Vers un "CTRL +F" amélioré pour tout type de document numérique? Techniques et enjeux de la recherche de motifs

Aurélien Bénel, Sylvie Calabretto, Véronique Eglin, Jérôme Gensel, Elisabeth Murisasco, Jean-Marc Ogier, Thierry Paquet, Jean-Yves Ramel, Florence Sèdes, Nicole Vincent

Vers un « CTRL+F amélioré » pour tout type de document numérique ?

Techniques et enjeux de la recherche de motifs

aurelien.benel@utt.fr ; sylvie.calabretto@insa-lyon.fr ;
veronique.eglin@insa-lyon.fr ; jerome.gensel@imag.fr ;
elisabeth.murisasco@univ-tln.fr ; Jean-Marc.Ogier@univ-lr.fr ;
thierry.paquet@univ-rouen.fr ; jean-yves.ramel@univ-tours.fr ;
florence.sedes@irit.fr ; nicole.vincent@math-info.univ-paris5.fr

Résumé. Notre article s'intéresse aux techniques et enjeux d'une recherche de motifs généralisée dans les documents numériques. La première partie rappelle que pour être pleinement un *document* (au sens de preuve), le document numérique doit être accessible y compris dans ses fragments. La seconde partie porte sur la recherche de motifs dans des documents nativement numériques (expressions rationnelles, XPath, cas des arborescences multiples). La troisième porte sur les documents numérisés (formulaires, textes imprimés, textes manuscrits, graphiques, photographies, enregistrements audio et vidéo).

1 POUR QUE LE DOCUMENT NUMERIQUE SOIT UN DOCUMENT

Une définition de la notion de « document », donnée dans les années 50, a récemment été remise à l'honneur par le réseau *Pédauque* et la *Document Academy*. L'auteur, Suzanne Briet, part du sens étymologique de "preuve" du mot "document". Elle en déduit que si un animal à l'état naturel n'est pas un document, il l'est, par contre, une fois catalogué et exposé dans un zoo. Prenant l'exemple d'une antilope ramenée au jardin des plantes, elle développe ensuite les procédés permettant tout à la fois de conserver ce "document" et de permettre son accès (mise en cage, catalogage, enregistrement du cri, empaillage à sa mort), deux conditions nécessaires à l'établissement de documents secondaires (articles de journaux, nouvelles à la radio et au cinéma, communications

scientifiques, monographie avec planches, encyclopédie, analyses, traductions).

Cette question de l'accès au document, centrale dans le domaine de la documentation, a rapidement amené les professionnels du domaine à considérer des unités documentaires plus fines que celles nécessaires à la conservation. Par exemple, dans une revue scientifique ou les actes d'un colloque, c'est l'article et non le volume qui fera l'objet d'une description (auteurs, titre...) et d'une indexation matière. Si l'on prolonge la question de l'accès au premier document en observant ses citations dans les documents secondaires, les unités documentaires en jeu sont encore plus fines. Ces citations sont les traces visibles des fragments, en nombre beaucoup plus grands, sélectionnés et annotés par les lecteurs.

En fin de compte, un document numérique ne serait pas réellement un document, si l'on n'offrait pas des manières de faciliter l'accès à des fragments de ce document susceptibles d'être sélectionnés et annotés pour être ensuite cités et discutés dans d'autres documents. Cet article se propose de discuter l'utilisation dans ce but de la capacité des ordinateurs à rechercher des "motifs" formels dans les documents qu'ils soient nativement numériques ou qu'ils soient le résultat de la numérisation d'un document analogique.

2 RECHERCHE DE MOTIFS DANS LES DOCUMENTS NATIVEMENT NUMERIQUES

Au-delà du simple "CTRL+F" permettant de rechercher les occurences d'une chaîne de caractère, reconnaître un "motif" revient à chercher ce qui obéit des règles, à une grammaire formelle. Noam Chomsky définit cinq types de grammaires formelles de complexité croissante : les grammaires à choix finis, les grammaires rationnelles, les grammaires hors-contexte, les grammaires contextuelles et les grammaires générales. Cette typologie, conçue au départ pour le traitement des langues en linguistique, a connu plus de succès en informatique dans le traitement des langages.

2.1 Expressions régulières et chemins dans des arbres

Certaines de ces grammaires formelles peuvent être utilisées, non plus pour la validation du document, mais pour la recherche de fragments instanciant un motif. Ainsi, les grammaires rationnelles (par exemple les expressions régulières simples), reconnaissables par un automate à états

finis, permettent par exemple de retrouver dans les RFC de l'IETF les tables des matières (cf. Fig. 1). Ou encore des grammaires hors-contexte (comme des expressions XPath simples), reconnaissables par un automate à pile permettent de retrouver dans une thèse en XML DocBook tous les paragraphes citant un ouvrage (cf. Fig. 2).

