Proposition de recherche doctorale

Système de dissémination de Flux RSS

Mots clés :

- **Directeur de thèse** : Mesaac Makpangou
- **Co-encadrant(s)** :
- **Unité de recherche** : Laboratoire d'informatique de Paris 6
- **Ecole doctorale** : École Doctorale Informatique, Télécommunications, Électronique de Paris
- **Domaine scientifique principal**: Divers

Résumé du projet de recherche (Langue 1)

Avec l'essor du paradigme web 2.0, Internet est devenu un vaste espace de publication de contenu dynamique. Les applications web 2.0 comme les blogs, les réseaux sociaux et les mashups permettent aux utilisateurs qui étaient des consommateurs passifs, de générer facilement du contenu web. En conséquence, la quantité d'information disponible sur le web est devenue importante mais aussi le délai entre la mise en ligne d'un contenu et sa découverte par les utilisateurs a augmenté. Pour réduire ce délai de découverte et faciliter la collecte et la mise en ligne de l'information, les fournisseurs mettent à disposition de leurs clients des flux RSS. RSS (Really Simple Syndication) désigne un format XML utilisés pour la syndication de contenu web. Un flux RSS est document composé d'une suite d'éléments d'information appelés "item". L'item est composé essentiellement de trois parties: le titre de l'information, le résumé de l'information et un lien qui mène vers l'article complet. A chaque apparition d'une nouvelle information, le fournisseur ajoute un item sur le flux. Le flux RSS est donc un conteneur d'items et possède lui même un titre, une description et un lien qui constitue l'URL du flux. Les consommateurs souscrivent sur un flux à travers cet URL et téléchargent régulièrement le contenu du flux disponible sur le fournisseur. Une fois que le flux est téléchargé, le consommateur effectue le trie pour récupérer les informations pertinentes. Les consommateurs abonnés sur plusieurs flux, reçoivent les informations provenant de tous ces flux, par conséquent, Il deviennent rapidement submergés par les données provenant de ces flux. Ces consommateurs sont alors confrontés au problème consistant à trier la masse importante de données reçues. En effet, ce sont les fournisseurs qui choisissent le contenu à publier sur les flux ainsi les informations diffusées auront des degrés d'importance différents selon les utilisateurs. Les outils d'exploitation des flux RSS (lecteurs RSS et aggrégateurs) n'offrent pas la possibilité de filtrer le contenu des flux. Les solutions existantes dans l'état de l'art proposent des mécanismes de filtrages sémantiques ou algébriques (fondés sur des algèbres de requêtes RSS). Ces solution contraignent l'utilisateur à connaître les sources de flux pour pouvoir y appliquer des filtres. Cette approche limite l'exploitation de RSS. De plus les fournisseurs de contenu célèbres voient l'accroissement continue du nombre d'abonnés sur leur flux. Ces fournisseurs restent ainsi confronter au problème consistant à réduire la bande passante consommée par le téléchargement des clients et la surcharge générée par les demandes de téléchargement. En effet, dans l'état de l'art, des auteurs affirment que le site Boing Boing qui dispose de 11500 abonnés sur ses flux RSS et Atom doit faire face à une consommation de bande passante de 22 Go par jour pour seulement son trafic de flux; pour le site BBC News la demande en bande passante est de 989 Go. Le problème s'accentue avec les sites ayant un plus grand nombre d'abonnés. Par exemple TechCrunch le plus célèbre des blogs américains traitant le sujet "technology" possède 1.750.000 abonnés à ses flux. Sa fréquence de mise à jour étant d'une fois par heure, si en moyenne une mise à jour a une taille de 50 Ko, il se retrouve confronter quotidiennement à une consommation de 2.100 Go en bande passante. Ce problème constitue un frein pour le développement des flux RSS. Des travaux existants proposent une collaboration des serveurs via un réseau P2P pour la dissémination des flux RSS vers les clients. Ces solutions sont limitées par le manque de filtrage des informations contenues dans les flux publiés. Ces solutions doivent être revues et améliorer pour fournir à l'utilisateur les informations pertinentes. Ceci réduira davantage la bande passante utilisée. Par ailleurs, la simplicité de RSS permet d'envisager de nouvelles utilisations des flux. Ainsi RSS peut être utilisé pour faire de la veille informationnelle sur le web. Cette dernière consiste à suivre l'actualité sur un sujet donné (élections présidentielles, réchauffement climatique, crise financière, etc.). Pour ce faire les veilleurs doivent s'abonner sur les flux RSS susceptibles de traiter les thèmes recherchés. A ce niveau deux problèmes majeurs se présentent. Le premier problème consiste à trouver les fournisseurs appropriés pour un sujet donné et le deuxième problème est d'être averti de l'existence de nouveaux fournisseurs de contenu sur le sujet en question. En effet le premier problème découle du fait de la multitude de sources existant sur le web, il est donc impossible de retrouver toutes ces sources. Le deuxième problème vient du fait que le web est devenu très dynamique donc le contenu publié change fréquemment. Les moteurs de recherches permettent à ces veilleurs de trouver les sites pertinents pour un sujet donné par le biais des méta-données décrivant de manière statique le contenu des pages. Mais comme énoncé tantôt, le contenu des sites changent régulièrement, de ce fait la description issue des méta-données peut devenir très large ou même diverger complètement du contenu publié. De plus les algorithmes de classement des pages web utilisés par les moteurs de recherche ne sont pas avantageux pour les flux RSS qui sont loin d'être célèbres. Il faut alors concevoir un support capable de trouver les flux pertinents mais aussi de découvrir les informations correspondant à un sujet de veille. Enfin, pour certains consommateurs, il est impossible de suivre en temps réel les mises à jour survenues au niveau des flux. Ceci peut être dû à des connexions intermittentes dans certaine zones isolées. Ces consommateurs souhaitent alors disposer des informations produites pendant leur absence. Malheureusement les solutions existantes ne tiennent pas compte de cet aspect, elles se focalisent plutôt sur la distribution des mises à jour récentes. Ainsi un mécanisme de surveillance, de récupération et de stockage de flux pour une utilisation future est nécessaire pour ces types de consommateur. {{Objectif}} L'objectif fixé pour la thèse consiste à élaborer un système de recherche et de découverte d'information dans les flux capable de mettre en relation les fournisseurs et les consommateurs de flux RSS. Le système à proposer devra permettre l'accès aux informations contenues dans les flux à travers des requêtes personnalisées.