Algorithme distribué de filtrage et de routage orienté contenu

Yosra Barouni, Prométhée Spathis, Serge Fdida

Algorithme distribué de filtrage et de routage orienté contenu

Yosra Barouni Prométhée Spathis Serge Fdida

Université Pierre et Marie Curie, Paris Universitas
104 Avenue du président kennedy
Paris, 75016, France

Dans cet article, nous présentons un modèle générique conçu pour capturer les propriétés distinctives des diverses approches existantes en matière de distribution de contenu. Nous utilisons la terminologie introduite dans notre modèle pour proposer un algorithme de filtrage et de distribution de contenu. Notre algorithme procède par réécriture des requêtes de client dans un format simple et générique. Une fois réécrites, les requêtes sont dirigées vers les contenus les plus pertinents. Pour ce faire, notre algorithme utilise un processus de prédiction basé sur l'optimisation d'un vecteur regroupant trois métriques de filtrage de contenu.

Keywords: réseaux de contenu, routage, filtrage, optimisation.

1 Introduction

Le succès d'Internet est largement dû à la quantité de contenus mis à disposition de tous. Les mesures de trafic ont toutes montré que l'accès à ces contenus est le service qui domine les usages de l'Internet. Avec son modèle orienté hôte, l'Internet n'a pourtant pas été architecturé pour permettre la livraison de contenu à son échelle. L'infrastructure logicielle avec les protocoles de réseau en vigueur ou matérielle tels que les équipements en place ne sont pas conçues pour répondre aux besoins des services ayant trait au contenu. Pour être rendus, les services actuellement en vigueur requièrent qu'une connexion soit préalablement établie sur la base de la localisation des entités impliquées.

Les réseaux orientés contenu font référence à la capacité donnée aux applications et à leurs utilisateurs d'accéder aux données indépendamment de leur localisation [PGM+06]. Ces réseaux ont fait l'objet d'un intérêt soutenu à l'origine d'un nombre important de propositions dans des domaines aussi différents que le multicast, les systèmes pair-à-pair [LCP+04], les réseaux de distribution de contenu [JEVO07] ou les systèmes de publication/abonnement [LP03]. Ces approches diffèrent selon les applications pour lesquelles elles ont été conçues, ainsi que selon les types de contenus visés voire en fonction de la topologie ou la taille du réseau.

Au vue de ces différences, il n'existe pas de solution qui s'accommode de la complexité résultant de l'interaction entre systèmes hétérogènes tout en permettant de tirer profit des avantages propres à chacune des approches proposées à ce jour. Il est donc nécessaire de dégager une architecture globale qui prend en considération l'hétérogénéité des infrastructures de gestion de contenu à la fois au niveau du routage et du filtrage de contenu ainsi que dans la distribution de contenu aux utilisateurs. Afin d'aboutir à une telle architecture, nous avons procédé à la conception d'un modèle générique de routage qui regroupe les éléments essentiels au processus de routage par contenu. Ce modèle a été conçu de façon à capturer les propriétés distinctives des différents schémas de routage par contenu existants tout en s'adaptant au contexte d'applications variées.

Modèle Générique Une architecture de réseau orienté contenu peut être vue comme étant constituée de quatre types d'entités : les fournisseurs d'origine, les clients, les routeurs de médiation et les serveurs de contenu. Chacune de ces quatre entités se différencient par le rôle qui leur est attribué dans le processus de routage orienté contenu. Notons qu'un même noeud du réseau peut cumuler plusieurs de ces rôles. Par

exemple, dans un réseau pair à pair, un nœud peut à la fois assumer le rôle de fournisseur, de client et de routeur de médiation. Un noeud qui assume le seul rôle de client est totalement passif dans le processus de recherche et de filtrage de contenu.