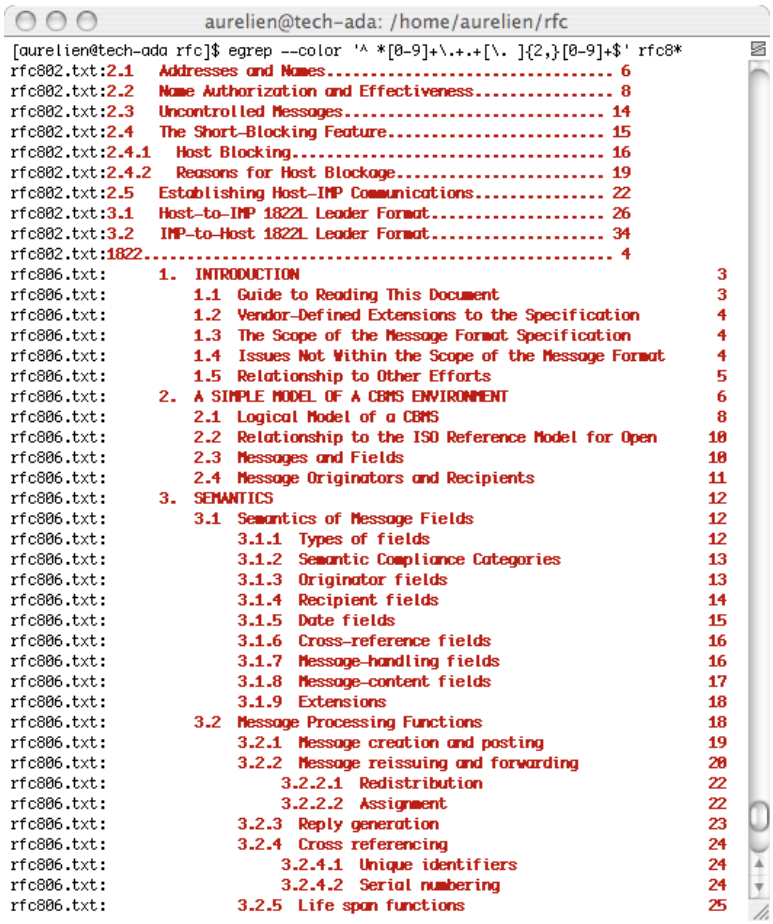

Figure 1 : Recherche avec *egrep* de tables des matières dans un texte brut

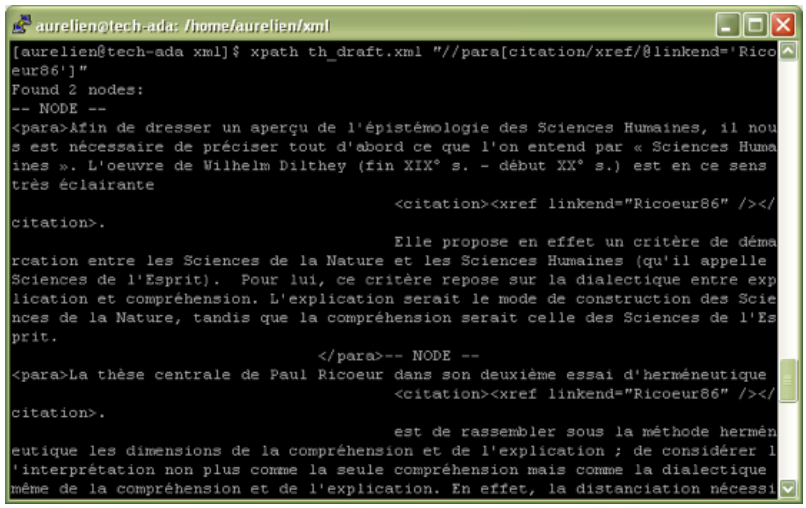

Figure 2 : Recherche avec *XPath* des paragraphes d'un document DocBook citant un ouvrage donné

2.2 Arborescences multiples

Pour des besoins d'usage différents, plusieurs structurations d'un même document peuvent être définies simultanément. Quand plusieurs structures partagent un contenu identique celles-ci sont généralement appelées structures *concurrentes* ou encore *parallèles*. Un document est dit *multi-structuré* plusieurs structures documentaires reliées entre elles via un contenu commun ou d'autres types de relations inter-structurelles. Les structures concurrentes sont alors des cas particuliers des documents multi-structurés. Les requêtes sur les documents multi-structurés combinent plusieurs structures simultanément.

Nous pouvons distinguer deux types d'approches pour modéliser les documents à structures multiples. Dans la première approche, les structures sont encodées séparément au format XML. Les principaux inconvénients de cette solution sont la redondance du contenu et la perte d'informations relatives aux relations entre les structures. La deuxième approche est fondée sur l'encodage de toutes les structures dans un même document XML sans répliquer le contenu, en les superposant les unes par rapport aux autres. Ainsi, le métalangage SGML offre une fonctionnalité optionnelle CONCUR permettant de faire référence à plusieurs DTD.

