REE REVUE DE L'ELECTRICITE ET DE L'ELECTRONIQUE

N° et date de parution : 72 - 01/02/2007
Diffusion : 3000
Périodicité : Mensuel
LAREVUEDELE_72_81_310.pdf
Page : 81
Taille : 95 %

Analyse de l'information par fusion multimodale

Mots clés

Nouvelles technologies de l'information et des communications NTIC, Nouvelles technologies de l'analyse de l'information NTAI, Analyse et fusion de l'information

■ Jean-François MARCOTORCHINO
Thales Land & Joint Systems, Laboratoire de Statistiques Théoriques et Appliquées (LSTA), Université Paris VI
Avec des encarts de Christian Malis [1], Patrick Constant [2], Henri Maître [3]
Thales [1], Pertimm [2], Telecom Paris [3]

1. Introduction

Dans les deux numéros récents de la revue REE (N°8 et N° 9 de septembre et d'octobre 2006), Alain Appriou et Pierre Borne ont réalisé un dossier sur la fusion d'information et son analyse avec l'aide de spécialistes des domaines concernés (Fusion type capteurs et Théorie des Possibilités et des Fonctions de Croyance : Didier Dubois et Henri Prade ; Réseaux de Neurones : Pierre Borne, Mohamed Benjereb et Joseph Haggège, Séparateur à Vaste Marge : Stéphane Canu etc..). Ce dossier se présente, sous forme d'un tour d'horizon assez complet et scientifiquement étayé, intitulé « Du traitement Numérique à la Gestion des Connaissances : de Nouvelles Voies d'Investigation (1ère et 2 ème Parties) ». Si ce dossier couvrant deux numéros de la REE, traite, avec une réelle maîtrise, de techniques assez peu vulgarisées, quoique connues des spécialistes, il s'adresse à des problématiques qui relèvent presque uniquement du traitement de données numériques. Il existe, néanmoins, un large champ technologique dont les approches citées précédemment peuvent participer bien sûr, mais qui va bien au-delà, ne serait ce que parce qu'on y aborde également des données et des informations de nature non structurée (Parole, Textes, Sons, Images etc...) en supplément de données numériques classiques (que les anglo-saxons nomment « Numerical Data »). Ce domaine porte un nom : il s'agit de « **l'Analyse Avancée de L'Information par Fusion Analytique des données Multimodales** ».

En effet il faut faire le constat aujourd'hui, que l'information et les données dont on dispose et qu'on aura de plus en plus à consulter, à extraire et à analyser ne seront plus seulement disponibles sous forme de tableaux de chiffres, mais également sous forme de documents textuels, d'images, de vidéos ou d'enregistrements de paroles et discours, voire de musiques ou de sons. Dès lors, c'est le traitement simultané et synchrone de ces informations de nature et de type multimodal [1] qui doit prévaloir, plutôt que leurs traitements séparés et diachroniques, qui ont abouti jusqu'ici à un traitement juxtaposé de ces informations. Cette situation a créé, de facto, des cultures de spécialisation tant sur les recherches faites par type de données que pour les traitements associés. De fait, ceci va plus loin en ce sens que cette spécialisation par « silo » a engendré des affectations quasi univoques d'experts aux disciplines : linguistes traitant du texte, spécialistes de l'image et de l'imagerie traitant des images et des vidéos, spécialistes du son et musicologues traitant du son. Le propos sur lequel nous nous orienterons dans ce texte est de promouvoir une approche visant à traiter « à la fois » données texte et données numériques, données parole et données texte, données image et données texte, voire tous ces types en même temps.

En effet, il y a quelques dizaines d'années le combat pour la « **Gestion de l'Information** » afin d'automatiser des processus de gestion des données, répétitifs et coûteux en main d'œuvre, était au cœur des enjeux industriels avec, comme concrétisation effective, l'hégémonie sans partage obtenue au fil des années par les Etats-Unis et les Japonais sur l' « **Informatique** », dévolue désormais à l'Inde et à la Chine.

Le vrai challenge se situe aujourd'hui au niveau, bien supérieur, de l' « **Information en tant que telle** » et dès lors de sa « Maîtrise [2] », en vue de faciliter et d'optimiser les prises de décisions qui en dépendent. Conséquemment et incidemment, la non-sujétion à des logiciels et systèmes étrangers pour l'analyse à valeur ajoutée de cette information, demeure l'une des garanties majeures de notre Indépendance Nationale.