Le service rendu par un réseau orienté contenu nécessite la coopération entre deux types d'infrastructures : une première pour la gestion des opérations liées au stockage et à la réplication des contenus, opérations prises en charge par les serveurs de contenu. La présence de tels serveurs vise la minimisation du temps d'accès aux données et la diminution de la charge de traitement des fournisseurs d'origine en cas de demandes simultanées multiples. Une seconde infrastructure concerne la gestion du routage des contenus entre fournisseurs d'origine ou serveurs de contenu et les clients. Les fonctions de routage sont réalisées au niveau des routeurs de médiation en assurant la dissémination des requêtes de client et celle des annonces de fournisseur. Pour chacun de ces messages, les routeurs de médiation maintiennent des états relatifs aux informations véhiculées par ces messages.

Plan de l'article Après avoir établi un modèle générique englobant l'ensemble des propriétés propres aux architectures de distribution de contenus qui nous sont apparus significatives, nous nous basons sur la terminologie introduite dans ce modèle pour décrire un algorithme de routage orienté contenu basé sur le filtrage des contenus distribués en réponse aux requêtes de client. L'algorithme que nous proposons repose sur un vecteur de métriques pertinentes du point de vue du filtrage et sur un processus d'optimisation linéaire du filtrage des contenus retournés. La suite de cet article s'organise de la façon suivante : dans la section 2, nous présentons notre contribution en détaillant notre algorithme de filtrage et d'optimisation du vecteur de filtrage. La dernière section conclut par une analyse de notre travail et présente les objectifs de la suite de nos travaux.

2 Filtrage orienté contenu

Dans cette section, nous proposons un schéma de routage en rupture avec les architectures de routage traditionnelles puisque centré exclusivement sur la recherche des contenus par leur description. Notre approche surpasse ces architectures conçues sur le modèle orienté hôte qui repose sur l'identification et la localisation des nœuds par l'utilisation respective de noms et d'adresses uniques. Pour s'abstraire d'un système de nommage ou d'adressage, nous proposons un routage consistant à filtrer de proche-en-proche les requêtes de client. Ce filtrage est réalisé à deux niveaux. Les routeurs de médiation sont répartis dans des domaines interconnectés par des routeurs spécifiques que nous appellerons en raison de leur situation, des routeurs d'interconnexion. Au sein d'un même domaine, la dissémination des informations décrivant les contenus est à la charge des routeurs de médiation et se fait sur une base périodique et de manière proactive. Les routeurs de médiation situés à la bordure d'un domaine ont également la charge de stocker les contenus disponibles au sein de ce domaine. Cette charge supplémentaire permet aux routeurs de bordure de ménager les capacités de stockage des routeurs de coeur tout en les consacrant exclusivement aux processus de filtrage et de routage. Les routeurs d'interconnexion quant à eux s'échangent les descriptions de contenu qu'ils se chargent d'agréger. Cette structuration hiérarchique des routeurs permet à notre approche de passer le facteur d'échelle.

2.1 Modèle formel

Avant de présenter notre algorithme de filtrage, nous introduisons le format des structures de données suivantes : les requêtes de client, les annonces de fournisseur où sont décrits les contenus mis à la disposition des clients et les tables de filtrage que calculent les routeurs de médiation sur réception des annonces de fournisseur. C'est sur la base de ces tables que les routeurs de médiation acheminent les requêtes de client dans le réseau selon le contenu recherché.

Un routeur de médiation i situé à la bordure d'un domaine maintient une copie de $C_i = \{c_{i,1}, c_{i,2}, ...\}$, l'ensemble des contenus populaires du point vue des clients qu'il connecte au domaine[†]. A cet ensemble C_i correspond $\mathcal{D}_i = \{d_{i,1}, d_{i,2}, ...\}$ l'ensemble des descriptions de chacun des contenus de C_i. Chaque élément d_i se compose d'un ensemble fini de mots-clés k_j. Un routeur de médiation j situé au cœur du réseau se

[†] Le mécanisme de sélection et de réplication des contenus selon leur popularité dépasse le cadre cet article

contente quant à lui de maintenir l'ensemble $\mathcal{D}_j$ pour chacun des contenus qu'il sait joindre. Une requête issue du client c notée RQ_c est un message contenant un identifiant unique et un ensemble de mots-clés décrivant l'intérêt du client c pour les contenus qu'il recherche : $RQ_c = \{k_1, k_2, ...\}$. Une annonce du fournisseur f notée AD_f contient la description $\mathcal{D}_f$ des contenus que f héberge.