Grâce à cette option plusieurs types de balisage (structures) peuvent être utilisés au sein du même document SGML. Les balises sont identifiées par un préfixe indiquant la DTD où elles sont définies. Malgré son intérêt, cette option, n'a été implémentée que très rarement et ne proposait pas de solution pour interroger plusieurs structures simultanément. De plus SGML est aujourd'hui totalement remplacé par son successeur XML, avec une spécification moins complexe. En XML une seule DTD peut être référencée à la fois, et même si le mécanisme des espaces de noms permet de distinguer entre des grammaires différentes dans un même document, il ne peut pas remplacer la fonctionnalité CONCUR de SGML. La Text Encoding Initialive a traité le problème d'encodage de hiérarchies multiples. Dans la version XML des directives de la TEI, un ensemble de techniques d'encodage de plusieurs hiérarchies dans un même fichier XML a été présenté. Ces techniques se basent sur le choix d'une structure qui sera encodée en premier, ensuite les éléments des autres structures sont insérés dans le document. Pour que le document XML reste bien formé, les éléments insérés sont cassés entre les éléments de la première structure et diverses techniques (milestone elements, fragmentation, virtual joins, etc.) sont proposées pour reconstruire virtuellement les structures.

Plus récemment, il a été proposé d'utiliser le formalisme RDF (Resource Description Framework) pour encoder des structures de documents textuels. L'objectif de cette proposition est de tirer profit du modèle de graphe du langage RDF, pour pouvoir encoder des structures complexes comportant des entrelacements entre les éléments et de les interroger par une adaptation d'un langage de requête de RDF.

D'autres travaux se sont intéressés à la définition de nouvelles syntaxes permettant d'améliorer les possibilités de structuration des langages existants. MECS (Multi-Element Code System) a été le premier langage à supporter le chevauchement entre éléments. TexMECS est un autre langage basé sur MECS qui permet de définir des structures complexes où un élément peut avoir plus d'un parent. Cependant, ces projets ne s'intéressent pas à la problématique de l'interrogation de documents à structures multiples. LMNL (Layered Markup and aNnotation Language) définit une syntaxe basée sur la notion d'intervalle permettant l'encodage de structures multiples dans lesquelles les éléments peuvent s'entrelacer. Le projet LMNL envisage de définir un langage de requêtes adapté s'appuyant sur XPath .

D’autres chercheurs se basent sur le modèle GODDAG pour créer une représentation interne de structures XML concurrentes. La structure de graphe obtenue permet de faire abstraction du problème de chevauchement. Pour interroger les documents sous cette représentation, les auteurs proposent une extension du langage XPath. Cette extension est définie pour tenir compte des nouveaux types de relations entre les éléments de structures différentes. Il est par exemple possible d’atteindre tous les ancêtres d’un nœud dans toutes les hiérarchies (axe xancestor). De même, sont définis les axes xdescendant, xfollowing, xpreceding, et pour traiter les chevauchements, les axes overlapping, following-overlapping et preceding_overlapping.

Une proposition récente, MSXD (MultiStructured XML Documents) définit un modèle décrivant des structures XML indépendantes construites autour d’un contenu textuel identique. Les relations entre les éléments des structures sont modélisées par un ensemble de contraintes définissant un schéma du document multi-structuré. Le contenu textuel est répliqué dans chaque structure. Le schéma a pour rôle l’instanciation d’un ensemble de règles (les relations de Allen) permettant de décrire les positions relatives des fragments textuels dans les structures. Enfin, une extension de XQuery à la multistructure a été proposée.

Dans le modèle MSDM (Multi-Structured Document Model), le document multi-structuré est une entité dans laquelle sont mises en relation différentes structures. En plus des relations exprimant le partage du contenu, le modèle définit un autre niveau de relations entre les structures au sein du document multi-structuré, pour permettre d’expliciter des liens spéciaux entre les éléments de structures différentes. Outre ce modèle, le formalisme MultiX a été défini pour l’encodage des documents multi-structurés ainsi qu’un ensemble de fonctions XQuery pour l’interrogation des documents multi-structurés encodés en MultiX.

3 RECHERCHE DE MOTIFS DANS LES DOCUMENTS NUMERISES

Dans le cas des documents numérisés par procédé optique, la recherche d'expressions régulières ou de chemins dans les structures arborescentes ne semblerait disponible qu'après rétro-conversion du fac-similé numérique. Mais, si les logiciels de reconnaissance optique de caractères (OCR) affichent des taux de reconnaissance de 99%, cela signifie que demeure encore une erreur toutes les deux lignes. Quant à reconnaître

dans n'importe quelle mise en page la structure logique du document, cela reste encore hors d'atteinte des logiciels du commerce. Dans les sections suivantes, nous verrons comment des travaux de recherche récents tentent pour chaque type de document numérisé, de perfectionner les techniques de rétro-conversion, ou de proposer d'autres techniques de recherche de motifs sans passer par un rétro-conversion préalable.

