Techniques de localisation géographique d'hôtes dans l'Internet

Bamba Gueye *, Artur Ziviani **, Mark Crovella ***, Serge Fdida*

** Université Pierre et Marie Curie*
Laboratoire d'Informatique de Paris 6
Paris France
*** Laboratório Nacional de Computação Cientifica (LNCC)*
Petrópolis - RJ - Brasil
**** Department of Computer Science*
Boston University

L'inférence de la localisation géographique d'hôtes dans l'Internet permet l'émergence de nouvelles applications très variées. Jusqu'à présent, la localisation d'un hôte cible est fournie par la position des hôtes références, hôtes dont on connaît les positions géographiques. Ainsi le nombre d'endroits possibles où on peut localiser un hôte cible est égal au nombre d'hôtes références, conduisant ainsi à un espace discret de réponses. Nous proposons une technique de Localisation Géographique basée sur la Multilatération (LGM) pour inférer la position géographique d'un hôte cible. La multilatération permet d'obtenir un espace continu d'endroits possibles où on peut localiser un hôte contrairement aux approches précédentes. LGM transforme les mesures de délai en distance géographique `surestimée`, malgré les délais supplémentaires dûs aux congestions, et la non linéarité des chemins entre les hôtes. LGM utilise la multilatération avec ces distances géographiques surestimées pour inférer la position de l'hôte. Les résultats obtenus montrent que LGM est plus performante que les précédentes techniques de localisation et, de surcroît, est capable d'attribuer une zone de confiance à chaque hôte localisé.

Keywords: Localisation, multilatération, mesures de délai

1 Introduction

Avec le développement des nouvelles technologies de l'information, des nouveaux services ont fait leur apparition, notamment des services dits de "proximité" basés sur la localisation des clients. Nous pouvons citer comme exemple la publicité ciblée, la sélection automatique de la langue à la connexion, la diffusion de contenu suivant une politique géographique, et l'acceptation d'une transaction bancaire seulement à partir d'un endroit pré-établi. Les techniques de localisation basées sur des mesures tentent de déterminer la position géographique d'un hôte en se basant sur la connaissance de son adresse IP. C'est ainsi que [PS01] proposent d'utiliser la position de l'hôte référence, hôte dont on connaît la position géographique, le plus proche en terme de délai comme possible localisation de l'hôte cible. Avec cette approche l'ensemble des endroits où on peut localiser un hôte cible est limité par le nombre d'hôtes références. Nous obtenons ainsi un ensemble discret de réponses.

Nous proposons une nouvelle technique de Localisation Géographique basée sur la Multilatération (LGM) pour résoudre ce problème. En effet, la multilatération permet d'estimer une position en utilisant un nombre suffisant de distances à partir de quelques points immobiles. Dès lors, elle fournit un ensemble continu d'endroits où on peut localiser la cible au lieu d'un espace discret de réponses. Connaissant la distance géographique *surestimée* entre la cible et chaque hôte référence, LGM fournit à l'instar du système de positionnement par satellites (GPS) [EM99] une estimation de localisation.

Pour l'évaluation de LGM nous avons utilisé les mesures de délai d'hôtes localisés à travers les États Unis et l'Europe de l'Ouest. Les résultats montrent que LGM est plus performante en précision que les

précédentes techniques de localisation géographique, basées sur des mesures. Ainsi l'erreur médiane de distance obtenue est inférieure à 25 km pour l'ensemble des hôtes localisés en Europe de l'Ouest, et 100 km pour ceux localisés aux États Unis. Nous avons remarqué que dans la plupart des cas, la zone de confiance que LGM fournit est raisonnable, car assimilable à la superficie d'un petit pays comme la Belgique en Europe ou un petit état comme le Maryland aux États Unis.

