Christian Fluhr (New Phenix) : "Nous voulons associer la reconnaissance de formes à l'image et à la vidéo"

Retour au sommaire de la lettre

Directeur du Laboratoire d'Ingénierie de la Connaissance Multimedia Multilingue (CEA) et à l'origine notamment du moteur de recherche Spirit, Christian Fluhr travaille depuis plus vingt ans sur la recherche d'information intelligente. En 2004 il a créé la société New Phenix qui commercialise aujourd'hui une solution de recherche d'images utilisant à la fois les avancées technologiques de la recherche sémantique et de la reconnaissance automatique de formes... Il nous parle de sa vision de la recherche d'images sur le Web.

Directeur du LIC2M, Laboratoire d'Ingénierie de la Connaissance Multimedia Multilingue du CEA et à l'origine notamment du moteur de recherche Spirit, Christian Fluhr travaille depuis plus vingt ans sur la recherche d'information intelligente. En 2004 il a créé New Phenix (http://www.new-phenix.com/), entreprise privée issue d'un essaimage du LIC2M. En étroite collaboration avec le laboratoire, New Phenix commercialise aujourd'hui une solution de recherche d'images qui utilise à la fois les avancées technologiques de la recherche sémantique et de la reconnaissance automatique de formes. D'autres applications à la pointe du web 2.0 sont à venir. Une belle promesse d'avenir pour une technologie française qui a le vent en poupe...
Il a accepté de répondre à nos quetsions et nous l'en remercions.

Pouvez-vous vous présenter en quelques mots ?

Je suis Directeur de recherches au CEA et travaille depuis 1971 sur la recherche d'information et en particulier sur l'utilisation de technologies linguistiques pour la recherche d'information. J'ai notamment fait partie des chercheurs à l'origine du système Spirit.
Au début de l'année 2002 j'ai créé au CEA un laboratoire qui mixte le traitement linguistique et le traitement de l'image. Il s'agit du LIC2M, Laboratoire d'Ingénierie de la Connaissance Multimedia Multilingue. Ce laboratoire a pour but de traiter de l'information multimedia et multilingue, en combinant des technologies de détection automatique, qu'elles aient comme support le media texte, le media son ou le media image. Nous faisons travailler de manière étroite les spécialistes de ces différents domaines qui auparavant s'ignoraient, ce qui permet une bonne fertilisation croisée de leurs compétences. Il est également vrai que l'expérience des ingénieurs qui travaillent sur le texte est beaucoup plus longue que celle des ingénieurs qui travaillent sur l'indexation automatique d'images, qui est une discipline plus récente et qui a moins de dix ans. La fertilisation croisée a cela de positif qu'elle permet de capitaliser sur les expériences des différents spécialistes.

Vous avez participé à un essaimage ?

J'ai été en quelque sorte un pionnier de l'essaimage puisque le moteur de recherches Spirit, dont le cœur de technologie était issu de mon laboratoire, a été créé par la société Systex en 1979. Il est toujours commercialisé aujourd'hui par la société Technologies SA. C'est l'un des produits de recherche d'information ayant eu la plus longue vie.
Début 2004 les recherches menées au LIC2M ont permis de redévelopper complètement une nouvelle technologie basée sur les mêmes concepts, pour les parties son et image. Pour la partie parole nous travaillons en partenariat avec le LIMSI. Nous avons alors créé une *Start-up* nommée New Phoenix. Cette *start-up* a pour but d'industrialiser et de commercialiser à l'international les technologies issues du laboratoire. New Phoenix a un accord de licence exclusive mondiale pour la diffusion des technologies issues du LIC2M. Le CEA par sa filiale de capital-risque possède une participation dans la société et inversement la société finance des activités de recherche dans le laboratoire.

Cette société travaille-t-elle exclusivement sur la recherche d'images ou sur l'ensemble des technologies développées au sein du laboratoire ?

Elle travaille sur l'ensemble des technologies, sachant que les technologies de recherche d'images sont plus faciles à mettre au point que les technologies linguistiques et étaient donc prêtes avant. Les premières ventes se sont donc faites sur la recherche d'images. Mais très vite le marché de la recherche d'images a fait émerger un besoin dans le domaine de la recherche linguistique. Les clients en effet avaient deux besoins : en premier lieu il leur fallait, pour des questions de droits, un moteur qui permette de trouver une image semblable à une image donnée. Par ailleurs comme la plupart de nos clients font de la vente d'images sur Internet et qu'il s'agit d'un marché mondial, ils avaient aussi besoin de comprendre la langue dans laquelle les images avaient été décrites.

Donc il faut du multilingue ?

Jusqu'à ce qu'ils s'équipent avec notre technologie, c'était très coûteux pour les sociétés qui étaient obligées de traduire en anglais toutes les descriptions. Avec les technologies interlingues, les descriptions demeurent dans la langue d'origine et chacun interroge le moteur dans sa propre langue.

Comment est-ce que ça marche ? Y a-t-il de la reconnaissance de formes ? Les images sont-elles indexées linguistiquement ?