3.1 Textes imprimés

Ce type de document rassemble les documents imprimés structurés contenant majoritairement du texte (on peut citer en exemple les journaux, les documents à structure complexe comme les sommaires de revue ou les pages de magazines, et les documents imprimés patrimoniaux, les documents scientifiques contenant également mais minoritairement des objets mathématiques ou quelques composantes graphiques). Bien que majoritairement textuels, ces documents sont d'une très grande diversité : qu'il s'agisse des polices utilisées ou des codes de mise en pages et d'agencement des textes.

L'analyse de documents structurés concerne le traitement numérique d'images de documents à forte teneur textuelle dont les contenus respectent un modèle de présentation et d'organisation permettant la compréhension par le lecteur. Cette activité s'est initialement développée autour d'une problématique de rétro-conversion de documents s'intéressant prioritairement aux documents modernes dont l'objectif était la sauvegarde et la ré-édition à la demande. L'objectif des premiers travaux visait ainsi la réalisation de systèmes de lecture complètement automatisés de documents papiers, numérisés au moyen d'un scanner ou d'une caméra. L'extraction de la structure physique de cette catégorie de document est une étape clé pour l'analyse de la mise en page et son interprétation (en structure logique). Dans ce type d'application, la détection de ces informations de contenus permet de lire le document, de le représenter sous des formats logiques pivots (XML, SGML, XSLT…) le rendant rééditable à la demande et donc ré-utilisable électroniquement avec des logiciels de composition et de production de documents ou dans une base de données. Avec l'arrivée récente des nouvelles problématiques liées aux documents imprimés anciens du patrimoine - où la grande variété de langues utilisées, les différentes variantes typographiques employées amènent un facteur supplémentaire de diversité - mais également des documents imprimés modernes de mises en pages très complexes superposant le texte à l'image et mélangeant les styles et les polices au sein d'un même paragraphe de texte, il a fallu repenser

l'interprétation et la représentation des structures pour pallier au manque de méthodes actuelles.

De récentes campagnes de numérisation ont conduit à la production de nombreux documents qui constituent une partie de notre patrimoine national, notamment les documents anciens ou les archives.

Numériser n'est pas simplement créer une image numérique, capturer et transformer un document en pixels. C'est mettre en oeuvre tous les traitements disponibles aujourd'hui (et élaborés dans les temps futurs) pour satisfaire au mieux le besoin ou le souhait de l'utilisateur. Numériser conduit à un nouvel objet dont les processus de mise à disposition pour le lecteur utilisateur sont encore objets de recherche, d'autant plus que de nouveaux usages apparaissent et que la demande sociétale va encore beaucoup évoluer.

Depuis décembre 2004 et l'annonce de Google qui prévoit la numérisation massive de plus de 15 millions de livres du patrimoine culturel mondial, la numérisation du patrimoine a connu un essor spectaculaire en occupant très largement les médias… Et pourtant il s'git d'un domaine qui est encore en pleine émergence nécessitant des techniques de pointe de traitement de l'image numérique en pleine mutation actuellement. L'ambition de tels systèmes se porte ainsi sur de nouvelles formes de diffusion des savoirs et des connaissances et de nouvelles formes de préservation d'un patrimoine précieux. Les établissements publics comme la Bibliothèque nationale de France (BNF), l'Institut national de l'audiovisuel (INA) ou la Réunion des musées nationaux (RMN), pionniers de la numérisation, tirent, en effet, une part importante de leurs recettes de ce qu'on pourrait appeler la commercialisation de ces savoirs à des fins d'éducation, de recherche et de tourisme.

Dans cette même lignée, se créait, il y a deux ans, le projet Euopeana, un prototype de bibliothèque en ligne développé par la Bibliothèque nationale de France, faisant appel aussi bien à des technologies avancées de reconnaissance automatique de caractères (OCR) appliquées aux documents patrimoniaux qu'à des outils d'ICR (Intelligence character recognition), qui concernent l'écriture manuscrite comme l'identification de signatures. Les enjeux sont de taille si on en croit la volonté des initiateurs du projet de concevoir des logiciels capables de reconnaître tous les types de caractères, depuis les incunables jusqu'aux écritures mayas. Il s'agit naturellement d'aller au-delà d'un simple exercice

d'archivage. Dans les mêmes temps, la BNUE (Bibliothèque numérique Européenne) prenait son envol en France en projetant de rassembler en France une collection d'ouvrages numérisés, riche de 5 à 6 millions de volumes accessibles à la lecture en ligne ou, pour la plupart, au téléchargement, permettant un accès aisé à l'ensemble de la culture française, mais surtout tisser des liens culturels à l'échelle du continent et même du monde.

