Des nouvelles du Centre d'ATO de l'UQAM

Jean-Guy Meunier et Louis-Claude Paquin

Un Centre pour l'analyse des textes par ordinateur

Depuis 1984, le Centre d'ATO a privilégié une forme particulière d'information: le texte. Les projets touchaient divers aspects du traitement informatiques de corpus textuels provenant de divers «environnements» : journalistiques, administratifs, médicaux, légaux, sociologiques, historiques, pédagogiques, littéraires, etc. Les projets faisaient appel à des méthodes d'analyse avant tout centrées sur le traitement symbolique et, en l'occurrence, linguistique. À l'aide des progiciels développés au Centre, SATO, DÉREDEC, D_EXPERT, FX et ACTE, des méthodologies ont été mises au point, pour l'extraction de la connaissance dans les textes, pour l'indexation, etc. De plus, des analyseurs ont été construits pour traiter des problèmes textuels spécifiques : analyseurs morphologiques, syntaxiques, lexicaux, gestionnaires assistés de fouilles textuelles, dépisteurs de patrons thématiques, etc. On peut penser ici à FX-S, TERMINO, LCMF, SIPO, etc.

Le traitement par ordinateur de l'information dans une perspective cognitive

Depuis juin dernier, le champ d'intérêt du Centre s'est élargi aux multiples facettes du problème du traitement par ordinateur de l'information dans une perspective cognitive. Par information, il faut entendre ici le contenu signifiant des systèmes sémiotiques qui sont utilisés par des agents (machines ou humains) à des fins de communication et d'adaptation à leur environnement. Par traitement, il faut comprendre, de manière générale, les processus par lesquels ces agents produisent, structurent, conservent, rappellent, diffusent et transmettent cette information.

Outre les formes sémiotiques linguistique et textuelle, le Centre d'ATO s'intéresse maintenant aux formes iconique et structurée (au sens de base de données). Du point de vue des méthodes et des expérimentations, le Centre soutient surtout des recherches qui privilégient une perspective cognitive, c'est-à-dire qu'elles recourent à des modèles psychologiques, linguistiques, informatiques, épistémologiques, mathématiques, logiques ou encore sémiotiques. Et le mode d'expérimentation privilégié est la construction et le design formel couplés à des simulations et des applications.

Dans cet horizon cognitif, en vue de traiter l'information de manière «intelligente», deux modèles de représentation sont proposés. Le premier modèle est dit symbolique. La représentation y est construite comme un langage qui exprime un contenu édicté dans des règles ou des schémas inscrits dans la programmation. La représentation dans le second modèle est a-symbolique et se réalise directement dans des structures de mémoire associative dont le contenu est idiosyncratique et résulte d'un apprentissage. Selon que l'on adoptera l'un ou l'autre modèle, le traitement de l'information par ordinateur s'en trouvera évidemment profondément modifié.

Les principaux axes de recherche du Centre d'ATO

Le premier axe de recherche est lié à l'analyse et la construction de modèles capables de traiter diverses formes linguistique, textuelle, iconique et structurée de l'information.

Le Centre continue ses travaux dans le traitement linguistique de l'information. Il peut s'agir de questions théoriques et formelles aussi bien que de produire et de construire des outils linguistiques : lexiques, dictionnaires, correcteurs automatiques, analyseurs morphologiques,

syntaxiques et même sémantiques, etc. Certains travaux peuvent même s'attaquer à des dimensions plus discursives et énonciatrices. Et l'on pourra s'intéresser au français comme aux langues étrangères.

L'aspect textuel d'un document ne se réduit pas à sa stricte dimension linguistique. Que ce soit au simple plan matériel de sa gestion ou au plan structural discursif, narratif ou argumentatif, un texte pose des problèmes spécifiques de traitement. Aussi, outre les recherches déjà entreprises, le Centre développe-t-il des recherches sur des systèmes de traitement et d'analyse de grandes masses de données textuelles. Ces systèmes effectuent des opérations de rappel, d'indexation, de classification, d'archivage, mais surtout d'analyse descriptive et thématique.

