Entretiens avec des responsables de projets de moteurs Open Source

Retour au sommaire de la lettre

Pour en savoir plus sur les développements en cours dans ce domaine, nous avons interrogé deux responsables de projets Open Source européens. Wray Buntine (responsable d'un projet visant à créer un moteur sémantique Open Source sur une infrastructure peer to peer) et Dawid Weiss (chercheur polonais ayant contribué au développement d'un nouveau système de clustering) ont accepté de répondre à nos questions.

Wray Buntine – Directeur scientifique – Projet ALVIS (http://www.alvis.info)

Pouvez vous décrire le projet ALVIS ?

ALVIS est un projet de recherche européen visant à construire une infrastructure peer-to-peer dédiée à la recherche "sémantique"dans un environnement Open Source. Le consortium se compose de onze partenaires venant de six pays de l'Union Européenne ainsi que de la Chine et de la Suisse. Ces partenaires collaborent dans des domaines aussi variés et vastes que les topologies du réseau, les algorithmes de routage, les probabilités appliquées à la recherche de l'information, l'analyse linguistique ou encore la bioinformatique.
Les moteurs de recherche globaux nécessitent de disposer de moyens énormes. C'est pourquoi le projet a choisi de se fixer des objectifs plus modestes et auto-financés. Deux activités complémentaires sont poursuivies et intégrées : la construction d'outils ouverts pour fabriquer des moteurs de recherche de grande qualité spécialisés sur un sujet spécifique, et la mise en place d'une infrastructure compatible avec un réseau peer-to-peer. Les moteurs de recherche spécialisés seront dotés d'une certaine dose de reconnaissance sémantique basée sur une technologie d'extraction de l'information. Cela devrait satisfaire les utilisateurs que nous avons initialement ciblés, c'est-à-dire les organismes institutionnels, les groupes de recherche universitaires et les bibliothèques numériques. L'infrastructure peer-to-peer va rendre possible l'interopérabilité, le partage des ressources de recherche et la distribution des fonctionnalités de recherche. Nous souhaitons également nous appuyer sur des programmes en Open Source existants tels que Lucene, Terrier et Gate.

Qui est à l'origine du projet et quels sont les objectifs poursuivis ?
Le projet a été créé par le Professeur Henry Tirri et par le Docteur Wray Buntine en 2002. Il a pris un tour plus concret lorsque d'autres partenaires ont rejoint le projet, notamment Claire Nedellec de L'INRA, Adeline Nazarenko de l' Université Paris-Nord, Paris 13 et Exalead, la société parisienne spécialisée dans la recherche. La plupart d'entre nous sommes des fervents supporters de l'Open Source qui pensons que la recherche a besoin d'être menée de façon organisée et stratégique pour que l'Open Source devienne une réalité. De plus, nous sommes convaincus que cette recherche ne doit pas chercher à concurrencer des géants tels que Google et Yahoo ! Il faut commencer en ciblant une niche. Le bulletin de juin 2005 (volume 39) du forum SIGIR présente nos arguments en la matière, le business modèle que nous souhaiterions peut-être développer. Lorsque nous avons lancé le projet, nous pressentions que la recherche en Open Source allait faire beaucoup d'émules sur le web. Par exemple, Nutch est apparu à la même époque. Nous voulions que cette nouvelle technologie puisse s'appuyer sur un modèle réaliste capable de servir son développement.

De quels financements disposez vous ?
Le programme est co-financé par le 6ème programme-cadre de la Commission Européenne. Les 11 partenaires participent aux coûts liés au programme. Pour de plus amples informations, vous pouvez vous référer à la page web http://www.alvis.info. Vous y trouverez la liste des partenaires et des informations concernant le financement du projet.

Pouvez vous décrire les principales technologies qui ont été développées ?
Nous construisons une architecture pour que des développeurs de solutions Open Source puissent contribuer à créer des éléments d'un système. Nous ne pensons pas qu'un système de recherche monolithique ait un avenir. Il faut que divers groupes Open Source aient la

possibilité de proposer leur expertise et d'apporter au projet des composants en matière de crawling, d'indexation, de traitement des résultats ou d'extraction d'information. C'est pourquoi, nous avons besoin d'une architecture et de standards ouverts capables de fédérer tous ces éléments. Certes, il existe déjà des standards de qualité mais il en faut davantage. Les autres domaines où nous travaillons sont les suivants. Pour les moteurs de recherche spécialisés, nous assemblons et réutilisons des outils d'extraction de l'information, de hiérarchisation thématique et de ranking (basés sur le calcul des probabilités), ainsi que d'autres techniques qui devraient nous permettre de créer un "petit" système Web sémantique. Pour les systèmes peer–to-peer (P2P), le groupe EPFL [Ecole Polytechnique Fédéral de Lausanne] dirigé par Karl Aberer travaille sur une technologie capable d'adapter le peer-to-peer à nos moteurs de recherche thématiques.