Pris dans leur ensemble, les documents numérisés peuvent provenir de sources manuscrites, imprimées (à partir de la Renaissance), mais également audiovisuelles, photographiques, muséologiques, ou archivistiques (plans cadastraux, registres paroissiaux, archives sonores, films…). Pour ces documents et plus particulièrement pour les documents écrits, les principaux enjeux visent à s'intéresser à leur mise en ligne, leur valorisation par l'exploitation de nouvelles métadonnées (par une analyse fine des contenus sur laquelle nous reviendrons dans la seconde partie de l'article) afin de produire des environnements de travail, parfois collaboratifs, basés sur des outils technologiques avancés d'aide à l'expertise à destination d'usagers historiens ou littéraires ou encore de conduire à la réalisation de véritables éditions électroniques à destination de publics très variés.

Au delà même du texte, les images de documents anciens contiennent quantité d'autres informations accessibles visuellement qu'il est tout aussi intéressant d'extraire pour permettre d'offrir de nouveaux types d'outils aux utilisateurs : recherche et indexation des illustrations, lettrines, enluminures, tampons, graphiques, permettant selon le cas, de reconstruire l'alphabet de lettrines utilisées par un éditeur, d'étudier l'usure des tampons, d'étudier les dessins des lettrines ou encore d'identifier les faussaires qui imitaient certaines illustrations d'éditeurs... Des travaux récents sont en cours dans ce contexte : ils relèvent de la mise au point d'outils logiciels permettant la comparaison d'images comme on en trouve en indexation d'images naturelles.

3.2 Formulaires

Les industriels se sont rapidement emparés des techniques pour développer des produits de lecture automatique. En effet, les applications telles que le tri postal, la reconnaissance des montants des chèques ou l'analyse automatique de formulaires pour la vente par correspondance ou pour les grandes administrations permettent un gain de productivité et raccourcissent les délais de réponse.

Lorsque les documents analysés sont de simples textes avec une structure hiérarchisée bien définie, la lecture automatique de ces documents est guidée par un modèle explicite ou implicite du type de document et la conversion numérique est aisée. Mais ce n'est pas le cas pour les documents où l'information n'est pas toujours très organisée et le contenu est hétérogène (mélange d'imprimé, de manuscrit et de graphiques). On retrouve ici les formulaires, les documents postaux ou techniques, les chèques. On retrouvera ces problématiques dans les documents composites (les magazines) ou les documents d'archives.

Dans le domaine du traitement automatique du courrier d'entreprises, par exemple, on peut distinguer plusieurs contraintes induisant une conversion numérique parfois très complexe. Parmi ces contraintes, on retient : la très grande variété de documents (texte manuscrit ou imprimé, qualité, couleur et texture de papier différentes), les impératifs de temps réel (temps de traitement limité), l'adaptation au mode de capture par système de caméra linéaire (on devra développer les outils d'analyse d'image à la particularité de cette prise d'image pour optimiser les temps de calcul), l'obligation de résultats (le système doit être le plus performant possible pour éviter les coûteuses interventions manuelles) et parfois certaines contraintes particulières comme la grande diversité des images. On retrouve aussi d'autres contraintes particulières liées à des applications industrielles spécialisées où un bon nombre d'images est réparti en nombreuses catégories de courriers des clients : Courriers internes manuscrits ou dactylographiés, formulaires, plans, cartes bleues, listings… Ces images sont très différentes du point de vue de leur taille, de leur orientation, des couleurs du fond et du texte, de la position de texte dans l'image, de la taille des caractères et des types d'écritures. L'échec de reconnaissance de ce type de documents s'explique généralement par un dysfonctionnement des étapes de prétraitements et en particulier des étapes de segmentation et de localisation des zones informantes (localisation du bloc adresse sur objets plats par exemple, localisation du code postal…).

3.3 Textes manuscrits

Depuis ses débuts dans les années soixante, la lecture automatique (ou reconnaissance) de l'écriture a connu de grands progrès notamment pour alléger un travail manuel très fastidieux de lecture employé dans divers secteurs, tels que le tri du courrier, la lecture des montants des chèques, ou la lecture de formulaires. Ce type de reconnaissance consiste à interpréter l'écriture contenue dans une image. Longtemps focalisées sur des problèmes de reconnaissance de mots pour ces applications phares du

domaine, la reconnaissance de l'écriture à partir d'images vise aujourd'hui le traitement et l'analyse de documents beaucoup moins contraints, tant du point de vue de la mise en page que du point de vue du vocabulaire de travail.

Alors que les techniques d'OCR appliquées aux documents imprimés structurés fournissent des taux-records de reconnaissance avoisinant les 99%, les techniques de lecture de l'écriture hors-ligne sont toujours assujetties à la qualité des tracés et la régularité des formes de plus en plus incertaine avec les nouvelles applications actuellement en jeu (lecture de manuscrits libres, de brouillons d'auteurs…).