Ce papier est organisé comme suit. La section 2 dresse l'état de l'art du domaine et montre les contributions que LGM a apportées. Dans la section 3 nous présentons la technique LGM et ses différentes caractéristiques. La section 4 illustre les différents résultats obtenus en appliquant LGM. Enfin la section 5 conclut notre travail et présente quelques perspectives pour le long terme.

2 Techniques de localisation géographique

2.1 État de l'art

La RFC 1876 [DVGD96] propose d'ajouter des informations de localisation dans les noms DNS (Domaine Name Server). Cependant cette proposition ne fut pas largement adoptée, car les administrateurs n'étaient pas trop motivés pour ajouter des enregistrements de localisation dans les bases de données DNS. Padmanabhan et Subramanian [PS01] quant à eux ont développé trois techniques pour inférer la localisation géographique d'un hôte. La première technique GeoTrack infère la localisation de l'hôte cible à partir de son nom DNS ou bien celui de l'hôte le plus proche. Les noms DNS dans Internet contiennent parfois certaines indications sur la localisation. Par exemple bcr1-so-2-0-0.Paris.cw.net indique un routeur localisé à Paris. Toutefois l'estimation de localisation peut être imprécise car le dernier routeur reconnaissable qui donne sa position comme estimation n'est pas forcement proche de la cible.

La deuxième technique, GeoCluster, se base sur l'hypothèse que tous les hôtes qui se trouvent à l'intérieur d'un même cluster sont co-localisés. Connaissant la localisation de quelques hôtes qui s'y trouvent, grâce à une base de données contenant des associations d'adresses IP et leurs localisations, elle déduit la localisation du cluster en entier. Son efficacité dépend de la véracité des informations se trouvant dans la base de données utilisée. Ces informations étant fournies par les utilisateurs sont peu fiables.

La troisième technique, GeoPing, est la plus proche de LGM, et exploite une possible corrélation entre délai et distance géographique. L'hypothèse de base de GeoPing est que des hôtes ayant un délai similaire par rapport à d'autres hôtes fixes (des serveurs sondes par exemple) tendent à être situés dans une même zone géographique. Ainsi, la localisation de l'hôte cible est assimilable à la position de l'hôte référence, qui a la mesure de délai la plus similaire à l'hôte dont on veut déterminer sa localisation. Le nombre d'endroits possibles, où on peut localiser l'hôte cible, est alors limité au nombre d'hôtes références, d'où un espace discret de réponses. Par conséquent, le nombre et le placement des hôtes références jouent un rôle important dans la précision de l'estimation de localisation [ZFdRD04]. L'amélioration de la technique GeoPing passe par une augmentation du nombre d'hôtes références. Dans la section 4 nous comparons LGM à l'approche DNS et à la technique GeoPing.

2.2 Contributions

LGM est la première technique dans le domaine de la localisation à utiliser la multilatération pour inférer la position d'un hôte. Ses principales contributions sont :

- LGM établit une relation dynamique entre les adresses IP et leur localisation géographique grâce à des mesures de délai faites périodiquement entre les hôtes références.
- L'apport principal de la technique LGM est sa capacité à transformer les mesures de délai en distances géographiques surestimées, utilisées par la multilatération. Ainsi, en utilisant la multilatération, nous obtenons un espace continu d'endroits où on peut localiser un hôte contrairement aux autres techniques de localisation basées sur des mesures de délai.
- LGM fournit également une zone de confiance pour chaque hôte localisé offrant aux applications qui l'utilisent la possibilité d'évaluer la précision de l'estimation par rapport à leurs exigences.

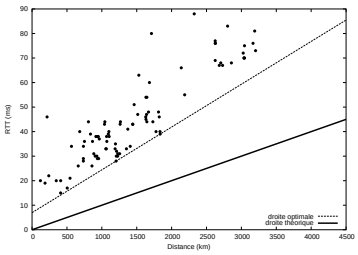


FIG. 1: Exemple montrant la relation entre distance géographique et délai.