2.2 Algorithme de filtrage

Le format des requêtes de client pouvant varier selon les systèmes de distribution de contenu, les requêtes sont transformées à leur entrée dans l'infrastructure de routage en une structure générique basée sur un ensemble de mots-clés. Cette structure permet de garder une relation de correspondance simple et suffisamment expressive entre la requête initiale et celle résultant de cette transformation. L'utilisation d'un format générique permet de garantir la compatibilité entre les différentes applications à partir desquelles les clients lancent leurs recherches de contenu.

Le processus de filtrage passe par le calcul d'un vecteur de filtrage noté (D,P,M) dont les coordonnées correspondent chacune à une métrique :

- D est la distance exprimée en nombre de sauts qui sépare le routeur des contenus potentiels répondant aux requêtes reçues
- P mesure la popularité du contenu exprimé en nombre d'accès total pour des requêtes similaires
- M représente le taux de satisfaction des requêtes.

Le calcul de la première métrique utilise les tables de routage classiques des routeurs. Le calcul de la popularité d'un contenu utilise les statistiques d'accès au contenu obtenues à partir de l'historique maintenu pour ce contenu. Enfin, nous nous basons sur un système de gestion de la *réputation* des contenus pour calculer les valeurs de la troisième métrique. Dans ce système de réputation, les clients retournent des informations concernant les contenus reçus en réponse à leurs requêtes. Le système se charge alors d'appliquer un processus d'apprentissage par historique tout au long de la vie d'un contenu.

Les valeurs des trois métriques que nous venons de présenter sont maintenues dans la table de filtrage des routeurs de médiation. A chaque entrée de cette table est associé l'identifiant de l'interface de sortie qui permet de joindre le contenu décrit par l'entrée correspondante. Le filtrage revient donc à rechercher l'entrée qui concorde avec les mots-clés k_i contenus dans la requête RQ_c d'un client c. Une fois l'entrée adéquate identifiée, la requête est alors relayée sur l'interface de sortie associée à cette entrée. Les routeurs de médiation calculent leur table de routage en exécutant un algorithme d'optimisation du vecteur (D,P,M) que nous présentons dans la section suivante. Afin de maintenir la taille de ces tables efficace, nous avons recours à l'agrégation de leurs entrées par similarité orthographique et sémantique. Un mécanisme similaire permet aux routeurs de traiter de manière groupée les requêtes identifiées comme similaires afin de minimiser les capacités de traitement mobilisées et le nombre des messages échangés à travers le réseau.

L'algorithme de filtrage est exécuté de proche en proche. Cette propriété permet de prendre en compte les changements de topologie ou les défaillances des noeuds du réseau, évitant ainsi le calcul de nouvelles routes. Cette propriété s'avère également utile dans le cas des réseaux large échelle pour lesquels les temps nécessaires à la propagation des mises à jour concernant la disponibilité des contenus peuvent être longs. Il est donc important de procéder au filtrage des requêtes au fur et à mesure de leur avancement dans l'infrastructure de routage.

L'algorithme 1 résume le processus de filtrage simplifié qu'exécute un routeur de médiation *i* à la réception d'une requête RQ_c (voir section 2.1).

