

FORMATION D'INGENIEURS
INFORMATIQUE
Option de 3ème année
2011-12

COMMUNICATION HOMME-MACHINE
ET
DOCUMENTS ELECTRONIQUES

Responsables : Basarab Matei et Adeline Nazarenko

matei@math.univ-paris13.fr

adeline.nazarenko@lipn.univ-paris13.fr

Exploring a 'Deep Web' That Google Can't Grasp

New York Times, Alex Wright, Feb. 22, 2009

One day last summer, [Google](#)'s search engine trundled quietly past a milestone. It added the one trillionth address to the list of Web pages it knows about. But as impossibly big as that number may seem, it represents only a fraction of the entire Web.

Beyond those trillion pages lies an even vaster Web of hidden data: financial information, shopping catalogs, flight schedules, medical research and all kinds of other material stored in databases that remain largely invisible to search engines.

The challenges that the major search engines face in penetrating this so-called Deep Web go a long way toward explaining why they still can't provide satisfying answers to questions like "What's the best fare from New York to London next Thursday?" The answers are readily available — if only the search engines knew how to find them.

Now a new breed of technologies is taking shape that will extend the reach of search engines into the Web's hidden corners. When that happens, it will do more than just improve the quality of search results — it may ultimately reshape the way many companies do business online.

Thématique

Dans une société où l'information devient un élément de plus en plus critique, stocker, retrouver, explorer, mettre en forme et recommander l'information représente un enjeu majeur. Même si le domaine a connu de grands succès depuis 15 ans, le traitement de l'information doit encore faire face à des défis majeurs :

- Les moteurs de recherche et les systèmes de recommandation grand public doivent aujourd'hui gérer les contenus multimédia : retrouver des sources similaires à un texte, une image ou un air de musique connu.
- Le champ des outils d'accès à l'information professionnels doit répondre à des besoins de plus en plus précis et spécialisés : la qualité de l'information produite devient un enjeu majeur.
- L'exigence en termes de qualité et de fiabilité amène à repenser l'accès à l'information : il ne s'agit plus seulement de retrouver une page web, il faut analyser et qualifier l'information au regard des besoins des utilisateurs.

Tout le monde connaît les moteurs de recherche, incontournables pour trouver de l'information sur le Web, mais on sait moins qu'ils continuent à évoluer. L'innovation consistant notamment à

- développer de nouvelles architectures réparties (par ex. peer-to-peer) pour gérer le volume et la variabilité des données d'internet dans un monde où la pluralité de l'information est à la fois une chance et un défi;
- concevoir des interfaces qui assistent l'utilisateur dans sa recherche d'information et l'aident à analyser les documents retrouvés par le moteur de recherche (outils de visualisation, extraction des termes saillants, etc.) ;
- donner une nouvelle place aux utilisateurs dans des dispositifs où ils deviennent eux-mêmes acteurs dans l'exploration hypermédia et de jeux adaptatifs comme dans les chaînes d'annotation des contenus (web 2.0).

Au-delà du Web, on retrouve ces problématiques dans la gestion des données des intranets. Les entreprises sont aujourd'hui de plus en plus sensibilisées par ces problèmes dont seuls les grands groupes se préoccupaient, il y a encore 5 ans. La veille et la e-réputation sont devenues des enjeux majeurs pour le monde économique d'aujourd'hui.

L'actualité montre l'enjeu de ces questions :

- Beaucoup d'industriels travaillent dans ce secteur, depuis les créateurs et diffuseurs de contenu (jeux vidéo, éducation numérique, e-citoyenneté, presse, culture et arts, design, etc.), jusqu'aux PME éditeurs de logiciels et aux grands groupes soucieux de veille économique ou sociétale.
- Le pôle de compétitivité francilien Cap Digital¹ anime toute la filière économique des contenus numériques. Il rassemble les acteurs de ce domaine (plus de 650 adhérents, PME, grands groupes, organismes de recherche²), soutient l'innovation et cherche à améliorer la compétitivité de ce secteur d'activité, très présent en Ile de France.
- Beaucoup de projets innovants ont été lancés et continuent à l'être : citons à titre d'exemple, un grand projet national comme Quaero, les efforts en faveur de la numérisation du patrimoine, l'installation de Google en Ile de France et le projet de création de l'institut de la Vie Numérique³.

¹ <http://www.capdigital.com>)

² Dont l'Université Paris 13.

³ <http://www.capdigital.com/institut-vie-numerique>

On recherche des informaticiens !

Les informaticiens sont aujourd'hui au cœur de ces évolutions de société où la technologie joue un rôle majeur. Il leur faut maîtriser les technologies de base du secteur (les systèmes d'informations, les technologies du web et du traitement des données) et les concepts sous-jacents qui leur permettent de suivre les évolutions technologiques.

