são as mais indicadas, devido à própria natureza dos dados que são aqui apresentados, modelados através de um grafo.

Recorremos à ferramenta CLUTO [19] por esta disponibilizar a técnica do corte mínimo-máximo e, ainda, uma grande facilidade de utilização e uma variedade substancial na definição de parâmetros. Para além disso, devolve como resultado a quantidade de aglomerados gerados, os objectos que fazem parte de cada aglomerado e, ainda, medidas da qualidade de aglomeração, quando a solução é comparada com informação externa conhecida à priori acerca das classes em que os objectos poderiam ser aglomerados.

A validação dos resultados de aglomeração é uma tarefa complexa, apenas, possível quando existe informação externa adicional acerca das previsões sobre qual o aglomerado que os elementos integram.

Definem-se vulgarmente duas medidas que avaliam a qualidade da aglomeração: a entropia e a pureza. Tipicamente, valores baixos de entropia e valores altos de pureza correspondem a boas soluções de aglomeração.

A entropia mede a distribuição dos elementos de um aglomerado pelas classes existentes. Para um aglomerado S_r com tamanho n_r a entropia desse aglomerado é dada por $E(S_r) = \frac{1}{\log q} \sum_{i=1}^{q} \frac{n_r^i}{n_r} \log \frac{n_r^i}{n_r}$, em que q é a quantidade de classes disponíveis e n_r^i o número de elementos da classe i atribuídos ao aglomerado r. Por seu turno, a pureza mede a

extensão de objectos atribuídos a um aglomerado da classe mais representativa, sendo dada por $P(S_r) = \frac{1}{n_r} \max_i(n_r^i)$.

Para uma solução de aglomeração podemos obter valores para a entropia e pureza com base nos valores individuais de cada aglomerado, multiplicando pelo peso associado ao tamanho de cada aglomerado, obtendo-se assim: $Entropia = \sum_{r=1}^{k} \frac{n_r}{n} E(S_r)$ e $Pureza = \sum_{r=1}^{k} \frac{n_r}{n} P(S_r)$.

A comparação dos valores de entropia e pureza apenas faz sentido quando se pretende estabelecer comparação da qualidade de aglomeração quando se variam as técnicas para obter os aglomerados. Efectivamente, foram comparadas várias técnicas de aglomeração, tendo sido a de corte minímo-máximo que obteve melhores resultados. A explicação e exposição dessas experiências estão fora do quadro da presente comunicação, razão pela qual nos abstemos de as apresentar.

Utilizamos, os valores da entropia e pureza para encontrar classes, previamente definidas, que incluam os servidores, como é o caso dos AS e das cidades. Desta forma, pode-se extrapolar qual das classes aproxima melhor a aglomeração calculada. Preparámos o CLUTO com os dados do nosso grafo de distâncias fim-a-fim entre todos os servidores WWW, aplicando a técnica de aglomeração referida. Utilizando as cidades como as classes que definem os elementos a integrar, com 20 aglomerados obtiveram-se valores para a entropia de 0,307 e para a pureza 0,577.

Surpreendentemente, para os AS verificaram-se valores de entropia e pureza ligeiramente piores (0,330 para a entropia e 0,502 para a pureza). Tendo em conta que os sistemas autónomos AS são zonas administrativas de redes, as cidades mantêm uma aproximação melhor dos aglomerados que os AS.

Variando o número de aglomerados verifica-se uma variação proporcional para a pureza (inversamente proporcional para a entropia), uma vez que, quanto mais próximo for o número de aglomerados, do número de classes, mais semelhante são os grupos de elementos.

Posicionamento dos robôs

Em termos de localização dos aglomerados, que nos permitirá descobrir a melhor localização dos robôs, faz todo o sentido calcular o centro de gravidade dos elementos que compõem cada aglomerado (centróides), através do cálculo do ponto médio das coordenadas geográficas dos elementos, uma vez que foi estabelecida alguma correlação entre os RTTs e a distância geográfica.

Torna-se evidente que os *centróides* irão corresponder a localizações abstractas e sem contexto geográfico, não deixando de ser, efectivamente, o ponto crítico que optimiza a proximidade com os elementos do aglomerado. Mesmo que aqueles pontos correspondam a localizações reais, condicionantes de várias ordem podem não ser compatíveis, em termos operacionais, com a instalação de robôs nesses pontos.

Uma hipótese de solução, imediata, parte da interpretação dos centróides resultantes da aglomeração dos servidores como os pontos que optimizam a descarga dos servidores desse aglomerado. Assim, os robôs deverão ser instalados nas localizações geográficas mais próximas desses pontos de forma a minimizar o desvio em relação aos valores óptimos dados pelos centróides.

DISCUSSÃO

A insuficiência de resposta das técnicas tradicionais de Recuperação de Informação no contexto da WWW, conduziu à necessidade de reformulação da sua arquitectura, assumindo uma visão distribuída e cooperativa de componentes. Em particular, os módulos de recolha de páginas destes componentes requerem uma afinação cuidada, no sentido, de permitir taxas de recolha elevada, assegurando uma frescura elevada de páginas nos repositórios locais.