Quel type de système de ranking des résultats avez-vous adopté ?
Nous testons le classement thématique (les utilisateurs peuvent spécifier des domaines thématiques) ainsi que de nouvelles méthodes statistiques qui se substituent au bon vieux Pagerank de Google. Cette partie de notre système est encore en développement.

Quelles sont les difficultés que vous rencontrez dans ce projet ?
Tout ce qui a été mentionné ci-dessus ! Nous pensons fermement être dans la bonne voie et des membres majeurs de la communauté se sont ralliés à notre point de vue. Toutefois, les obstacles en matière de conception et d'ingénierie qu'il nous faut surmonter sont nombreux. Nous voulons que notre premier système soit disponible au public dans quelques mois ; c'est là notre première difficulté à surmonter à court terme. A plus long terme, la grande difficulté sera sûrement de réunir une plus grande communauté autour de cette architecture.

Quel avenir voyez vous pour les technologies de recherche en Open Source ?
Selon nous, l'Open Source est inévitable. Son développement est lent mais régulier. Il faut que des organismes majeurs tels que W3C finissent par s'y intéresser et par apporter leur contribution pour que les meilleurs standards soient mis en place. Il faut que l'on permette à de nombreux groupes de recherche talentueux de se concentrer sur ce qu'ils font de mieux, que ce soit pour développer un système de récupération des données de grande qualité, ou un outil d'extraction et d'étiquetage [*tagging*] de l'information. Une plateforme de recherche en Open Source est une application dont rêvent de nombreux groupes de recherche en informatique mais au jour d'aujourd'hui, ils en sont toujours exclus. La recherche en Open Source ne sera faisable que si nous avons tous la possibilité d'apporter nos composants isolés en utilisant des standards et une architecture qui permette de les fédérer. Par ailleurs, l'Open Source ne sera réellement réalisable que lorsque des organismes tels que ACM, la Bibliothèque nationale de France ou Médecins Sans Frontières décideront que cela vaut la peine d'apporter en place leur propre "nœud actif" au réseau. Cependant, nous ne considérons pas que l'Open Source concurrence directement des géants comme Google ou Yahoo! puisque ceux-ci fournissent déjà un service gratuit de qualité.

Dawid Weiss – Chercheur à l'université de Poznan (Pologne) - (http://sourceforge.net/projects/carrot2/)

Pouvez vous décrire le projet "Carrot" ?
Le projet "Carrot", ou plutôt "Carrot2", a pour objectif la mise au point d' un cadre dans lequel diverses approches de text mining peuvent s'inscrire et qui soit réutilisable, efficace et pratique. Nos premiers objectifs sont les suivants : facilité d'utilisation, possibilité d'évolution rapide, efficacité et compatibilité avec plusieurs langages. Carrot2 est un cadre, cela signifie que nous essayons de faire en sorte que le projet soit générique et que les diverses composantes qui le constituent soient réutilisables. Ceci étant dit, notre principale application concerne en premier lieu un aspect précis du text mining : le clustering des résultats d'une recherche. Le cadre comporte quelques éléments qui permettent de regrouper contextuellement et automatiquement les résultats identiques d'une recherche. Cela ressemble beaucoup aux fonctionnalités de Vivisimo (ou de leur rejeton Clusty). Cependant, contrairement à Vivisimo, nous fournissons aussi une explication disponible en Open Source de l'algorithme et de la façon dont fonctionnent nos composants. Carrot 2 tire principalement sa réputation d'une composante spécifique de clustering : Lingo Classic. Nous sommes très fiers de ce produit et nous pensons qu'il est capable de concurrencer avec succès certains géants tels que Vivisimo et iBoogie.

Qui est à l'origine du projet et pourquoi l'a-t-il lancé ?
C'est en 2000 que le projet a été créé. J'ai écrit une thèse sur le clustering des résultats d'une recherche à l'université de Poznan en Pologne (University of Technology). Le projet, qui se nommait déjà Carrot, avait à l'origine pour but de vérifier certains concepts et certaines idées mises en œuvre dans le projet Grouper, un autre système de clustering développé par l'Université de Washington.
Lorsque Staszek Osinski et quelques autres ont rejoint le projet un an plus tard, l'architecture du système a été redessinée et le projet rebaptisé Carrot 2. Puis, un an après, nous avons publié le projet en Open Source sur SourceForge. Depuis lors, des gens du monde entier nous ont donné leurs codes et leurs composants. Staszek Ozinski et Dawid Weiss sont toujours responsables du projet et continuent, ensemble, de développer des algorithmes de clustering.