Améliorer la reconnaissance de l'écriture hors-ligne constitue un enjeu de taille si on se réfère au nombre de plus en plus important de documents manuscrit en circulation aujourd'hui. De très nombreux impacts socio-économiques sont alors visés par l'amélioration de ces techniques allant jusqu'à redonner une deuxième vie à certaines collections anciennes manuscrites du patrimoine très récemment numérisées. De nouvelles applications d'accès aux contenus, de navigation dans de grandes bases d'images de manuscrits ont ainsi pu voir le jour ces dix dernières années. On peut ainsi citer les nouveaux enjeux de la recherche : (i) l'extraction d'information dans les documents manuscrits ; (ii) la navigation dans les documents anciens ; (iii) l'authentification des mains ; (iv) l'aide à la transcription.

Ces applications loin de se contenter d'une analyse exclusive du tracé des formes en présence, se portent également sur l'analyse de la mise en page du document. L'analyse de la mise en page d'un document manuscrit est une étape indispensable et préalable à son indexation par le contenu. La variabilité inhérente à l'écriture manuscrite repousse les techniques développées pour l'analyse des documents structurés à leurs limites. Cette problématique appelle de nouveaux développements pour parvenir, par exemple, à analyser des courriers libres, des brouillons d'auteurs, des manuscrits anciens du Moyen Age, etc.

Longtemps cantonnée à quelques mots, une signature ou une adresse, la reconnaissance de manuscrits est appliquée aujourd'hui à des documents complets. Les travaux de recherche actuellement se sont tournés vers le traitement de manuscrits du patrimoine, qui peuvent être des documents d'archive comme les fiches militaires, des brouillons d'auteurs ou encore des écrits paléographiques. De nombreux projets voient actuellement le jour autour de nouveaux corpus manuscrits en

passe d'être numérisés. L'ensemble de ces recherches portent sur des documents manuscrits de qualité variable qui ne présentent pas toutes les mêmes difficultés en termes de qualité d'acquisition des images, de formats d'images, de qualité des contenus.

Après une période de relative stagnation dans le milieu des années 90, les activités dans le domaine de la reconnaissance de l'écriture manuscrite en-ligne sont de nouveau sur le devant de la scène. Ce renversement de tendance se justifie pleinement par l'essor important de toutes les technologies liées aux systèmes mobiles de télécommunications et aux nouveaux systèmes communiquants. Avec l'arrivée récente de nouveaux dispositifs d'interaction Homme-Machine tels que les PDA, les téléphones portables nouvelles génération, les Smartphone, les tableaux interactifs, le « e-papier », les TabletPC ou ardoises électroniques, le papier digital, ou encore le stylo-caméra, l'utilisation de la souris et du clavier ont été remis en cause. En effet, ces nouveaux terminaux peuvent être utilisés pour la prise de notes ou le dessin de schémas, par exemple : l'utilisation d'interfaces orientées stylo est alors plus naturelle que celle d'un clavier et d'une souris, puisqu'elle se rapproche du dessin sur une feuille de papier. De plus, ce type de périphérique permet également la prise de notes en environnement contraint, en station debout, par exemple. Il s'agit là d'un jalon significatif, dans l'évolution de notre société où des usages nouveaux sont en train de naître basés sur les concepts de nomadisme et d'accès permanent à l'information, en tous lieux et en tous temps, avec un minimum d'encombrement pour l'usager et un maximum d'ergonomie.

Plusieurs défis se posent alors. Ils concernent tout d'abord des aspects matériels liés à l'environnement de saisie des données mais aussi toutes les méthodes et modèles ou normes de représentation, de manipulation et d'interprétation de l'information ainsi saisie.

Dans le cadre de la reconnaissance en-ligne, l'écriture est analysée à travers des échantillons d'encre, constitués d'un ensemble de coordonnées ordonnées dans le temps : cette dynamique est enregistrée sous la forme d'"encre électronique". Il est ainsi possible de suivre le tracé, de connaître les posés et levés de stylo et éventuellement l'inclinaison et la vitesse. Il faut évidemment un matériel spécifique pour saisir un tel échantillon, c'est le cas notamment des stylos électroniques ou des stylets sur agendas électroniques ou sur tablets PC. Un tracé en ligne n'est pas perçu comme une forme pleine et n'a pas d'épaisseur. Certains facteurs externes liés à l'hésitation dans la production de l'écriture, le manque d'application, ou encore la variabilité de l'échelle (taille des caractères non régulière)

compliquent fortement la reconnaissance. C'est la raison pour laquelle, les données initiales, enregistrées en ligne, sont plus structurées qu'en hors ligne. L'analyse de la suite chronologique de points (en ligne) plutôt que l'analyse d'une image de traits (hors ligne) permet d'éviter tous les problèmes liés à une caractérisation rigoureuse et à une reconstruction complète des tracés nécessaires à une analyse hors ligne.