Num sistema com as características enunciadas, julgou-se, por isso, pertinente um estudo prévio rigoroso acerca das topologias de rede, de forma a permitir descobrir nuvens densidades de servidores para poder deduzir um centro de massa em que o tempo de acesso fosse mínimo, a partir desse ponto, para todos os servidores dentro da nuvem.

A aproximação realizada para calcular o centro de massa da nuvem, obtida a partir das coordenadas geográficas dos seus elementos constituintes, for realizada a partir do cálculo da correlação entre as distâncias geográficas e os tempos médios de ida e volta (RTT) das comunicações entre servidores. A não concordância do centróide com um lugar geograficamente válido para a instalação de um robô, obriga a deslocá-lo para a cidade mais próxima dos elementos da nuvem, que minimiza a variação em relação ao valor óptimo que o centróide representa.

A geração do grafo de distâncias fim-a-fim vem, efectivamente, aproximar substancialmente a estimativa dos RTTs entre todos os servidores, embora se tenha concluído pela necessidade de utilizar um maior número de sondas. Embora, não disponhamos das condições necessárias para aumentar o número de sondas utilizado, é do nosso conhecimento mais recente a existência de "reflectores" do *traceroute* dispersos por diversas localizações no mundo, com os quais é teoricamente possível adquirir rotas a partir desses pontos.

Os resultados do trabalho realizado, estão naturalmente, comprometidos com a qualidade da informação sobre a localização geográfica dos IP, nomeadamente a que obtivemos através do *NetGeo*. É nesse sentido que consideramos a possibilidade de como trabalho futuro, desenvolver uma ferramenta semelhante de forma a corrigir as incorrecções e utilizar dados mais actualizados e precisos sobre a localização geográfica dos IPs

Finalmente, com o esquema de determinação de localizações óptimas de robôs para a optimização do seu processo de descarga, esperamos poder contribuir para a concretização de um sistema fidedigno e eficiente para a recuperação de informação aplicados à WEB portuguesa.

BIBLIOGRAFIA

1. C. J. van Rijsbergen. Information Retrieval, http://www.dcs.gla.ac.uk/Keith/Preface.html, 1979.
2. Sergey Brin and Lawrence Page, The anatomy of a large-scale hypertextual web search engine, 1998.
3. J. Cho e H. Garcia-Molina. Parallel crawlers, Proc. of the 11th International World Wide Web Conference, 2002
4. Steve Waterhouse. JXTA search: Distributed search for distributed networks, Sun Microsystems, Inc., 2001.
5. Marc Najork and Allan Heydon, High-Performance Web Crawling. Chapter 2 in J. Abello et al. (editors), Handbook of Massive Data Sets, Kluwer Academic Publishers, 2002.
6. José Exposto, António Pina, Joaquim Macedo, Albano Alves e José Rufino, Um Modelo Cooperativo e Distribuído para a Recuperação de Informação na WWW, 6ª Conferência sobre Redes de Computadores (CRC'2003), Bragança, Portugal, 2003.
7. Fasulo, D., An analysis of recent work on clustering algorithms. http://www.cs.washington.edu/homes/dfasulo/clustering.ps, 1999.
8. Caida - The Cooperative Association for Internet Data Analysis, http://www.caida.org/.
9. Leopoldo Silva, Joaquim Macedo, António Costa, Orlando Belo e Alexandre Santos. NetCensus: Medição da evolução dos conteúdos na web. Departamento de Informática, Universidade do Minho, 2002.
10. R. R. Siamwalla, R. Sharma, e S. Keshav, Discovering internet topology, in Proc. IEEE INFOCOM '99, 1999
11. National Geospatial-Intelligence Agency, Geographic Names Data Base, http://earth-info.nima.mil/gns/html/index.html.
12. CAIDA, NetGeo - The Internet Geographic Database, http://www.caida.org/tools/utilities/netgeo/, 2002.
13. Edward Minieka, Zoriana Dyschkant Kurzeja, Statistics for Business with Computer Applications, South-Western College Publishing, 2001.
14. V. Jacobson, traceroute, ftp://ftp.ee.lbl.gov/traceroute.tar.Z, 1989.
15. Harrenstien, K., Stahl M., and Feinler, E., NICNAME/WHOIS, RFC-954, SRI, 1985.
16. P. Mockapetris, Domain Names--Concepts and Facilities, RFC 1034 (Standard: STD 13), 1987.
17. C. Davis, P. Vixie, T. Goodwin and I. Dickinson, A Means for Expressing Location Information in the Domain Name System, RFC 1876, 1996.
18. Intergraph Corporation, GeoMedia Professional, http://imgs.intergraph.com/gmpro/, 2002.
19. George Karypis, CLUTO A Clustering Toolkit Manual, 2003.