Maîtriser l'Information signifie beaucoup de choses, parfois contradictoires : naviguer dans le « trop plein » de données, filtrer le nécessaire à la compréhension de situations, corréler des informations provenant de sources ou de capteurs différents, en étalonner la pertinence, fusion-

[1] **Multimédia** est l'adjectif qui caractérise les Sources (Télévision, Journaux et livres, Base Iconographique, Mails sur Internet, Pages Web, Enregistrements de paroles, CD et déchargements de musiques etc..), **Multimodal** est celui caractérisant les Types (Textes, Images, Vidéos, Paroles, Sons et Musiques, Données de signal, Données Numériques et Enregistrements dans les bases de données). Quand il n'y a aucune ambiguïté les deux adjectifs sont remplaçables l'un par l'autre.

[2] *« **Maîtrise** » étant justement pris ici au sens d'« analyse avancée et profonde »*

REVUE DE L'ELECTRICITE ET DE L'ELECTRONIQUE

N° et date de parution : 72 - 01/02/2007
Diffusion : 3000
Périodicité : Mensuel
LAREVUEDELE_72_81_310.pdf
Page : 82
Taille : 95 %

ner avant analyse et synthétiser après, ou l'inverse, mesurer la portée des interprétations etc... le champ des possibles est immense.

Le changement de paradigme qui crée ce besoin nouveau de maîtrise de l'Information est résumé dans les quatre concepts suivants : *Quantité, Multiplicité-Hétérogénéité, Rapidité-Simultanéité et enfin Sécurité :*

• **Quantité :** en effet, de nos jours, on parle de Giga-octets voire de Tera-octets disponibles alors qu'il y a 10 ans on parlait encore de Mega-octets, les flux d'informations transitant par Internet ont une croissance exponentielle, les pages du Web se comptent en milliards, les bases de données disponibles « Datawarehouses » se mesurent en centaines de Giga-octets pour une Entreprise de taille moyenne et en Tera-octets pour les plus grandes etc..

• **Multiplicité et hétérogénéité :** aujourd'hui, il ne s'agit plus seulement de gérer des données qui se présentaient sous forme structurée dans des fichiers plats ou dans de petites bases de données, mais bien de traiter et d'interpréter, en plus d'informations structurées, émanant de bases énormes, des mails, des dépêches ou des textes sous forma` électroniques arrivant en flux croissants, des images, des cartes et des représentations graphiques de toutes sortes et de plus en plus complexes, voire des enregistrements audio ou vidéo. Cette multiplicité des sources dans des formats de plus en plus différents sous forme de données de nature très disparate est la caractéristique claire du champ d'application auquel doit s'adapter cette maîtrise de l'information.

• **Rapidité et Simultanéité :** de plus la composante temporelle est un facteur clef de la maîtr´e de l'information, aujourd'hui. Qu'elle soit tactique ou stratégique, la prise de décision se doit d'être extrêmement rapide. Quasi-temps réel pour l'aspect tactique, en heures pour l'aspect stratégique. Or, analyser en séquence les informations qui arrivent, par des media différents, ce que l'on pratique couramment aujourd'hui, prend beaucoup de temps et est générateur d'erreurs, d'où l'importance fondamentale de la notion de **simultanéité** de cette Analyse à valeur ajoutée. C'est en fait le vrai changement qui se prépare. Cette rapidité et cette simultanéité étant d'autant plus importantes que les menaces et les risques, dans certains contextes, sont de nature asymétrique et peu susceptibles d'être modélisés facilement ou de rentrer dans des formalismes doctrinaires ou dans des processus de traitement suffisamment exhaustifs.

• **Sécurité :** un facteur à ne pas négliger et de plus en plus nécessaire dans le contexte actuel (domaines : militaire, économique, santé etc...) est l'obligation faite de protéger les données et surtout leurs analyses et les résultats de ces analyses. En effet chaque information (ou donnée) a une valeur intrinsèque faible si elle reste isolée, sa confrontation et sa comparaison à d'autres, quelles que soient leurs natures lui permettra de prendre de la valeur (« asset value ») au fur et à mesure de la profondeur du processus d'analyse. C'est donc bien les résultats qu'il convient de protéger au maximum, même si protéger les données est déjà un plus. Il faut donc que des processus de protection des flux lors de transferts et de protection des niveaux d'autorisation pour ceux qui doivent en connaître soient élaborés préalablement ou au minimum de façon concomitante à ces analyses. C'est une nouvelle nécessité d'aujourd'hui qu'il convient d'intégrer le plus vite possible.

