L'écrit et le document

Jacques Labiche

Ce numéro spécial de la revue Traitement du signal, réalisé à la demande du Comité de programme du Colloque National sur l'Ecrit et le Document, a reçu le soutien et un appui sans faille du comité de rédaction de la revue, qu'il en soit remercié.

Une sélection de quinze communications les plus représentatives du domaine a été réalisée par le comité de programme. Les auteurs sélectionnés se sont engagés à remanier en profondeur leurs communications pour les transformer en articles adaptés aux lecteurs de la revue TS; ils ont parfaitement réussi je crois, qu'ils en soient remerciés ici.

Le domaine comporte différents thèmes : de la segmentation à la reconnaissance de documents complexes en passant par l'authentification de signatures et la multimodalité; différents types d'outils sont utilisés : des outils de traitement d'images aux outils de l'intelligence artificielle en passant par les outils statistiques. Les articles décrivent des avancées dans l'un des thèmes en utilisant différents types d'outils pour une application particulière.

En imaginant que l'on souhaite réaliser une chaîne de traitement de documents complexes, contenant aussi bien du texte que des informations graphiques, on est amené à concevoir la succession des différents traitements, avec d'éventuels rebouclages. Ces traitements concernent la saisie en ligne ou hors ligne, la segmentation en entités textuelles ou graphiques, la reconnaissance de ces entités, la reconnaissance de la structure physique puis de la structure logique, et enfin l'analyse et l'interprétation du document considérée dans sa totalité. Aussi, les articles seront présentés ici en les ordonnant par rapport à la contribution qu'ils peuvent apporter à cette succession des différents traitements que doit mettre en œuvre cette chaîne.

Pour donner une grille de lecture supplémentaire, il semble utile de présenter en introduction les objectifs, conclusions et perspectives du Colloque National sur l'Ecrit et le Document 1994.

CNED'94
3ème Colloque National sur l'Ecrit et le Document
Rouen, 6, 7 et 8 juillet 1994

L'objectif du colloque CNED'94, à l'initiative du Groupe de Recherche en Communication Ecrite (GRCE) et du groupe AFCET Traitement Automatique de l'Ecrit (TAE), organisé par le laboratoire La3i de l'université de Rouen, a été de mettre en contact des industriels et des chercheurs universitaires ou privés afin de tenter de faire le bilan, aussi bien des besoins utilisateurs, que des nouvelles solutions apportées par les laboratoires dans le domaine de la reconnaissance de l'écrit et du document.

Le brassage d'idées et concepts issus de problématiques différentes a été très important du fait de la pluridisciplinarité naturelle au thème de l'écrit et du document qui concerne effectivement aussi bien la psychologie, la didactique, l'informatique, le traitement d'image, la graphologie, ou l'intelligence artificielle que la cartographie ou les statistiques.

Seules 38 communications orales sur 60 propositions ont été présentées au colloque pour assurer un bon niveau scientifique.

Les sessions du colloque ont été dédoublées :

- Reconnaissance des chiffres manuscrits
- Modélisation de la lecture et de l'écriture
- Signature et authentification
- Applications neuronales
- Reconnaissance de l'écrit
- Analyse de dessins techniques, plans et documents cartographiques
- Ecrit et multimodalité
- Analyse et reconnaissance de documents
- Concepts et outils pour les Systèmes d'Information et de Communication
- Segmentation de documents

Pour ces sessions on retrouve des problématiques communes. La première a été abordée par le biais de la stratégie, il s'agit de la coopération de méthodes. La deuxième concerne plus particulièrement la classification, il s'agit de l'apprentissage.

Le besoin d'une stratégie plus efficace se fait sentir dans les domaines de l'analyse de documents et de l'analyse de l'écriture manuscrite. En effet, si l'on enregistre des progrès significatifs pour la plupart des techniques : statistique, neuronale, traitements d'image ..., il apparait qu'aucune méthode n'est suffisamment robuste et qu'il est nécessaire de faire coopérer plusieurs de ces méthodes plus ou moins globales, avec accès au contexte, jusqu'à l'obtention d'un résultat « cohérent ».

Ce mot « cohérent » a d'ailleurs été le mot le plus utilisé durant le colloque. Cette cohérence des données extraites ne peut se justifier que par un accès à un niveau sémantique le plus élevé possible correspondant au problème posé, et nécessite des connaissances sur les contenus attendus.

La deuxième problématique importante abordée durant le colloque est l'apprentissage. Dès que l'on a à traiter un problème réel, dès que les données sont bruitées, dès qu'il y a une saisie digitale et une variabilité des données, les problèmes de classification deviennent très délicats. Il est nécessaire d'apprendre avec une importante capacité de généralisation, au moyen, par exemple, de réseaux neuronaux. L'évolution dans le domaine concerne l'apprentissage incrémental et la coopération de réseaux de neurones.

