L'infrastructure distribuée de Google : comment faire tourner nos algorithmes sur des milliers de machines?
Les actions de Google autour du thème femmes et informatique

Alice Bonhomme-Biais

Retranscription de la conférence

Alice Bonhomme-Biais abordera ici deux thématiques consécutives. La première, plus technique, traitera de l'élaboration d'un moteur de recherche et des problèmes de passage à l'échelle qui y sont liés.
La seconde lui permettra de raconter son expérience en tant qu'ingénieure femme chez Google et aux USA, ses rencontres avec les groupes d'ingénieures femmes et les actions au sein de Google pour promouvoir les femmes en informatique.

Pour réaliser un moteur de recherche, le premier élément est un crawler, programme destiné à rapatrier toutes les pages Web et les stocker sur des machines appartenant à la structure. À partir de cette sauvegarde, on réalise un index de ces pages, par mots. Ensuite, on passe au ranking. Pour chaque mot indexé, on trouve des milliers de pages Web : quelle hiérarchie, lesquelles servir en priorité à l'utilisateur ? Et enfin, comment servir l'information souhaitée et sélectionnée, de manière agréable, à l'utilisateur en quelques millisecondes ?

Pour ce qui est du rapatriement du crawler, on procède en suivant pour chaque page les liens vers lesquels elle pointe. Une des difficultés importantes, dans ce cas, étant les cycles, les boucles, qu'on doit être capable de détecter pour ne pas tourner indéfiniment. On doit également construire un système tolérant aux fautes et respectant les instructions des sites pour les robots. A l'heure actuelle, il y a encore certains contenus qu'on n'est pas capables de trouver et de rapatrier.
On tient également compte de la fréquence de mise à jour des sites pour choisir la fréquence à laquelle on vient les visiter.

L'indexation se fait à partir d'un dictionnaire plus gros qu'un dictionnaire classique, puisqu'il va comprendre les orthographes erronées, des noms propres, acronymes, nombres, etc. On liste à partir de cet index tous les mots mais également leur position sur chaque page (ce qui sert aux recherches sur des expressions ou groupes de mots).
Pour ce qui est de l'indexation, on se heurte à la difficulté des nombreuses langues existantes, et notamment de celles qui utilisent d'autres systèmes de représentation des caractères. Il faut donc être capable d'identifier la langue pour ensuite découper en mots.
On commence également à travailler sur les lieux et les dates.

Une fois l'index réalisé, on établit le ranking, l'ordre de pertinence. On utilise pour cela deux éléments : la pertinence par rapport à la requête (fréquence de termes dans la page, comparée également à la fréquence moyenne ; position dans la page ; texte d'ancrage, etc.), et la qualité intrinsèque de la page (une page de bonne qualité pointe en général vers d'autres pages de bonne qualité, mais il faut aussi détecter les liens créés pour influencer les requêtes Google).

Une fois ces résultats obtenus, on doit les servir à l'utilisateur, et rapidement. L'index est donc répliqué un grand nombre de fois et conservé en mémoire directement sur des grappes de PC.
Enfin, on doit montrer les résultats à l'utilisateur, tout en prenant soin de garder une interface simple qui a fait le succès de Google.

La question centrale est ensuite d'implémenter ces solutions à très grande échelle et avec une pérennité assurée. On parle alors de passage à l'échelle, de nombre de machines, de disques durs, de données accumulées et produites. On a donc également de plus en plus de gens.
Sur chaque grappe de PC, on a une infrastructure distribuée générique qui permet à n'importe qui de paralléliser et distribuer son application.
Pour ce qui est de cette infrastructure distribuée, on a besoin de stocker les données de manière fiable, de faire tourner les processus sur un grand nombre de machines, et que ce soit simple.
Les solutions développées pour cela :
• Google File System, utilisé sur toutes les machines, pour stocker notamment des fichiers de plusieurs gigaoctets (par blocs de 64 Mo, tous répliqués trois fois),
• Global Workqueue, un système global de répartition des processus sur les machines,
• Map reduce, qui permet de faire tourner facilement des analyses de données sur plusieurs machines, en distribuant et parallélisant automatiquement le traitement de données tout en étant tolérant aux fautes.

Récemment, a été mise en place une solution permettant de croiser recherche, lieu et temps, pour identifier notamment des évènements ponctuels, concerts, etc. Cela crée de nouveaux problèmes, d'indexation, mais aussi de hiérarchisation des résultats, problèmes pour lesquelles, au démarrage, on ne dispose pas des masses de données habituellement disponibles pour Google.

De manière plus personnelle, quelle expérience de femme ingénieure à Google, et plus généralement aux Etats-Unis ?
Quelques mois après son arrivée, elle fut contactée parce que l'université de New York organisait une conférence sur les femmes et l'informatique, à laquelle Google désirait envoyer des représentantes. Sa première question fut : pourquoi faire ? Après tout, il n'existe pas de groupes ou de conférences sur les hommes et l'informatique. Elle s'y rendit cependant

pour se faire une idée directement. Il y avait notamment une intervention particulièrement intéressante à la fin de laquelle, ayant un grand nombre de questions, elle levait la main pour intervenir. Elle réalisa alors que c'était la première fois qu'elle posait une question en public pendant une conférence : le public étant essentiellement féminin, elle ne s'était même pas posé la question de le faire ou non.
Elle se dit alors : effectivement, nous interagissons ici différemment et ces groupes "Femmes et informatique" peuvent avoir un intérêt.
Elle resta donc en contact avec ces groupes, et ils interviennent régulièrement sur les campus. Quand ils faisaient des présentations techniques dans des amphithéâtres, elle remarqua qu'ils n'avaient jamais de questions d'étudiantes. Alors qu'après l'intervention, elles avaient de nombreuses questions intéressantes.
Autre exemple, après une présentation, une étudiante est venue la voir disant qu'elle était intéressée et efficace en informatique, mais contrairement à ses collègues, peu encline à passer ses jours et ses nuits exclusivement sur l'ordinateur. Elle avait envie de travailler à Google mais se demandait si elle pouvait postuler. Elle ne serait sans doute pas allée voir un homme, qui aurait représenté dans son esprit ces personnes qui passent leur vie sur la machine.
Ces groupes essaient donc toujours d'avoir une représentante parmi les ingénieurs, pour servir de modèle d'une part, et d'interlocutrice d'autre part.
Chez Google, l'objectif n'est pas tant de recruter plus de femmes mais surtout de faire en sorte que plus de femmes postulent. Et son impression est que plus la notoriété de Google augmente, plus le nombre de postulantes diminue. Un nombre important d'étudiantes ne postule sans doute pas en pensant que ce n'est pas pour elles.

D'autre part, Alice Bonhomme-Biais souhaitait parler également des actions menées en direction des minorités, dont font partie les femmes dans le domaine de l'informatique.
Il existe la bourse Anita Borg (une des premières femmes à obtenir une thèse en informatique), depuis 2003, visant le financement d'étudiantes. Il existe trois comités différents (USA, Europe et Australie) auprès desquels déposer des dossiers, pour des financements d'environ 5000 euros, mais aussi des invitations à des rencontres.
Sur la base d'actions plus indépendantes, de tutorat local à la participation à des festivals de sciences, ces idées peuvent également progresser. Notamment, une journée "Introduce a girl to engineering" a été mise en place permettant aux employés d'amener leur fille au sein de l'entreprise pour en découvrir les différents aspects, et notamment la journée d'un ingénieur. Grâce à ces actions, on peut espérer donner une image un peu moins masculine des sciences et de l'informatique.