L'enjeu des Nouvelles Technologies de l'Analyse de l'information (NTAI) est donc désormais de pouvoir croiser, au sens analytique du terme, des données de toutes sortes et de pouvoir les analyser conjointement en allant bien au-delà de leur simple juxtaposition, avec l'espoir que les unes affineront et désambiguïseront l'analyse des autres et réciproquement. C'est aussi avoir beaucoup plus de matière pertinente pour appliquer par la suite les processus décisionnels et les principes d'aide à la décision.

De façon synthétique, les NTAI correspondent à l'ensemble des technologies permettant d'analyser de très grandes quantités de données, de natures, de formats, de provenances et de types très différents de façon très rapide et en simultanéité, dans un environnement sécurisé.

Quelques enjeux de l'analyse avancée de l'information au service de la sécurité

Christian Malis

Il existe actuellement dans le monde environ 40 millions de caméras de surveillance (chiffre en progression rapide), 1 caméra sur 7 étant semble-t-il installée dans un lieu public. On compte chaque jour 180 milliards de secondes de communications téléphoniques. La NSA surveille quotidiennement 5 millions de messages de tous types (internet, téléphone, etc...). La veille sanitaire à des fins de prévention des pandémies suppose la surveillance permanente de la presse quotidienne régionale mondiale et son croisement avec les milliers d´informations remontant 24H sur 24 des services d'urgence. Sur une zone de crise géopolitique, un analyste de renseignement peut recevoir plusieurs milliers de messages par jour, contenant des informations multimédias…

Ces quelques chiffres, qui donnent le vertige, permettent de prendre la mesure des besoins en moyens performants d'exploitation de l'information au service

REVUE DE L'ELECTRICITE ET DE L'ELECTRONIQUE

N° et date de parution : 72 - 01/02/2007
Diffusion : 3000
Périodicité : Mensuel
LAREVUEDELE_72_81_310.pdf
Page : 83
Taille : 95 %

de la Sécurité. Les domaines sont nombreux : contre-terrorisme, interventions militaires extérieures, sécurité sanitaire, etc. Dans ce contexte, les méthodes et solutions de détection de risques et de gestion de crise se développent considérablement en incorporant de plus en plus de techniques d'analyse et de fusion d'information.

Le sujet étant vaste, on donnera surtout des éléments sur les besoins relevant de la Sécurité Nationale. Dans ce dernier domaine, les autorités expriment de plus en plus de besoins précis et s'efforcent d'orienter les efforts de R&D. C'est tout particulièrement vrai de pays en pointe dans le renseignement et la lutte anti-terroriste comme les Etats-Unis, la France, la Grande-Bretagne. De manière générale, la préoccupation est de pouvoir extraire des éléments de connaissance synthétiques à partir de la masse complète d'informations soumises aux systèmes d'information et de décision. Consultons pour la France le récent *Livre Blanc du Gouvernement sur la sécurité intérieure face au terrorisme* [3]. Dans le cadre de la « bataille technologique » ont été identifiés deux besoins majeurs : le contrôle des flux de communication et la surveillance par moyens vidéo (la vidéosurveillance dans les lieux et les moyens de transport publics se généralisant). L'automatisation des tâches de détection devra nécessairement comporter un module destiné à la parole, permettant à terme la reconnaissance du locuteur, celle de la langue employée, la transcription automatique des conversations et l'extraction des faits marquants du discours, voire leur traduction automatique. Dans le domaine des recherches dans les grands entrepôts de données multimédias, les efforts devront porter principalement sur les capacités de tri sémantique, et sur la reconnaissance d'objets et de personnes dans un flux continu de photographies et de vidéos. Par ailleurs le développement de la vidéosurveillance ne peut se concevoir sans l'introduction de logiciels experts permettant la reconnaissance faciale, la détection de mouvements, la détection d'objets abandonnés ou le suivi de personnes. Seuls ces logiciels sont en effet à même de permettre une exploitation rapide et efficace de la masse d'informations recueillies. Franchissons l'Atlantique. Depuis septembre 2001, de nombreux programmes de fouilles massives de données à des fins de contre-terrorisme ont été lancés, sous l'égide des autorités en charge de la Sécurité intérieure (DHS, *Department of Homeland Security*, HSARPA, *Home-land Security Advanced Research Project Agency*, et même DARPA, *Defense Advanced Research Project Agency*). Selon des estimations de 2003, 1,1 milliards de $ étaient affectés annuellement à la R&D dans ce domaine. Le département spécialisé de la DARPA met l'accent, pour les besoins strictement militaires, sur la fusion d'informations multimodales (texte/ images/ son/ données). Déjà dans le domaine purement civil le « data mining » permet aux organismes financiers de réduire les fraudes à la carte de crédit en analysant les transactions en temps réel. Adaptés à la grande distribution, ces logiciels sont capables, après avoir analysé le comportement des consommateurs à travers leurs achats, de suggérer les grandes lignes d'une campagne marketing. Les programmes de data mining massif veulent transposer ces méthodes pour les besoins de la sécurité intérieure : il s'agit d'utiliser les données collectées sur des millions de personnes vivant aux Etats-Unis, pour tenter, par un croisement des fichiers gouvernementaux et privés, de débusquer d'éventuels terroristes.