Les conclusions du colloque du point de vue de ces deux problématiques sont :

- Aucun système d'analyse de l'écriture manuscrite ne pourra être totalement général, ne pourra être totalement multi–scripteur, il est alors nécessaire de définir des « typologies » d'écriture pour choisir la meilleure méthode. Plusieurs tentatives nouvelles ont été présentées au colloque et concernent l'aspect neuro physiologique du mouvement du scripteur, aussi bien que l'analyse « fractale » du tracé.
- Il semble impératif de faire appel aux sciences cognitives pour élaborer de nouvelles méthodes et stratégies inspirées des stratégies humaines en faisant appel dès que possible au contexte et en utilisant au besoin une approche plus ou moins globale, ou une approche multi–résolution, puis une vérification basée sur les informations de plus haut niveau sémantique possible.
- Aucune des méthodes ne s'avère suffisamment robuste, les logiciels utilisent de plus en plus des coopérations de méthodes; statistique, connexionnisme ... Cette coopération semble devoir être assurée par des systèmes à intelligence distribuée de type tableau noir ou multi agents avec une vérification de la cohérence comme « principe intégrateur ».
- L'analyse ne peut plus se satisfaire d'une aide contextuelle apportée par des outils de type statistique ou de type vérificateur d'hypothèses, mais il est indispensable de faire également la liaison avec des outils d'analyse de la langue. La validation a posteriori des hypothèses émises ne peut être explicitée que par des outils sémantiques et s'exprimera par une vérification de la cohérence sémantique (une démarche top–down peut être envisagée en parallèle

par émission d'hypothèses basées sur des règles sémantiques). Ces outils devront prendre en compte la problématique de l'apprentissage pour tenir compte de la variabilité et du bruit. Cet aspect amènera à créer pour les différents logiciels des bases de connaissances structurées par niveau sémantique.

En ce qui concerne le bilan des besoins des industriels de la bureautique et les besoins exprimés par les grands corps de l'état, il apparait, en particulier à la suite de la session « Concepts et outils pour les Systèmes d'Information et de Communication », que l'un des débouchés les plus importants pour la reconnaissance de documents et pour la reconnaissance de caractères concerne les S.I.C. Les logiciels devront prendre en compte cet aspect et devront pouvoir s'interfacer facilement avec d'autres logiciels et des bases de données. Ils devront s'intégrer facilement dans des systèmes faisant appel aux techniques du groupware.

PERSPECTIVES

« ... Mais lire avec la machine, c'est aussi et surtout écrire – c'est à dire corréler, sous forme électronique – à partir de cette lecture : c'est organiser, au sein de fichiers hypertextuels, des bases de données personnelles, des systèmes d'archivage électronique, des séries corrélatives, sans perte de mémoire; ... »

Bernard Stiegler

À cause du développement du trafic sur les autoroutes de l'information, de nouveaux questionnements sont apparus, et ont pu occulter partiellement les problèmes fondamentaux du devenir de ces autoroutes. Comme Herbert A. Simon le faisait remarquer, le problème principal est de savoir où aboutissent, et d'où proviennent, ces autoroutes; le véritable problème est de concevoir des « parkings intelligents » pour les informations, donc de concevoir un stockage numérisé intelligent associé à une interprétation–indexation permettant de retrouver ces informations.

Le problème de la reconnaissance du document sera alors de savoir reconnaître pour interpréter, puis stocker intelligemment les informations dans des structures évoluant dynamiquement.

On assiste en effet actuellement à une véritable révolution dans le domaine des systèmes d'information et de documentation; de nouvelles modalités de fonctionnement émergent chez les professionnels comme le fait remarquer Jean Michel président de l'ADBS (association des professionnels de l'information et de la documentation) :

- l'acquisition et le stockage d'information peut être une démarche individuelle de libre « promenade » dans les bases de données via internet qui fournissent les « facettes » de nos futures connaissances,
- l'acquisition et le stockage d'information peut être une démarche de réflexions profondes et de restructuration des connaissances par des professionnels pour des professionnels.

Pour la deuxième modalité, il est certain que les chercheurs et industriels concernés par la reconnaissance de l'écrit et du document devront se poser des questions sur l'extraction et la structuration des informations contenues dans le document, de sa structure matérielle à sa sémantique en passant par sa structure logique. Ne seront utiles que les systèmes incluant ces fonctionnalités de reconnaissance du caractère, du schéma, de la structure logique... permettant de gérer la saisie et le stockage intelligent.

En plus de cette ouverture de la reconnaissance de l'écrit et du document vers les métiers du traitement de l'information, il faut également s'ouvrir vers l'ergonomie du poste de travail, et compter avec la qualité des interfaces homme–machine, car il s'avère qu'aucun produit ne fonctionnera totalement en mode automatique, mais que l'on s'oriente vers des systèmes de plus en plus efficaces d'aide automatisée à l'utilisateur. Par ailleurs, la saisie et l'archivage intelligents amèneront des bouleversements dans les flux de documents et dans l'organisation des postes de travail des services documentaires.

Enfin, après ces interrogations sur les concepts et paradigmes fondateurs de la recherche dans notre domaine, il faut noter qu'aucun résultat significatif ne pourra être obtenu si nous oublions que nos outils et techniques de base sont et resteront du domaine du traitement du signal et de l'image.

Jacques Labiche
Président du colloque CNED'94, Secrétaire du GRCE et du TAE
labiche@greyc.ismra.fr

Le prochain Colloque « CNED'96 » se déroulera à Nantes les 3, 4 et 5 juillet 1996.
Informations sur www : http ://www.ireste.fr/cned96