Il semble néanmoins que les propositions actuelles liées à la reconnaissance de l'écriture en ligne se trouvent confrontées à une plus grande difficulté d'utilisation (apprentissage long, gestes en petit nombre et difficiles à mémoriser). Il n'en demeure pas moins que les enjeux du succès de ces méthodes sont très importants, notamment pour les applications où l'interaction Homme-Documents est en jeu, comme l'édition et l'annotation de documents électroniques, l'utilisation de tableaux interactifs et intelligents, ou encore la reconstruction de version idéale d'un document à partir de son brouillon.

3.4 Graphiques

Quel que soit le contexte ou le domaine concerné, le document graphique, ou à forte teneur graphique, est utilisé comme support de représentation d'informations dans le cas où le graphisme peut apporter une plus-value par rapport à une information purement textuelle. Ils contiennent toujours une quantité importante d'informations, peuvent être d'une grande complexité et leur domaine d'application est très étendu. Les documents graphiques sont généralement constitués de lignes, de régions pleines, de régions hachurées, de texte, de symboles ... Les domaines concernés sont très nombreux, qu'il s'agisse des cartes et plans (géographie, topologie, hydrographie, cadastre, plans d'accès…), d'informations techniques (schémas, plans d'architectures, plans d'évacuation, etc.), d'informations organisationelles (organigrammes), ou encore d'une superposition de ces différentes classes d'informations (bases de données urbaines, plans de réseaux téléphoniques, électriques, etc.). Quelles que soient les institutions au sein desquelles les documents graphiques servent de support décisionnel, les volumes impliqués sont généralement très importants (490 000 planches cadastrales, 2 500 000 plans du réseau téléphonique français…) et l'automatisation de leur analyse reste un défi tout à fait essentiel dans le cycle de vie des organismes concernés. Pour faire face à ce défi, différentes stratégies ont été retenues, suivant les usages et les communautés d'intérêt concernées. Certains organismes ont opté pour la gestion d'armoires électroniques sous forme de documents « morts » (issus d'une simple acquisition), alors que d'autres ont choisi de

procéder à une numérisation manuelle des informations, permettant ainsi de manipuler une information dynamique et mise à jour. Entre ces deux extrêmes, d'autres organismes ont opté pour des solutions intermédiaires, consistant par exemple à ne traiter que les documents « vivants » au sens de leur activité (certaines archives ont des taux d'utilisation ne justifiant pas toujours un traitement automatisé). Mis à part certains secteurs de haute technologie en CAO-DAO (aéronautique, automobile…), pour lesquels de grands volumes de documents ont fait l'objet d'une rétroconversion complète, ces différents choix ont conduit à la création de documents numériques de niveaux de structuration variables, et font émerger de nouveaux problèmes relatifs à la navigation dans ces bases documentaires dites « semi-structurées ». Le problème de base qui est posé est donc celui de la conversion d'une information faiblement structurée — telle que l'image d'un document papier, ou un fichier PDF, par exemple — en une information enrichie des structures qui la rendent exploitable directement au sein d'un système d'information. À côté de ces applications traditionnelles de la reconnaissance de graphiques, de nombreux nouveaux usages apparaissent actuellement, pour lesquels la question n'est plus tant de réaliser une rétroconversion de l'information véhiculée par le document, mais plutôt d'intégrer de manière la plus transparente possible les documents disponibles sous forme papier, voire sous forme électronique, mais faiblement structurée, dans la chaîne globale de la documentation de l'entreprise. On s'intéresse alors à l'utilisation des algorithmes de reconnaissance à des fins d'indexation par le contenu, voire de réingénierie – au moins partielle – du contenu. Notion d'usages qu'il convient de développer afin de montrer que l'analyse graphique n'est pas que de la rétroconversion « Scan to XML ».

Dans ce contexte, il faut noter que la réussite d'une chaîne de rétroconversion, même partielle, nécessite la mise au point d'outils de segmentation robustes basés sur une extraction d'indices de bas niveau précises et fiables extraits de l'image scannée. Un bon nombre de recherches portent actuellement sur l'estimation robuste de primitives vectorielles et la modélisation statistique des composants. Au-delà de l'analyse bas niveau, la reconnaissance de symboles est un des domaines phare de l'analyse des documents graphiques. Un grand nombre de travaux lui sont consacrés. Cette activité consiste à localiser et à identifier, dans un document graphique, les symboles qui y sont présents. Alors que de nombreuses méthodes de reconnaissance de formes nécessitaient ces dernières années une phase préalable de segmentation pour extraire des régions d'intérêt contenant les symboles à reconnaître, les travaux récents visent au contraire à développer des méthodes sans

segmentation, robustes au bruit, efficaces et suffisamment génériques, se basantsur les propriétés fondamentales des traits pour parvenir à cette reconnaissance. Cette généricité peut être mise en œuvre en partant du postulat récemment mis en évidence qu'au sein des différents types de documents graphiques, un certain nombre de points communs ont pu être relevés. Ils se ramènent à l'existence essentielle d'une organisation des données en couches fortement dépendante du domaine d'application auquel on s'intéresse (couche de symboles, de relations entre symboles, de légendes, d'annotations, de champs textuels…) et qui participe fortement à l'établissement de relations entre objets et participent à leur reconnaissance.