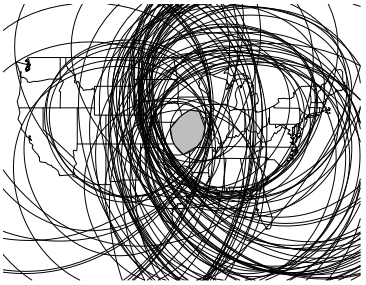


FIG. 2: Exemple de zone d'estimation de la localisation d'un hôte.

3 La Multilatération : idée générale

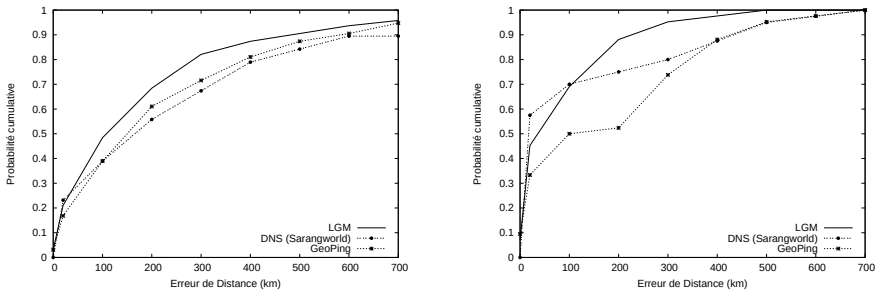
Soit un ensemble $\mathcal{H} = \{H_1, H_2, \ldots, H_K\}$ de K hôtes références. Connaissant les délais entre un hôte cible et ces hôtes références, notre principal objectif est d'estimer les distances géographiques correspondantes. En effet, le délai de bout en bout est la somme des délai de propagation et de transmission, et des délais d'attente dans les files des routeurs.Au délai effectif mesuré s'ajoute donc un délai supplémentaire induit par ces distorsions. Ainsi, l'estimation de distance fournie par LGM est appelée par définition distance géographique surestimée, car étant la somme de la distance géographique réelle et de la distance induite par les distorsions.

La Figure 1 illustre un exemple choisi parmi les résultats décrits dans la section 4. L'axe des abscisses représente la distance géographique réelle et l'axe des ordonnées le délai mesuré entre un hôte référence H_i et les autres hôtes références restants. Supposons qu'il existe un chemin linéaire entre l'hôte référence H_i vers les autres hôtes références restants et que les données sont contraintes à aucun autre facteur à part le délai de propagation. On devrait avoir une droite de la forme $y = mx + b$ où $b = 0$ puisqu'il n'y a pas de délai additionnel et m n'est rien d'autre que la vitesse de transmission des données dans le support physique. La "droite théorique" illustrée dans la Figure 1 montre ce cas. Cependant dans la réalité ce chemin linéaire existe rarement à cause des politiques de routage et des goulots d'étranglement. Ainsi pour modéliser la relation entre délai et distance géographique, nous définissons une "droite optimale" pour chaque hôte référence H_i comme la droite $y = m_i x + b_i$ qui est la plus proche, mais en dessous de tous les points (x,y) et dont l'ordonnée à l'origine *i*.e. b_i n'est pas négative. Ainsi [GZCF04] montre comment les droites optimale et théorique sont construites. Chaque hôte référence utilise sa propre droite optimale pour convertir le délai obtenu, entre l'hôte cible et lui, en distance géographique surestimée.

4 Resultats

Pour nos expériences, nous avons utilisé des hôtes de RIPE [RIP] localisés en Europe Occidentale et de NLANR [NLA] localisés aux États Unis. Chaque ensemble contient respectivement 42 et 95 hôtes. Nous construisons la matrice de délai de chaque ensemble qui contient le RTT minimum entre les hôtes. Les hôtes références de chaque ensemble jouent à tour de rôle l'hôte cible à localiser et les hôtes références restants tentent de le localiser.