Si les documents textuels et leur dimension linguistique occupent une grande place, la recherche ne saurait négliger l'information qui se présente sous forme iconique, telle les graphismes, les plans d'ingénieurs, d'électricien, d'architecte, d'arpenteurs, etc., ou les cartes géographiques. Grâce aux développements considérables qu'a connu la technologie électronique (écran sensible, archivage sur disque laser, lecteur optique, ordinateur parallèle, etc.), les documents iconiques sont de plus en plus fabriqués, traités et archivés par ordinateur. Si la digitalisation des documents iconiques est avancée, le traitement analytique demeure élémentaire et la recherche dans ce domaine est à faire. Aussi le Centre se sensibilise-t-il aux divers aspects de cette problématique.

Grâce aux développements considérables qu'a connu la technologie électronique (disque laser, lecteur optique, etc.), les documents iconiques tels les graphismes, les plans d'ingénieurs, d'électricien, d'architecte, d'arpenteurs, etc., ou les cartes géographiques sont de plus en plus fabriqués, traités et archivés par ordinateur. Si la digitalisation des documents iconiques est avancée, le traitement analytique demeure élémentaire et la recherche dans ce domaine est à faire. Aussi le Centre se sensibilise-t-il aux divers aspects de cette problématique.

La forme la plus classique de représentation de l'information dans les ordinateurs est celle des bases de données. Celles-ci présentent l'information sous une forme structurée aux paramètres réglés. Vu la nature spécifique de ce type d'organisation de l'information, son traitement a fait appel à des stratégies informatiques qui ont défini les approches classiques du traitement de l'information. Mais sa rencontre avec des perspectives cognitives en renouvelle grandement les méthodes d'exploration. En effet, la construction, la fouille et l'analyse des bases de données recourent de plus en plus à des stratégies de structuration syntaxique, sémantique et logique (par ex.: l'héritage), etc.

Le deuxième axe de recherches concerne le design des systèmes informatiques dans le traitement de l'information. Cet axe appartient à l'informatique pure et met en jeu des problématiques comme les suivantes :

- les stratégies algorithmiques : mathématiques, numériques, logiques, statistiques;
- les langages de programmation: formels, fonctionnels, orientés objets;
- le design et l'architecture des systèmes : systèmes experts, systèmes intelligents;
- les machines qui les réalisent : sérielles, parallèles, multimédias;
- les interfaces :
    - avec les humains : la convivialité, l'ergonomie;
    - avec d'autres machines : la télécommunication.

Le troisième axe de recherches est celui des modèles cognitifs dans le traitement de l'information en tant que tel. Tous les aspects précédents posent des problèmes théoriques touchant les structures cognitives que les systèmes doivent acquérir ou apprendre, représenter et manipuler. Aussi faut-il étudier la nature d'une représentation de l'information, les modèles formels logiques, mathématiques, linguistiques, algorithmiques, épistémologiques. Bien que cet axe soit plus théorique et abstrait, il

n'en demeure pas moins essentiel à la poursuite des objectifs généraux du Centre. Il est par ailleurs la garantie d'un renouvellement constant de sa visée générale.

Le projet de R&D avec ALEX INFORMATIQUE

En décembre dernier, la Compagnie ALEX INFORMATIQUE a signé un contrat de recherche avec le Centre ATO touchant les points suivants :

- L'acquisition de 5 ordinateurs parallèles VOLVOX (valeur approximative de $8 millions).
- L'acquisition de micro-ordinateurs PC compatibles 486 (5) et Macintosh FX (5) et de stations de travail Spark de SUN.
- Le versement d'un montant forfaitaire de $ 2 Millions à des fins de recherche.

Les ordinateurs que nous connaissons sont tous de type sériel. Les informations y sont traitées, l'une après l'autre; de même, les processus se succèdent les un après les autres. Les limites physiques de ces architectures ne nous permettent pas d'explorer des modèles associatifs et numériques qui par ailleurs s'avèrent intéressants dans le domaine des sciences cognitives. Le traitement parallèle ouvre une aire nouvelle dans la technologie informatique en permettant une immense augmentation de l'efficacité et de la productivité du traitement de l'information. En effet, le parallélisme permet non seulement de multiplier l'information qui subit un même traitement, mais aussi de multiplier les traitement s'effectuant sur l'information. Dans le premier cas l'accent est mis sur la segmentation de l'information alors que dans le second l'accent est mis sur la communication entre les processeurs. Le parallélisme, technologie qui deviendra de plus en plus accessible dans un avenir rapproché, s'avère un paradigme important et hautement stratégique.