Ces approches, typiquement américaines, suscitent à vrai dire des critiques : d'une part concernant l'atteinte potentielle aux libertés privées, ensuite relativement à leur efficacité - la recherche « en aveugle » sur des monceaux de données risque d'engendrer des masses ingérables de fausses alarmes. Mais, au-delà des vicissitudes entourant son développement, il est tout à fait probable que va se renforcer et se structurer un domaine du « data mining de sécurité nationale et de défense », encadré par des règles juridiques et méthodologiques strictes. Peut-être un jour le métier de « Data-interprète », comme il y a aujourd'hui des « photo-interprètes » pour l'analyse d'images, devra-t-il être créé.

L'enjeu est également économique. Comme pour la plupart des activités considérées comme stratégiques outre-Atlantique (l'Espace par exemple), l'offensive américaine est puissante et marche « sur ses deux jambes » : gouvernementale et civile. Une action gouvernementale mobilisant des investissements de R&D substantiels complète la conquête des premiers marchés effectués par des acteurs privés comme les géants du Business Intelligence pour le traitement d'informations structurées, ou plus récemment Google et Yahoo pour l'information non structurée sur la Toile. On a ainsi affaire à une véritable « machine de guerre » contre laquelle il faut lutter si l'on veut conserver un minimum de souveraineté.

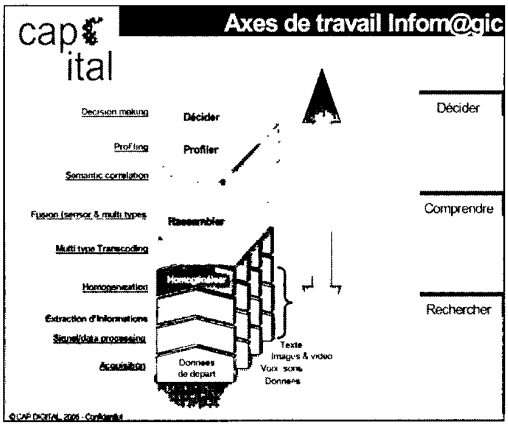
Mais en Europe, à l'initiative de la France, des propositions solides et structurées ont vu le jour. Dans le domaine de la R&T, la plus notable est Infom@gic, projet porté par le Pôle de compétitivité à vocation mondiale Cap Digital. Ce projet, conduit par Thales avec le concours d'EADS et de XEROX, de nombreuses PME innovantes (Pertimm, Vecsys, Temis, Sinequa etc.) ainsi que de laboratoires publics de pointe (LIP6, ENST, LIMSI, LIPN etc.) ambitionne des percées dans

[3] 2006

REVUE DE L'ELECTRICITE ET DE L'ELECTRONIQUE

N° et date de parution : 72 - 01/02/2007
Diffusion : 3000
Périodicité : Mensuel
LAREVUEDELE_72_81_310.pdf
Page : 84
Taille : 95 %

les trois domaines fondamentaux de l'ingénierie de la connaissance que sont la recherche d'informations numérisées, l'extraction de connaissances, la fusion multimodale et l'aide à la décision. Une plate-forme d'intégration technologique révolutionnaire, qui reprend un standard « open source » (UIMA d'IBM), doit en être le support logiciel et permettre l'élaboration de futures normes… d'origine européenne cette fois.



Infom@gic factuellement

- **Budget / Durée : 56 M€ / 3 ans (2005-2008)**
- **25 partenaires**
- **Schéma directeur du projet**
  - 2006 : Maquettage
  - 2007 : Démonstrateurs
  - 2008 : Plate-forme technologique, briques logicielles pré-industrielles, démonstrateurs applicatifs
- **Répartition**
  - Grands groupes, grands instituts : 58%
  - PME : 21%
  - Instituts, universités :21%