La Figure 2 montre un exemple extrait à partir de nos résultats obtenus et illustre la méthodologie de LGM. Elle illustre l'ensemble des 94 cercles utilisés pour estimer la localisation d'un hôte AMP situé à Lawrence, en Kansas aux États Unis. La région grise illustrée dans dans la figure 2 représente la zone d'intersection $\mathcal{R}$ de ces cercles. Le polygone approximant cette région $\mathcal{R}$ (*voir* [GZCF04]) est la zone de confiance que LGM associe à chaque estimation de localisation et son centre le point d'estimation de l'hôte cible. Après avoir trouvé la position d'estimation de chaque hôte cible, nous avons calculé l'erreur de distance qui représente la différence entre la position estimée et la position réelle de l'hôte cible τ. Nous avons comparé nos résultats avec ceux obtenus par une méthode basée sur les noms DNS (*voir* le projet SarangWorld Traceroute [Sar]) et par GeoPing qui utilise un espace discret de réponses [PS01]. La Figure 3 montre la fonction de probabilité cumulative de l'erreur de distance obtenue en utilisant LGM, la



(a) Hôtes situés aux États Unis.

(b) Hôte situés en Europe de l'Ouest.

FIG. 3: Erreur de distance de LGM, de la méthode DNS et de GeoPing.

méthode basée sur le DNS et GeoPing. LGM dépasse en précision et la méthode basée sur le DNS et la technique GeoPing. Ainsi, l'erreur médiane de distance obtenue pour les hôtes références localisés aux E.U est inférieure à 100 km tandis que pour l'Europe Occidentale elle est inférieure à 25 km. Alors que pour la technique GeoPing cette erreur est de 150 km pour les E.U et 100 km pour l'Europe Occidentale.

5 Conclusion

Cet article montre une comparaison des différentes techniques de localisation. Les résultats obtenus illustrent que LGM est plus performante que les précédentes techniques de localisation géographique. LGM montre que la transformation des mesures de délai en distance géographique surestimée est possible. Transformer les mesures de délai en distance géographique surestimée avec précision est un challenge en raison de beaucoup de particularités inhérentes à l'utilisation d'Internet. La localisation à partir ou vers des hôtes situés un peu partout dans l'Internet, par exemple PlanetLab [pla], est envisagé pour nos travaux à long terme. De même que l'utilisation d'une base de données, où on enregistre les couples IP-Localisation des hôtes déjà localisés, afin d'éviter des mesures répétitives.

Références

[DVGD96] Christopher Davis, Paul Vixie, Tim Goowin, and Ian Dickinson. A means for expressing location information in the domain name system. *Internet RFC 1876*, January 1996.

[EM99] Per Enge and Pratap Misra. Special issue on global positioning system. *Proceedings of the IEEE*, 87(1) :3–15, January 1999.

[GZCF04] Bamba Gueye, Artur Ziviani, Mark Crovella, and Serge Fdida. Constraint-based geolocation of internet hosts. In *Proc. of the ACM Sigcomm Internet Measurement Conference - IMC'2004*, Taormina, Sicily, Italy, October 2004.

[NLA] NLANR Active Measurement Project. `http://watt.nlanr.net/`.

[pla] *PlanetLab*. `http://www.planet-lab.org`.

[PS01] Venkata N. Padmanabhan and Lakshminarayanan Subramanian. An investigation of geographic mapping techniques for Internet hosts. In *Proc. of the ACM SIGCOMM'2001*, San Diego, CA, USA, August 2001.

[RIP] RIPE Test Traffic Measurements. `http://www.ripe.net/ttm/`.

[Sar] Sarangworld Traceroute Project. `http://www.sarangworld.com/TRACEROUTE/`.

[ZFdRD04] Artur Ziviani, Serge Fdida, José Ferreira de Rezende, and Otto Carlos Muniz Bandeira Duarte. Improving the accuracy of measurement-based geographic location of Internet hosts. *Computer Networks, Elsevier Science*, 2004. Accepted for publication.