Au niveau basique, on recherche des caractéristiques basées sur les couleurs, sur la forme et sur la texture. Il ne s'agit pas d'un corpus de formes, base constituant un référentiel. Ce qui est utilisé aujourd'hui dans l'exploitation c'est de la ressemblance globale entre images utilisant un peu de chacun des critères : couleur, forme et texture, sachant que c'est la forme qui est la plus difficile à identifier.

Quelles sont les performances des systèmes ?

L'évaluation en recherche d'images est souvent une évaluation quelque peu empirique. Cependant grâce au programme Technovision, nous sommes en train de mettre en place des campagnes d'évaluation validées scientifiquement, comme cela se fait pour les moteurs de recherche textuels avec TREC.

Comment fonctionne généralement la recherche d'images sur Internet ? Quelles sont les technologies sous-jacentes ?

A un niveau basique, la recherche se fait sur du texte associé aux images et ne tient pas compte de la forme. L'indexation est réalisée sur les meta-données. Quand Google images fait une indexation, le moteur utilise soit des mots qui figurent dans le nom du fichier, soit des mots qui apparaissent dans le texte à côté. Le texte est associé de manière indirecte quand il s'agit du texte proche de l'image ou bien de façon délibérée quand les éditeurs d'images complètent un formulaire de description, comme c'est le cas sur Flickr, le moteur de recherche d'images participatif de Yahoo. D'une façon générale les moteurs de recherche grand public n'utilisent pas de technologie de reconnaissance de formes.
Le deuxième niveau c'est pour l'instant ce qui est utilisé par les clients de New Phoenix. Nos clients sont des agences qui vendent des photos sur internet. Ces photos leur sont communiquées par des auteurs ou par d'autres agences. Ils reçoivent les photos avec des descripteurs rédigés par les auteurs et ces descripteurs ne sont pas toujours très pertinents. Les auteurs pensent souvent que plus il y a de mots, plus ils ont de chances de vendre leurs photos, ce qui est une erreur et ne manque pas de générer beaucoup de bruit dans la recherche d'images. La description sous forme de mot-clés est donc très aléatoire.
Parmi les vingt à cinquante premières photos fournies par le moteur, si l'une ressemble un peu à ce qu'on recherche, on la signale au système. Ensuite cette photo sert de modèle pour la recherche d'images. Ce n'est pas de l'apprentissage, mais un calcul de distance entre deux images pour dire qu'elles se ressemblent globalement sur le plan de la texture, de la couleur ou de la forme.
Imaginons maintenant qu'un utilisateur trouve dans un magazine une photo dont il ne possède pas les droits de publications et qu'il souhaite l'utiliser à cette fin. Le service que nous proposons est le

suivant : il lui suffit de scanner cette image, de la soumettre au site et demander une photo qui lui ressemble et dont il peut, de fait, acquérir les droits.

Quel est le taux de satisfaction des clients et quelle est la concurrence sur le marché français ?

Nous n'avons pas réalisé d'évaluation de satisfaction numérique. Par contre nous avons constaté que les agences qui sont passées à ce type de technologie ont augmenté leur chiffre d'affaires. La recherche linguistique affine beaucoup la recherche sur les meta-données textuelles associées aux images. Si par exemple votre requête est le mot "glace", le système vous demandera si vous voulez une photo de dessert glacé, un miroir ou bien une photo de glacier de haute montagne. En revanche, si vous demandez "ice" en anglais, le système ne posera pas de question, car le mot n'est pas ambigu.

Quel est le niveau de concurrence actuel ?

Pour ce qui est de la concurrence, en France une autre société propose des utilitaires de ressemblance d'images. Il s'agit de LTU, une start-up de l'INRIA. Ils ont été rachetés récemment par une entreprise japonaise et se sont reconvertis dans le domaine de la sécurité avec une application qui utilise des technologies de biométrie et de recherche d'objets. Aux Etats-Unis il y a la société Virage, qui a été rachetée par Autonomy. On trouve également Convera, avec son produit Retrievalware qui réalise de l'indexation automatique de contenus texte et multimedia.

Quelles sont les perspectives technologiques de développement de New Phoenix ?

Ce qui va passer en exploitation sous peu et qui plaît beaucoup aux utilisateurs, c'est un module de classification automatique. Dans la mesure où on est capable de calculer une distance entre deux images, il existe des algorithmes qui sont capables de regrouper les objets qui se ressemblent, donc de créer des classes. Si vous prenez le mot "café", il s'agit encore d'un mot ambigu. Une classification permettra de regrouper dans des classes des tasses de café, des devantures de café etc.

Cela fonctionnerait-il sur Internet ?

Pour l'instant ça fonctionne chez New Phoenix à titre de démonstration. Mais nous sommes en train d'étudier les usages possibles. Nous avons effectué un test en interne sur Google images. Un clic sur une icône permet alors d'effectuer ce type de classification et déclenche l'utilitaire proposé par New Phoenix.
En tant que laboratoire nous nous intéressons à l'ensemble des utilisations possibles. Cependant nous n'avons pas encore décidé quel serait le meilleur mode de commercialisation de ce produit, puisque notre clientèle est composée essentiellement de professionnels. Techniquement il pourrait parfaitement être vendu à des particuliers et s'utiliser pour trier les résultats d'une recherche. Toutefois si un partenaire souhaite monter une opération de commercialisation, New Phoenix suivra.