2. Typologie des Problématiques

2.1. Fusion « Physique » et Fusion « Analytique » des Données

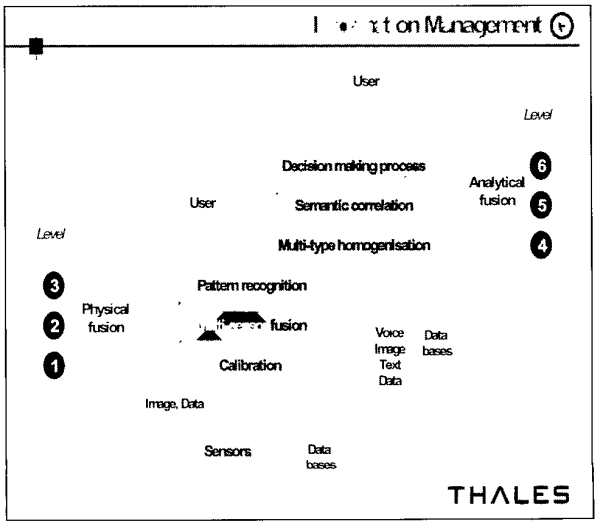
Comme nous venons de le voir, c'est dans notre capacité à intégrer les quatre concepts de quantité, multiplicité, vitesse et sécurité que nous pourrons adresser la problématique générale de l'Analyse approfondie de l'Information. Or cette problématique n'est pas uniforme, il existe, en effet, deux types d'approches de l'analyse et de la fusion de l'information à la fois séparées et néanmoins voisines quant aux buts finaux :

Approche I : Une approche très « **physique** » orientée capteurs, temps réel et vision tactique, et opérationnelle (redressement d'image, fusion de signal, association pistes /cibles, détection et désambiguïsation de signatures radars, sonars, capteurs images, caméras infra-rouge, images satellites etc...), identification et authentification de cibles, de la voix, de l'empreinte digitale etc... Les données dans ce cas sont codables en formats numériques purs (signaux/bruits, pixels, luminance, formes, fréquences, intensité, etc...). Certaines des méthodes citées précédemment et présentées dans les numéros 8 et 9 de la REE et plus particulièrement les articles de Didier Dubois et Henri Prade sur la Théorie des Possibilités, la Théorie des fonctions de Croyance s'inscrivent parfaitement dans ce contexte de « Fusion Physique ».

Approche II : Une approche plus « **Analytique** » orientée vision stratégique, moins temps réel que la précédente, mais fondée sur la gestion et l'analyse de grands volumes de données en stock ou en flux, capable d'intégrer et de fusionner selon des approches de codages « pivots » des informations à caractère plus sémantique et surtout de natures et types très différents : Numériques, Textuelles, Paroles, Images, Vidéos, Cartographies etc..

Nous nous proposons de traiter essentiellement dans cet article de la deuxième approche, celle que nous avons appelée « Analytique », par opposition à l'approche dite « Physique », couverte davantage par les deux numéros de la revue REE de septembre et d'octobre, même si certaines techniques et méthodologies présentées (comme le support à vaste marge et les réseaux de neurones peuvent trouver également leur place dans le cas de la « Fusion Analytique »).

D'un point de vue global la situation peut se présenter de la façon suivante, (document interne Thales) :



Les niveaux 1,2,3, relèvent de ***l'Approche I****, les niveaux 4, 5, 6 relèvent de* ***l'Approche II***

REVUE DE L'ELECTRICITE ET DE L'ELECTRONIQUE

N° et date de parution : 72 - 01/02/2007
Diffusion : 3000
Périodicité : Mensuel
LAREVUEDELE_72_81_310.pdf
Page : 85
Taille : 95 %

Pour l'Approche I, la situation peut être succinctement décrite selon le schéma suivant :

• Dans la **couche 1**, par exemple, il existe une forte tradition de compétences françaises, nos grands industriels français (Thales, SAFRAN, EADS etc..) ont fourni des capteurs qui couvrent une bonne partie du spectre offert par nos concurrents Anglo-saxons. Les investissements se poursuivent, les plans de développements prévus sont en accord avec les tendances technologiques du niveau international, et seules les contraintes sur le montant des investissements en jeu, nous obligerons à faire des choix, nous empêchant de ce fait d'avoir une couverture totalement exhaustive, des « possibles ». La problématique est moins technique et technologique que financière.

• Dans la **couche 2**, nous avons des chercheurs qui rivalisent avec les chercheurs américains et qui ont même, dans certains cas, amélioré leurs travaux (Dubois, Prade, Smets améliorant les travaux de Dempster-Shafer ou de L. Zadeh, par exemple). Ce niveau est plus scientifique et technologique que technique.

