Proposition de recherche doctorale

An Intelligent Publish/Subscribe System in a BigData environment

Mots clés :

- **Directeur de thèse** : Cédric DU MOUZA
- **Co-encadrant(s)** :
- **Unité de recherche** : Centre d'Étude et de Recherche en Informatique et Communications
- **Ecole doctorale** : École Doctorale Informatique, Télécommunications, Électronique de Paris
- **Domaine scientifique principal**: Divers

Résumé du projet de recherche (Langue 1)

{{Lieu:}} laboratoire CEDRIC du CNAM {{Equipes:}} ISID & Vertigo {{Encadrants:}} C. du Mouza, N. Travers ({dumouza, nicolas.travers}@cnam.fr) {{Mots-clés:}} Publish/Subscribe, indexation, passage à l'échelle, continu, diversité, Big Data {{{Contexte}}} Internet est aujourd'hui un support économique reconnu et utilisé pour la diffusion d'informations à large échelle. Afin de réduire l'intervalle de temps nécessaire entre la publication de l'information sur un site web ou un réseau social et sa consultation par les utilisateurs, les systèmes de notifications en continu ont pris une ampleur considérable sur la toile. L'approche {{« publication/souscription »}} (Pub/Sub) [3] pour la diffusion contrôlée et efficace d'informations sur le web est devenue la référence du domaine de notification en continue. Les fournisseurs d'information diffusent l'apparition de nouvelles informations (par exemple un article dans un journal électronique) à travers entre autres des flux (feeds) RSS [1] ou ATOM [2] auxquels les clients intéressés peuvent s'abonner grâce à des portails web ou des logiciels (lecteurs RSS/ATOM) spécialisés. Ce processus permet au final à chaque utilisateur de surveiller « en temps réel » l'évolution d'informations publiées sur le Web. Naturellement, le nombre de ces sources de données grandit chaque jour et le nombre d'utilisateurs explose (e.g., Twitter suit une croissance exponentielle). De fait, le passage à l'échelle des systèmes de notification de type « publication/souscription » est un défi réel aussi bien au niveau qualitatif que quantitatif. En effet, il faut traiter un {{énorme volume de données en continue}}, tout en étant capable de délivrer des {{informations pertinentes}} à un nombre d'utilisateurs toujours plus grand, sans les submerger d'informations hors sujet ou redondantes. Les portails d'agrégation spécialisés comme Blastfeed.com, Plazoo.com et Technorati.com sont de plus en plus confrontés à des problèmes de passage à l'échelle et proposent uniquement des possibilités de filtrages rudimentaires pour l'utilisateur. Nous avons réalisé une étude préliminaire approfondie du comportement des flux RSS, dans laquelle nous nous sommes particulièrement intéressé à leur comportement et à la structure et contenu des items publiés [4]. En nous appuyant sur cette étude, nous avons proposé des {{structures d'indexation de souscriptions}} (requêtes utilisateurs) basées sur des mots-clés et adaptées à la {{notification de messages en continu}} [5]. Toutefois, malgré cette étape de filtrage par mots-clés qui permet de réduire le nombre de messages notifiés aux utilisateurs, la masse d'information reste phénoménale. De fait, l'utilisateur reste submergé par des {{informations redondantes ou largement similaires}}. Ainsi, nous souhaitons étudier une nouvelle approche de filtrage intelligent, complémentaire des travaux précédents, reposant sur la {{diversité et la nouveauté}} des résultats sur des données produites en continu [6,7,8]. L'idée est que les messages déjà notifiés (l'historique de la souscription) à un utilisateur peuvent servir de filtre pour les messages futurs. Les notions de nouveauté (information non encore notifiée dans l'historique) et de diversité (information globalement différente de l'historique) sont les concepts clés sur lesquels repose ce sujet de thèse. {{{Sujet de thèse:}}} Un des deux objectifs de cette thèse est donc d'étudier la {{diversité et la nouveauté}} de l'information sur des données en continue {{dans un cadre Publication/Souscription}}. Il s'agit d'un sujet précurseur dans ce contexte, très porteur, et qui rencontre un succès croissant dans la communauté base de données dans d'autres contextes. Le grand volume de données et de souscriptions fait de ce sujet un réel défi d'importance. Une première approche envisagée est de mutualiser les historiques de toutes les souscriptions pour permettre un passage à l'échelle évident. Le calcul de ces fonctions de similarité, nouveauté, diversité, restent malheureusement couteux et des techniques d'optimisation de calcul, des regroupements d'historique ou de souscription, et des techniques de réduction de l'espace seront à prospecter tout au long de la thèse. De ce point de vue, nous nous focaliserons particulièrement sur une approche de pré-filtrage, inspiré de méthodes basées sur des seuils de contrôle [9] réduisant positivement l'espace de recherche. Le dernier point envisagé dans cette thèse est l'étude du {{passage à l'échelle}} de notre solution {{dans le cadre du BigData Management}}. En effet, bien que nos précédents résultats aient été étudiés au niveau local, une distribution des données et des calculs sont indispensables pour un réel passage à l'échelle au niveau du Web. Un des objectifs de cette thèse consistera à proposer une distribution des structures d'indexation et de calcul de similarité. Une étude apronfondie des systèmes de BigData tel que, Memcached [10], Hadoop [11,12,13], MapReduce [14,15], Hbase [16,17], sera nécessaire pour trouver la solution adéquate. Une attention particulière sera donnée à la gestion de données en mémoire et en continue. C'est pourquoi nous nous intéressons plus particulièrement à Pig (Hadoop en flots de données) et Memcached (key-value store in full-memory).