3.5 Photographies, enregistrements vidéo et audio

Dans le cas d'une collection d'images, le processus de segmentation s'attachera à identifier des régions ou formes, auxquelles on associera des motifs tels que couleur, texture, etc. Les plus souvent, les techniques d'extraction d'information liées à l'analyse d'une seule image visent en effet à la segmenter en régions et à extraire pour chacune des informations concernant les motifs qu'elle contient. C'est ce qui est proposé notamment par les systèmes d'interrogation de bases d'images QBIC et NETRA. Le standard MPEG-7 propose, dans sa partie dédiée à la vidéo, des schémas de description des couleurs des régions d'une image ou d'un segment, par détermination de différents espaces de couleur, en utilisant les couleurs dominantes ou des histogrammes. MPEG-7 peut également rendre compte de la texture des régions, à un bas niveau de description correspondant à des filtres de Gabor, ou à un plus haut niveau en utilisant trois caractéristiques : la régularité, la direction et la granularité. Enfin, la forme des régions d'une image peut être décrite par une représentation des contours basée sur la courbure multi-échelle ou sur des histogrammes de formes.

En ce qui concerne les documents audiovisuels, les systèmes de gestion ou de présentation de vidéos ont bénéficié des progrès réalisés dans la compression des données (par exemple, les formats MPEG-1, 2 et 4), dans les vitesses de transfert, les systèmes d'exploitation et les capacités de stockage. Ces évolutions ont conduit au succès de divers types d'applications autour de ce média : vidéo à la demande, vidéoconférence, vidéo personnelle… Simultanément, des efforts de recherche ont conduit à l'extension des SGBD afin qu'ils soient capables de gérer des données de type vidéo, et ce non plus à travers de simples BLOB (Binary Large OBjects). En effet, les SGBD sont des outils

puissants pour résoudre les problèmes posés par la vidéo en prenant en charge la modélisation, le stockage, l'interrogation et, dans certains cas, la présentation. Les défis sont nombreux et de taille. La structure hiérarchique et la décomposition de la donnée vidéo en séquences, scènes, plans et images doit être décrite. Il en va de même pour les caractéristiques signal ou bas niveau (qui en général, sont extraites automatiquement par différents analyseurs ou extracteurs), et les caractéristiques haut niveau (ou sémantiques qui proviennent le plus souvent d'une annotation manuelle ou semi-automatique) ou bien encore les métadonnées associées. La composition de vidéos à partir de segments d'autres vidéos nécessite une algèbre. L'interrogation des vidéos doit permettre de consulter les différents aspects de la modélisation d'une vidéo (composition, caractéristiques bas et haut niveau, méta-données). Enfin, les aspects temporels et spatiaux sont également importants, ils servent notamment à rendre compte du mouvement des objets dans des scènes de la vidéo…Par ailleurs, le standard MPEG-7 s'intéresse plus particulièrement au problème de l'annotation. L'objectif général de MPEG-7 est de proposer des standards de description pour l'indexation et la recherche de segments audio-vidéos. Les descripteurs MPEG-7 sont définis, soit sous un format XML, soit sous un format binaire.

Dans le cas d'un média audio, des algorithmes de traitement du signal permettent de la segmenter selon une dimension temporelle et d'effectuer un étiquetage sémantique. Ce type de segmentation, basé sur des logiciels comme Transcriber, est multidimensionnel et permet par exemple, pour une facette et selon une dimension temporelle, de construire une liste de segments par type de signal (parole, musique, bruit), ou en identifiant et détectant des sons-clés (jingles, mélodies, silences, génériques, …). Cette liste se mémorise dans un descripteur XML. Sur l'ensemble de segments de niveau « physique », s'applique un processus que nous désignons par « étiquetage sémantique » qui consiste à définir un ensemble de facettes et de les valoriser pour chaque segment. Par exemple, on peut identifier les facettes : type de signal, type de signal vocal, type de signal musical, etc., qui peuvent être complétées par type de locuteur, les mots prononcés, … Ces facettes sont hiérarchisées en strates incluant les dimensions spatiales et temporelles. L'étiquetage sémantique génère la valorisation de la facette « type son ». La segmentation temporelle se traduit par le partitionnement de la dimension par des intervalles délimités par les bornes de début et de fin d'intervalle.

4 CONCLUSION

5 RÉFÉRENCES

[Revues] IJDAR, PAMI, .MVA, PR, PRL, PAA, CVIU, TS, IEEE Multimedia, IPM, JODL, Multimedia Tools and Applications

[Conférences] ICDAR, CFD, CIFED, CIDE, DAS, DLIA, IWFHR, GREC, WDA, ICFHR, DIAL, ACM Multimedia, CBMI, DocEng, ECDL , JCDL, MMM, ICME, Hypertext, MMM, RIAO