• Mais nous sommes vulnérables sur la **couche 3**, car même si nous ne sommes pas distancés dans la recherche théorique, l'aspect développement de composantes technologiques a été engagé sur une base volontariste par les Américains et nous risquons des dérives graves au niveau de la capitalisation intellectuelle.

Pour l'Approche II, la situation est plus mitigée :

• En ce qui concerne la **couche 5**, l'enjeu qui sera décrit en profondeur ci-après est crucial, nous avons des atouts méthodologiques (à base mathématique et linguistique) de très haut niveau : nous précédons les Américains au niveau méthodologique et théorique dans certains des pans scientifiques concernés, tradition française en mathématiques et linguistique oblige. Néanmoins le montant très important des investissements déployés par les USA sur ces domaines leur permettra, sans doute, de compenser leur léger retard théorique (sur certains points particuliers) par une capacité technologique accrue. D'autre part et ceci est bon à savoir, un certain nombre de leurs « gourous » dans leurs centres de recherche viennent de France. Ils en ont à IBM Research Yorktown, à ORACLE Research, Monterey notamment. La NSA, la CIA et le « Homeland Security Dept » bénéficient de leurs travaux et percées. Or l'enjeu ici est fondamental, quiconque aura la **primauté sur cette couche 5 et sur la couche 2** atteindra la vraie « Maîtrise de l'information ». Ce que nous tenterons de prouver par la suite.

• Pour la **couche 4**, français et anglo-saxons ont pris des voies différentes. Brièvement, les anglo-saxons homogénéisent l'hétérogène en privilégiant une approche « quantitative » des données, quelles qu'elles soient, les Français une approche « qualitative », il n'y a pas de dominance stricte d'une approche par rapport à une autre. Nous sommes seulement dans l'expectative réciproque. Mais il faut être vigilant à tout changement de leurs habitudes en ce domaine.

• Pour la **couche 6**, même si les anglo-saxons possèdent une très abondante littérature, considérablement plus volumineuse que la notre dans les domaines relatifs à ce niveau, et également une pléthore de laboratoires tant industriels (ATT, IBM, Oracle, SABRE Tech, Microsoft etc..) qu'universitaires (MIT, Princeton, Berkeley, Stanford, Imperial College...), leur domination réelle n'est que statistique car ils possèdent pratiquement toutes les revues scientifiques de ce domaine (ceci est d'ailleurs un point fondamental : **nous n'avons plus les d'bouchés de notre savoir et faire savoir** plus de 96 % des revues scientifiques de bon niveau sont anglo-saxonnes et contrôlées par eux, ceci expliquant parfois cela). Nos élites sont néanmoins compétentes et elles satisfont le niveau requis par la couche 6 : « Aide à la Décision et systèmes dérivés » quant aux besoins réels dans l'Analyse Avancée et Profonde de l'Information.

2.2. Les Processus de Fusion Analytique comme enjeu fondamental

Fort des constats établis par des cabinets d'audit et d'analyse conjoncturelle (Gartner Group, IDC, OVUM, etc, ...) que :

• 80 % des informations disponibles dans les entreprises sont de nature non structu ée (textes, mails, enregistrements audio),

• le trafic sur le Web est en croissance exponentielle dans les 1000 plus grandes entreprises internationales,

• la quantité de textes, d'articles, de notes, de mails, de dossiers, etc..., publiés, édités, ou stockés au sein des entreprises durant les cinq prochaines années représentera l'équivalent de ce qui aura été publié et édité durant les vingt années précédentes, ou comme l'attestent certains spécialistes de la DGA , il y a doublement de ce type d'informations tous les 18 mois, il est important de comprendre qu'il est **fondamental** d'exploiter ce capital informationnel. Il faut le faire en supplément et complément de ce qui est communément analysé à des fins décisionnelles aujourd'hui à partir des structures de « datawarehouses » et de bases de données structurées existantes. Si l'on rajoute à ceci l'ensemble des interceptions téléphoniques, les enregistrements audio, les enregistrements vidéo, les images photographiques et satellitaires, ce sont des masses énormes qu'il faut structurer, stocker et surtout analyser en mode complémentaire et simultané.