A quelle échéance ?

D'ici la fin de l'année ou au début de l'année prochaine. Ces technologies ont été présentées au salon CEPIC, salon des professionnels de la vente d'images, lors de sa dernière édition à Biarritz, et cette technologie a eu beaucoup de succès.

Vous intéressez-vous également à la vidéo ?

Oui, nous indexons aussi bien la bande image que la bande son. Les traitements appliqués à la vidéo utilisent des technologies de reconnaissance de forme et de reconnaissance vocale. Ce que nous voulons, dans un proche avenir, c'est associer des concepts aux images et aux vidéos. Nous savons actuellement associer aux images des concepts généralistes qui indiquent par exemple

qu'une photo a été prise à la mer ou à la montagne. Nous allons élargir la gamme de nos concepts afin de décrire peu à peu le contenu précis de l'image. Actuellement par exemple notre système est capable d'identifier une quinzaine d'animaux différents. Nous adoptons pour le multimedia une démarche similaire à la démarche utilisée en traitement de la langue. En linguistique on a dû, pour interpréter les textes automatiquement, construire des dictionnaires, des ontologies, des grammaires. Or l'équivalent n'existe pas pour le multimedia.

Vous voulez donc créer une ontologie du multimedia ?

Le terme ontologie est trop spécialisé pour s'appliquer à un corpus généraliste. Nous l'employons par commodité. Pour être plus précis, le concept que nous mettons en œuvre est celui de "grounded ontology". Il s'agit d'une ontologie reliée à la réalité et représentée par des images et du son.
Actuellement, en vidéo l'identification de l'image ne peut se faire grâce au son, à l'image, au contexte fourni par le commentaire ou les dialogues. Or nous souhaitons créer ces ressources et sommes en train d'élaborer une vaste ontologie du multimedia qui associera des concepts linguistiques multilingues et des images d'objets ou de personnes. Le son jouera également un rôle important dans cette ontologie. Les sons caractéristiques d'une scène donnée seront identifiés automatiquement : on peut par exemple reconnaître un match de foot par son environnement sonore ou encore une personne par sa voix. Nous nous basons sur des points caractéristiques pour reconnaître un objet. La pertinence des formes est validée humainement.

Ces ressources sont-elles constituées manuellement ?

Non. Ce n'est pas possible. Nous sommes en train de mettre en place des outils qui permettent d'extraire du web des données significatives, afin de procéder avec certitude à l'identification de l'objet filmé. Si par exemple lors d'un journal télévisé, le commentateur dit : "M. Chirac pensez-vous que…" on sait que la séquence suivante sera une image de Jacques Chirac. A partir de là nous sommes capables de choisir la séquence photo la plus significative qui servira de matrice d'apprentissage automatique au visage du Président. Nous faisons de même pour le son. Le plus important pour l'apprentissage est de constituer un corpus d'images suffisamment représentatives de leur classe. Pour reconnaître un contexte il faut être capable de procéder à de la segmentation d'images et cela se fait essentiellement par apprentissage. De plus, il y a interaction entre les ressources visuelles et les ressources langagières. Prenons par exemple la classe "oiseau". Si l'ontologie nous dit qu'un oiseau peut aussi être un merle, il suffira de reconnaître un merle pour en déduire qu'il s'agit d'un oiseau, grâce à la sémantique ; d'où l'intérêt de disposer de ces deux types de ressources.

La reconnaissance d'images issues du web ne peut-elle se faire que dans la perspective d'un post traitement ?

Non. Se mettre derrière un moteur de recherche grand public comme Google ou Altavista n'est pas réaliste. L'analyse des séquences d'images demanderait quelques minutes et les internautes ne sont pas toujours patients. Par contre ce qui est envisageable, c'est d'effectuer un traitement de ces données à la source. Cela reviendrait à mettre au point un moteur de recherche d'images qui indexerait une grande partie du web.

Quelques mots en guise de conclusion ?

Cette "grounded ontology" ne pourra être réalisée que grâce à une collaboration entre plusieurs laboratoires. Il s'agit de créer une ressource qui pourra être mise à la disposition de la communauté. Jusqu'à présent nous avions des spécialistes qui travaillaient sur le texte, d'autres sur le son, et d'autres encore sur l'image. L'avènement du web 2.0 qui amène les spécialistes de ces disciplines à travailler sur le multimedia, incite non seulement les chercheurs des différents laboratoires à travailler ensemble mais à tenir des contraintes des uns et des autres. C'est un énorme chantier sur le plan international. De plus, les enjeux ne sont pas seulement ceux de la recherche d'information. Ce qui est en jeu ce sont toutes les applications qui touchent au dialogue homme/machine et dialogue *machine to machine*. Une grande application de ces travaux concerne les applications multimodales et les objets communicants.

Interview menée par Marianne Dabbadie
Directrice Innovation i-KM
Laboratoire GERIICO – EA 1060