On recherche dans ce secteur des jeunes professionnels⁴ qui ont

- de bonnes compétences techniques,
- l'envie d'évoluer à la pointe de l'innovation pour introduire ce qu'il faut de nouvelles technologies dans les systèmes d'information des entreprises,
- le goût du travail pluridisciplinaire pour travailler en dialogue avec des spécialistes d'autres disciplines (ergonomes, veilleurs, économistes, par exemple), ce qui est une source d'enrichissement permanent.

C'est à cela vous prépare l'option CHM&DE en s'appuyant sur les compétences acquises par les étudiants au cours de leurs premières années de formation d'ingénieur. Il ne s'agit pas, dans le cadre de cette option, de former des spécialistes de la gestion de contenu multimedia mais de permettre à des ingénieurs dotés d'une formation solide en informatique générale de s'initier à ces questions et aux techniques sous-jacentes pour leur permettre d'intervenir en entreprise sur cette thématique, d'architecturer des solutions et éventuellement de collaborer avec des acteurs plus spécialisés.

Dans ce domaine de pointe où de nouvelles techniques sont en voie de développement, les étudiants intéressés par les aspects les plus innovants et attirés par la recherche en informatique peuvent compléter cette option par un parcours recherche en master 2. Le master informatique proposé par l'Université Paris 13 propose une spécialité « Ingénierie des textes et contenus numériques » (ITCN) qui partage des cours avec l'option CHM&DE et prolonge assez naturellement le programme de l'option.

Contenu

Cette option s'organise en 2 thèmes :

- Le thème 1 est une formation aux techniques de base de la communication sous ses différentes formes visuelle, orale et écrite ;
- Le thème 2, plus appliqué, montre comment ces différentes techniques sont mises en œuvre dans des applications documentaires conçues comme parties intégrantes d'une communication homme-machine.

Cette option s'appuie sur les compétences du LIPN en matière de traitement de l'information et analyse de contenu ainsi que sur un réseau de partenaires industriels⁵ qui sont consultés dans l'élaboration du contenu pédagogique, proposent des projets aux étudiants, donnent des cours ou des conférences et sont intéressés à recruter des stagiaires ou des jeunes recrutés.

⁴ Consultez les offres du jeune site Digilagents (<http://www.digilagents.fr>) et déposez-y votre CV.

⁵ Citons à titre d'exemple Exalead, Temis, Sinequa, ELDA, Thales, FranceTélécom, Mondeca, Ontoprise, IBM, Jouve, etc

Programme

THEME 1 : TECHNIQUES DE BASE DE LA COMMUNICATION HOMME-MACHINE

Traitement de données vocales (A. Nabeth Halber, Naturalvoice) (7,5h C, 7,5 hTD)

Introduction aux techniques de reconnaissance et synthèse de la parole. Indexation et gestion de données vocales.

Contrôle des connaissances : partiel, contrôle continu

Cours commun avec le master d'Informatique, spéc. EID

Traitement de données visuelles (B. Matei, LAGA) (10 h C, 10 h TP)

Acquisition et restitution de données visuelles, méthodes de base du traitement de données visuelles statiques, reconnaissance d'objets, indexation et recherche par le contenu, visualisation de l'information.

Contrôle des connaissances : Devoir

Fouille de données textuelles (A. Nazarenko, LIPN) (15 h C, 15h TD)

Techniques symboliques et numériques d'analyse du contenu textuel. Notions de pragmatique et de dialogue pour la communication homme-machine. Présentation et évaluation des logiciels et bases de données disponibles.

Contrôle des connaissances : partiel, contrôle continu

Cours commun avec le master d'Informatique, spéc. ITCN

THEME 2 : CONCEPTION D'APPLICATIONS DOCUMENTAIRES

Méthodes d'accès à l'information (H. Zargayouna, LIPN) (15h C, 10h TD, 10h TP)

Rappel des principes de la recherche d'information et de son évaluation. Méthodes avancées de recherche d'information (web sémantique et social, moteurs de recommandation, données multimédias) Développement et test d'un moteur de recherche.

Contrôle des connaissances : partiel, devoir implémenté

Cours commun avec le master d'Informatique, spéc. ITCN

Outils et manipulation de données textuelles (L. Audibert, LIPN) (15 h C, 15 h TP)

Outils de manipulation de texte (Perl et Java). Logiciels d'analyse de données textuelles et de traitement automatique des langues. Plateforme UIMA.

Contrôle des connaissances : partiel, devoir implémenté

Cours commun avec le master d'Informatique, spéc. ITCN