2.2.1. Caractérisation des Processus de « Fusion Analytique »

Si l'on veut caractériser la notion de « fusion analytique », il convient d'en définir au préalable les contours. La fusion analytique associe les termes : « fusion » et

REVUE DE L'ELECTRICITE ET DE L'ELECTRONIQUE

N° et date de parution : 72 - 01/02/2007
Diffusion : 3000
Périodicité : Mensuel
LAREVUEDELE_72_81_310.pdf
Page : 86
Taille : 95 %

« analyse » ce qui revient à dire qu'on va pratiquer d'abord des agrégations et des fusions de données de nature disparate et hétérogène avant d'en effectuer l'analyse groupée et simultanée. A ce niveau deux stratégies ou processus existent, qui essaient de rendre compatibles des données de natures différentes en pratiquant une conversion (transcodage) dans un espace de représentation unique. On appellera cet espace de compatibilité : « Espace Pivot » . Deux espaces sont pour l'instant considérés comme candidats : l'espace pivot du « tout numérique » et l'espace pivot du « tout textuel ».

- La première stratégie consiste donc à transformer toutes les données (qu'elles soient textuelles, images, paroles, signal, et numériques de base) en des **représentations vectorielles numériques**.
- La seconde stratégie consiste à transformer toutes les données en une représentation en scripts textuels (s'ils préexistent) ou en **réécritures sous forme de descriptions littérales** si nécessaire.

- En particulier, si l'approche « **tout numérique** » est intéressante par sa capacité à traiter de grandes masses de données ou d'informations, elle est souvent moins riche au niveau de la « **sémantique de l'analyse** », car elle repose presque uniquement sur des principes de corrélations ou d'implications statistiques, qui outre le fait qu'ils sont dépendants des mesures de corrélation ou d'implication choisis, ne présentent que des visions principalement quantitatives des faits et reposent assez peu sur l'aspect qualitatif descriptif ou explicatif de certaines situations. Par ailleurs, les processus de transcodage ou de conversion associés à cette première approche ne sont pas toujours simples. Selon le mode ou le type de données concernées, on peut ainsi passer du trivial à l'extrêmement complexe. Ainsi, les enregistrements numériques présents dans les bases de données types : « datamarts » ou « datawarehouses » sont déjà sous le **bon format** et aucun transcodage n'est nécessaire, il en va de même pour les enregistrements sous forme de signaux numériques (paroles et sons par exemple). C'est un peu plus complexe pour les images ou photos, car on doit les pré-analyser au travers de leur distribution pixellaire et ne considérer que leurs valeurs caractéristiques vectorielles d'extraction de contours ou de formants. En revanche, il en va tout autrement si l'on veut passer du texte à une vectorisation numérique (même si ceci se pratique déjà, voir les paragraphes ultérieurs), ce n'est pas un processus de transcodage trivial il est même parfois complexe et nécessite des traitements et des ressources préalablement élaborés.

- L'approche « **tout textuel** », présente quasiment des avantages et des inconvénients qui sont les symétriques « duaux » de ceux de l'approche « tout numérique ». En effet si l'on peut compter sur un **bien meilleur pouvoir « sémantique »** de l'analyse, car le texte permet des extractions de sens et des mises en perspectives explicatives bien meilleures que les seules données numériques, en revanche, il ne permettra pas des traitements sur des masses énormes de données (sauf en mode « moteur de recherche plein texte » qui n'est pas notre propos car nous nous intéressons à **l'analyse élaborée** de l'information et non à sa **recherche en surface** qui est un autre sujet). Par ailleurs, comme dans le cas précédent, le transcodage d'un type de données vers une représentation textuelle s'échelonne du « trivial » à de « l'hyper complexe ». En effet le « Texte déjà formaté » ne nécessite **aucune modification**, le transfert de la « Parole » vers le « Texte » nécessite, en revanche, un transcodage, mais il existe déjà, tant au LIMSI (Orsay) qu'à IBM Hawthorne, par exemple, des processus de transfert qui marchent aujourd'hui de façon très satisfaisante pour supporter certains types de fusion analytique. Au niveau du passage de l'« Image » vers le « Texte », les tentatives et les pistes de travail qui se situent plutôt au niveau de la recherche sur l'extraction automatique d'interprétations textuelles et de scripts générés automatiquement, n'en sont qu'à leur début (des travaux menés à l'Inria, ENST, INT, montrent que ce domaine s'anime). Reste un dernier problème lui aussi très difficile, le passage de « Données Numériques » à leur explicitation sous « forme littérale ». Ce cas peut avoir une relativement bonne solution dans certaines situations où les bases de données structurées correspondent à un type voisin du suivant :