Les machines VOLVOX sont de la classe des MIMD (*Multiple Instruction Multiple Data*); il s'agit des ordinateurs parallèles les plus généraux en ce que chaque microprocesseur peut faire tourner un programme de façon indépendante. La mémoire est distribuée à chacun des microprocesseurs. Ceux-ci sont appelés «transputer» parce que leur fonctionnement dépend d'échange de «messages» avec les autres micro-processeurs. Ceux-ci sont de type RISC (*Reduced Instruction Set Computer*); cette architecture tente d'obtenir que l'unité centrale de traitement (UCT) exécute une instruction à chaque cycle machine. Le VOLVOX comporte 64 transputers de modèle T800 de 32 bits avec quatre liens externes bidirectionnels qui possèdent une unité arithmétique à point flottant. Cette machine, construite par Archipel (France) n'a aucun périphérique. Elle est accessible à partir d'une machine hôte qui est dotée d'une carte à cet effet. Sur cette carte est logé un transputer qui active les autres par message. Toutefois une des machines est dotée de huit ports de communication, de sorte qu'elle est accessible par un réseau de type «Ethernet». La topologie des connexions entre les transputers est variable; elle peut donc être choisie par le programmeur. La machine peut fonctionner sans système d'exploitation, la communication entre transputers doit alors être assurée par un logiciel de routage des messages. Il existe de bons compilateurs : C, FORTRAN, PASCAL et OCCAM le compilateur développé spécialement pour les T800.

Le défi scientifique du projet ALEX

Le défi scientifique que pose le parallélisme est de deux ordres différents. Le premier touche à des questions de génie logiciel. En effet, pour tirer le meilleur parti de cette architecture il faut créer un environnement logiciel adéquat (interface, langage de programmation, etc) qui soit d'assez haut niveau pour ne pas avoir à se préoccuper, entre autres, de la connexion des transputers et du routage de leurs messages. Développer cet environnement est en soi une tâche de recherche à laquelle plusieurs projets se consacrent. Le deuxième problème est d'un autre tout autre type. Il s'agit de la modélisation du domaine à informatiser. En effet, le parallélisme impose une reconceptualisation de l'objet à traiter. Par exemple, un analyseur linguistique, traditionnellement approché dans une

perspective sérielle, doit être radicalement repensé si on veut le soumettre à un traitement parallèle. Dans ce domaine, les ressources du Centre sont les plus pertinentes. La conceptualisation et la modélisation, étapes antérieures à la conception des algorithmes, s'avèrent cruciales pour l'exploitation des prodigieuses ressources computationnelles offertes par les architectures parallèles.

Dans cette perspective, la recherche a proposé des modèles dits "connexionnistes", génétiques, neuronaux, etc. et qui sont des hypothèses qui se sont avérées intéressantes en traitement parallèle. Leur fécondité a été vérifiée surtout dans les domaines de la vision, des banques de données, de la linguistique et dans de nombreux domaines où il faut tenir compte de multiples variables opérant simultanément et se modifiant dynamiquement. Leur implantation dans une machine parallèle en révèle la véritable portée opérationnelle. Par ailleurs, dans l'état actuel de nos connaissances, rien ne garantit que tous les problèmes qui ont été classiquement traités de manière sérielle peuvent être transformés avec profit pour un traitement parallèle.

Afin de relever ce double défi, une formation technique dans le domaine de la programmation parallèle sera offerte aux chercheurs du Centre et à leur équipe; un échange international avec des spécialistes du domaine sera aussi initié.

En conclusion, l'hypothèse que nous avançons est que certains des domaines de recherche que nous privilégions (traitement du texte, analyse du langage naturel, reconnaissance de formes visuelles, fouille documentaire etc) dans une perspective cognitive pourront être adaptés à un traitement parallèle. Nous croyons que nos recherches sur le parallélisme nous permettront l'adaptation de logiciels déjà opératoires ainsi que la conceptualisation et l'expérimentation de nouveaux modèles de traitement qui déboucheront sur des logiciels pour le futur. Les recherches qui seront menées au Centre entraîneront d'une part le renouvellement de certaines dimension de la problématique du traitement de l'information et, d'autre part, la formation d'une main d'oeuvre qualifiée dans les ordinateurs parallèle.

Rappelons en terminant que du fait qu’il se spécialise dans le domaine du traitement de l'information, le Centre est appelé à rendre des services techniques à diverses personnes du milieu. Il met donc ses ressources professionnelles et matérielles et son expertise, moyemmamt entente, à la disposition tant du monde universitaire que des organismes publics et privés.