Web, infrastructure, entreprise et recherche : convergences

Exposé invité - CORIA 2008

Florian Douetteau

Exalead S.A.
10 place de la Madeleine
75008 Paris
Florian.Douetteau@exalead.com

L'industrie de la recherche d'information a connu en 2007 une période de grande convergence capitalistique. Ceci est un signe indirect des coûts croissants de recherche et développement dans ce domaine : infrastructures matérielles, logicielles, accès aux utilisateurs, expertise des problématiques d'entreprise. Ces dépenses sont difficiles à internaliser au sein d'unités de recherche isolées et celles-ci doivent par conséquent accélérer leur interaction avec les acteurs industriels. Nous illustrons ceci autour de cinq domaines d'activité centraux : les modes d'interactions avec l'utilisateur, le profilage d'utilisateurs, l'architecture de gestion des savoirs dans l'entreprise, la maîtrise des contenus Web, les technologies de bases de données.

Modes d'interactions avec l'utilisateur. La recherche de mots-clefs, produisant une liste de dix résultats, s'est imposée comme la norme tristement absolue de tout système de recherche. De nombreux domaines de recherche prometteurs perdurent pourtant : questions et réponses en langage naturel, cartographies sémantiques... Pour chacun de ses domaines, un basculement des usages nécessiterait, outre recherche et idées, des ressources telles qu'un laboratoire de tests utilisateurs, la maîtrises de la création artistique.

Technologies de bases de données. Un système de recherche complet est amené à exploiter des nombreuses technologies de bases de données (relationnelles, textuelles, sémantiques, base XML, bases à colonnes haut volume) qui correspondent chacune à des compromis différents en terme de flexibilité, de requêtes, de transactions, de ratio entre lecture et écriture. Les recherches dans le domaine des technologies de recherche

sont fortement spécialisées, et n'ont pas l'occasion d'explorer les systèmes mixtes et leur intégration, faute de cas d'usages et plateforme logicielle suffisamment complète.

Profilage d'utilisateur. Ce domaine est devenu, via la manne publicitaire des liens sponsorisés, le principal moteur de développement de l'industrie de recherche d'information. Les problématiques de recommandation sont difficiles à appréhender sans des volumes conséquents de données utilisateur, jalousement gardées par les éditeurs de contenu. Les périphériques à capacité d'interaction réduite (mobile, télévision) offrent de nouvelles perspectives et demandent probablement de nouveaux paradigmes mêlant suggestion personnalisée, recherche et navigation.

Architecture de gestions des savoirs en entreprise. Les outils de recherche doivent pouvoir se projeter comme des composants logiciels et s'intégrer aussi bien à des besoins comme la gestion d'un intranet, qu'à l'aide à la décision, en passant par l'archivage numérique. Les technologies de manipulation d'ontologie métiers sont souvent mal comprises et gagnerait en recherche de simplification conceptuelle

Maîtrise des contenus Web. Nous estimons qu'actuellement seuls 5 à 6 pôles dans le monde maîtrisent la technologie permettent d'analyser et indexer une fraction pertinente du Web (plusieurs dizaines de milliards de pages). Une recherche réaliste sur l'analyse du graphe du web nécessite de collaborer avec des infrastructures lourdes.