2.3 Optimisation du filtrage

Comme présenté dans la section précédente, à chaque entrée d'une table de filtrage correspond un ensemble de mots-clés décrivant les contenus qu'il est possible d'atteindre en passant par l'interface de sortie associée à cette entrée. Pour chaque entrée, les routeurs de médiation maintiennent également dans un vecteur, la valeur des trois métriques : D, P et M. Déterminer l'entrée la plus pertinente du point de vue d'une requête consiste dans un premier temps à parcourir la table de filtrage en confrontant les mots-clés de la requête à ceux des entrées scrutées. Une fois la table parcourue, le routeur sélectionne parmi toutes celles retournées dans un premier temps, l'entrée la plus pertinente du point de vue de la requête en cours de

Algorithme 1 Algorithme de filtrage saut par saut au sein d'un router de médiation i

1: $d_{matching} \leftarrow null$
2: **for all** $d_{i,k} \in \mathcal{D}_i$ **where** k $\in [1, length(\mathcal{D}_i)]$ **do**
3: **if MatchEntryToRequest(** $d_{i,k}$ **,** RQ_c **)** ▷ Comparer les mots clés de l'entrée $d_{i,k}$ à ceux de RQ_c
4: **then** $d_{matching} \leftarrow d_{i,k}$ ▷ Ajouter l'entrée pertinente à la liste des résultats
5: **end for**
6: **MaximizeFilteringVector** $(d_{matching}, D, P, M)$ ▷ Entrée dont le vecteur (D,P,M) optimise la fonction (1)
7: $I_{output} \leftarrow$ **SearchOutputInterface**$(D_{opt}, P_{opt}, M_{opt})$ ▷ Interface de sortie associée au vecteur optimal
8: **ForwardRequest**(RQ_c, I_{output})

traitement. Pour ce faire, le routeur utilise le vecteur (D,P,M) pour optimiser la fonction coût suivante :

$$\text{Max}(w_D \cdot D + w_P \cdot P + w_M \cdot M) \wedge Min(D) \quad (1)$$

A chaque métrique est associé un poids que fait parvenir dans sa requête le client pour pondérer l'importance des métriques dans l'optimisation de la fonction (1) dont résultera le choix de la meilleure entrée. Ces poids sont notés w_D, w_P et w_M.

3 Conclusions et perspectives

Dans cet article, nous avons proposé un modèle générique de réseau orienté contenu qui capture les propriétés des différentes approches existantes. Sur la base de modèle, nous avons proposé un premier algorithme pour le filtrage des contenus que vient optimiser un second algorithme qui utilise un vecteur dédié aux métriques de filtrage. Notre solution présente l'avantage de passer le facteur d'échelle en tenant compte de la topologie dynamique du réseau. Notre solution présente également l'avantage d'être compatible quelque soit les applications de distribution de contenu. Cette indépendance est garantie par la transformation des requêtes de client dans un format générique. Le filtrage de ces requêtes est réalisé de proche en proche permet ainsi la prise en compte des changements de topologie ou des défaillances des noeuds du réseau. Enfin, notre approche garantit aux utilisateurs l'accès aux contenus les plus pertinents du point de vue de leurs requêtes.

Concernant les perspectives de nos travaux, nous suivons actuellement plusieurs pistes. Nous étudions des mécanismes basés sur les communautés intra ou inter-domaines pour gérer la réputation des contenus. Nous cherchons également à moduler la granularité des sous-requêtes résultant de la décomposition des requêtes. L'agrégation des résultats retournés est aussi un processus important que nous étudions dans le but d'accélérer la recherche des contenus et de minimiser le nombre de messages échangés entre routeurs de médiation.

Références

[JEVO07] B. Jabari, V. Englebert, J.P. Vigneron, and E.M. Oualim. Towards an agent-based grid platform 1st episode : the meta-grid. *UPGRADE '07 : Proceedings of the second workshop on Use of P2P, GRID and agents for the development of content networks*, pages 73–80, 2007.

[LCP+04] E.K. Lua, J. Crowcroft, M. Pias, R. Sharma, and S. Lim. A survey and comparison of peer-to-peer overlay network schemes. *Communications Surveys and Tutorials, IEEE*, 7(2) :72–93, 2004.

[LP03] Y. Liu and B. Plale. Survey of publish subscribe event systems. *Technical Report TR574, Indiana University*, 2003.

[PGM+06] T. Plagemann, V. Goebel, A. Mauthe, L. Mathy, T. Turletti, and G. Urvoy-Keller. From content distribution networks to content networks - issues and challenges. *Computer Communications*, pages 551–562, 2006.