Résumé du projet de recherche (Langue 2)

{{{Objectifs de la thèse:}}} Le défi ici est multiple : -* Contraintes de temps : il faut en un temps très court satisfaire la demande d'un grand nombre d'utilisateurs recherchant des informations. On cherchera donc à filtrer les items en évitant de parcourir toutes les souscriptions susceptibles d'être notifiées. Le filtrage intelligent se base donc sur un pré-filtrage basé sur la diversité et les historiques des souscriptions concernés ; -* Evolution continue des flux : dans un contexte de flux continus, le contenu notifié fait évoluer l'historique au cours du temps et change, de fait, la pertinence de la diversité et de la nouveauté. Une mesure évolutive de la pertinence devra être prise en compte ; -* Contexte unique : la combinaison du matching exact de souscriptions dans le contexte des flux [5], combiné au filtrage par diversité et nouveauté pour donner de la qualité au résultat est inédit dans le domaine Pub/Sub. Une analyse fine de cette combinaison est nécessaire pour améliorer la pertinence du résultat de ce filtrage ; -* Passage à l'échelle : La distribution des structures d'indexation et de filtrage intelligent dans le contexte du BigData est une étape indispensable pour gérer des données en continues provenant du Web. Des systèmes de passage à l'échelle seront donc étudiés.

Informations complémentaires (Langue 2)

{{{Equipe d'accueil:}}} Cette thèse s'inscrit dans la continuité d'une collaboration solide entre les équipes ISID et Vertigo du laboratoire CEDRIC (encadrement par C. du Mouza et N. Travers). Ce travail permettra également de conforter les nombreuses collaborations dans ce domaine avec le laboratoire FORTH de Crètes (Grèce) avec Vassilis Christophides. {{{Références:}}}

[1] RSS. Really Simple Syndication. http://www.rssboard.org/rss-specification [2] Atom. W3C. Atomenabled. http://www.atomenabled.org/. [3] "Distributed Event-Based Systems", 2007 by Gero Muehl, Ludger Fiege, and Peter R. Pietzuch, Springer-Verlag, Germany. [4] Z. Hmedeh, N. Travers, N. Vouzoukidou, V. Christophides, C. du Mouza, M. Scholl - Characterizing Web Syndication Behavior and Content, WISE'11, The 12th International Conference on Web Information System Engineering, October 2011, pp.29--42, Series LNCS, Sidney, Australia [5] Z. Hmedeh, H. Kourdounakis, V. Christophides, C. du Mouza, M. Scholl, N. Travers - Subscription Indexes for Web Syndication Systems , International Conference on Extending Database Technology (EDBT'12), March 2012, pp.311-322, Berlin, Germany [6] M. Drosou, E. Pitoura – Dynamic Diversification of Continuous Data. EDBT 2012, Berlin, Germany [7] Marcos R. Vieira, Humberto L. Razente, Maria C. N. Barioni, Marios Hadjieleftheriou, Divesh Srivastava, Caetano Traina Jr., Vassilis J. Tsotras – On Query Result Diversification. ICDE 2011 [8] A. Angel, N. Koudas – Efficient Diversity-Aware Search. SIGMOD 2011, Athens, Greece [9] N. Vouzoukidou, B. Amann, V. Christophides - Processing continuous text queries featuring non-homogeneous scoring functions. CIKM 2012: 1065-1074 [10] Brad Fitzpatrick. 2004. Distributed caching with memcached. Linux J. 2004, 124 (August 2004) [11] T. White. 2009. Hadoop: The Definitive Guide. O'Reilly Media, Inc. June 2009. [12] Apache Hadoop. http://hadoop.apache.org/. [13] Pig. Hadoop data-flow platform. http://pig.apache.org/ [14] Ashish Thusoo, Joydeep Sen Sarma, Namit Jain, Zheng Shao, Prasad Chakka, Suresh Anthony, Hao Liu, Pete Wyckoff, and Raghotham Murthy. 2009. Hive: a warehousing solution over a map-reduce framework. Proc. VLDB Endow. 2, 2 (August 2009), 1626-1629 [15] MongoDB. http://www.mongodb.org/ [16] Lars George - HBase: The Definitive Guide : Random Access to Your Planet-Size Data, O'Reilly Media, August 2011, Pages: 556 [17] Hbase. http://hbase.apache.org/