Etes vous soutenus financièrement?
Pas directement. Nous proposons aux entreprises des services de consulting et de personnalisation afin d'ajuster Carrot2 à leurs demandes spécifiques. Carrot2 est homologué sous les termes d'une licence BSD [une des licences du logiciel libre] et il peut donc être utilisé commercialement sans presque aucune restriction. Cependant, c'est un logiciel complexe qui nécessite fréquemment des ajustements et des réglages manuels. Nous faisons payer ces services. Nous compensons également financièrement notre choix de l'Open Source grâce à un produit dérivé que nous commercialisons et qui s 'appelle Carrot Search (http://www.carrot-search.com/). Carrot Search propose une version améliorée de Lingo qui est plus efficace et qui présente certaines fonctionnalités intéressantes qui sont absentes de la version disponible en Open Source.

Pouvez-vous décrire les technologies principales que vous avez développées ?
Carrot 2 est un ensemble de plusieurs composants reliés entre eux pour accomplir certaines tâches. Bien sûr, nous sommes très fiers de l'algorithme de clustering mais nous faisons également beaucoup d'efforts pour que le cadre soit ouvert et efficace. Ce cadre nous permet de facilement combiner des composants très divers.
Pour faire en sorte que des données alimentent nos algorithmes , nous avons développé des interfaces prenant en compte toute une palette de sources existantes de données textuelles telles que les moteurs de recherche Web, l'indexeur Lucene, les fils RSS/XML, les données de l'Open Directory etc…. Carrot 2 possède aussi des composants indispensables de traitement des données textuelles tels que l'analyse lexicale ou l'analyse basée sur la racine des mots.

Le nouvel outil Carrot s'adresse-t-il aux utilisateurs ou aux moteurs de recherche ?
La réponse n'est pas simple. Carrot 2 proprement dit est un cadre et s'adresse aux chercheurs et aux développeurs. Le composant Lingo, une partie essentielle du projet, peut être intégré à n'importe quel serveur ou à n'importe quel logiciel PC. C'est un composant dont le but est d'optimiser les solutions existantes de recherche d'informations. Au final cependant, notre cible est l'utilisateur car c'est lui que nous essayons d'aider avec notre technologie de clustering. Pour revenir à votre question, nous fournissons principalement une infrastructure aux développeurs mais Carrot 2 présente aussi quelques applications très simples qui peuvent intéresser l'utilisateur lambda.

De quelle nature sont les difficultés rencontrées lors de ce projet ?
Selon moi, la chose la plus difficile est de comparer notre algorithme aux autres. Il existe de multiples façons de regrouper un document textuel. Il n'est pas aisé de comparer Lingo à Vivisimo, par exemple. Les deux systèmes ont leur manière propre "d'interpréter" le résultat et il est souvent impossible de dire quel moteur regroupe "le mieux" thématiquement les données.
Nous rencontrons également les problèmes habituels : accroître l'efficacité, améliorer l'algorithme, rechercher les bugs etc…
De même, nous nous heurtons au problème du temps et de l'argent. Nous aimons le fait que le projet Carrot soit disponible en Open Source puisque cela lui apporte beaucoup d'avantages et de publicité. Toutefois, nous ne sommes que deux personnes à travailler activement sur le projet et cela nous prend énormément de temps. C'est pourquoi nous aimerions une compensation quelconque. C'est la raison pour laquelle nous avons créé une version commerciale : nous avons le plaisir de travailler en Open Source tout en offrant à des partenaires commerciaux potentiels des services à valeur ajoutée

Comment voyez vous l'avenir des technologies de recherche en Open Source ?
Il ne s'agit pas d'une question d'avenir. L'avenir se passe aujourd'hui, tout autour de nous. Regardez par exemple Google Labs, Microsoft ou les moteurs de recherche en Open Source comme Nutch ou Egothor ; ils ont des idées fantastiques qu'ils mettent déjà en pratique. La seule chose que je puis dire sur l'avenir, c'est qu'il faudra approfondir l'analyse sémantique, l'analyse contextuelle et la "restitution visuelle des résultats", c'est à dire le fait que les moteurs de recherche essaieront de prédire ce que vous recherchez et vous suggéreront d'une façon visuellement appétissante d'affiner votre interrogation. Le clustering des résultats de recherche est déjà disponible et vous pouvez déjà en profiter. Pour ma part, je ne m'en lasse pas.

Merci à tous pour vos contributions !