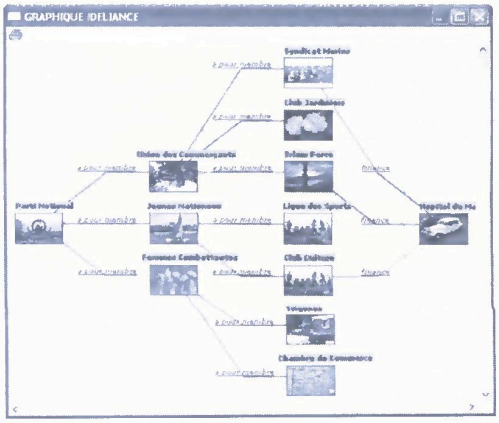
Noms	*Est agé*	*Habite*	*Gagne*	*Appartient*	*Possède*
Dubois	42	Paris	4500€	UMP	FORD
Dupond	35	Lyon	3500€	PS	VW
Dupré	41	Toulon	4800€	UDF	BMW
Durand	54	Nantes	3700€	VERTS	Renault

(Où chaque colonne « variable » du tableau correspond à un « verbe », chaque ligne correspondant à un « individu sujet», et où chaque modalité des variables correspond à une valeur de « complément » d'une phrase virtuelle) :

Phrase : sujet »=>« verbe »=> « complément »

Chaque valeur d'une cellule de la base peut si l'on connaît le numéro de la ligne et de la colonne correspondants être caractérisée par une phrase « trinômiale » simple du type :

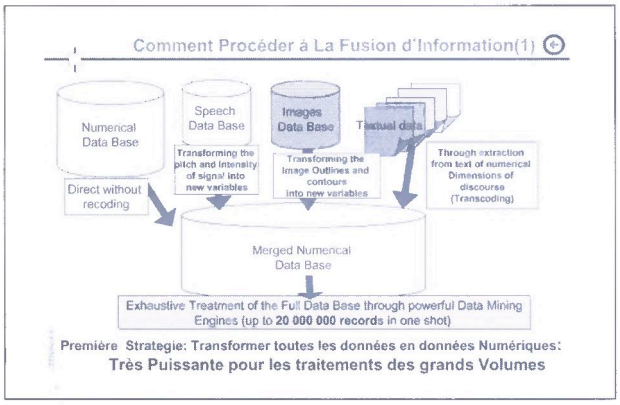
Durand « habite » à « Nantes »

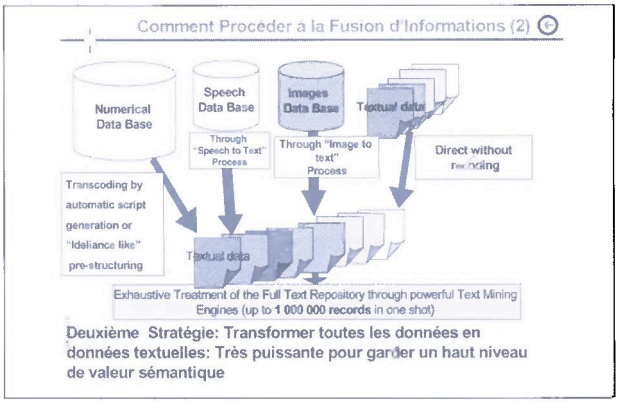


REVUE DE L'ELECTRICITE ET DE L'ELECTRONIQUE

N° et date de parution : 72 - 01/02/2007
Diffusion : 3000
Périodicité : Mensuel
LAREVUEDELE_72_81_310.pdf
Page : 87
Taille : 100 %

De cette approche qui nécessite M phrases, (M= nombre de colonnes) pour caractériser un individu, et N.M pour caractériser l'ensemble du tableau, on peut déduire des visualisations type réseau sémantique (approches Ideliance ou Analyse Notebook). (Voir figure jointe représentant un réseau Ideliance) :





En tout état de cause, on peut déjà pratiquer un certain nombre de tels transferts et les résultats obtenus sont par définition beaucoup plus interprétables que dans le cas de recours au « tout numérique » seul.

A titre d'illustration, nous avons donné ci-dessus les deux configurations schématisées du « tout numérique » et du « tout textuel » :

REVUE DE L'ELECTRICITE ET DE L'ELECTRONIQUE

N° et date de parution : 72 - 01/02/2007
Diffusion : 3000
Périodicité : Mensuel
LAREVUEDELE_72_81_310.pdf
Page : 88
Taille : 95 %

Les applications de cette nouvelle approche vont simplement révolutionner le futur de l'informatique en général. Par exemple, si l'on regarde les moteurs de recherche Internet actuels, on voit que plus l'on tape de mots, moins l'on a de réponses intéressantes, alors que c'est cognitivement le contraire: plus on parle à une personne, plus elle comprend et peut répondre intelligemment. Tous les moteurs de recherche actuels fonctionnent donc dans le cadre d'une informatique déjà dépassée, alors que cette nouvelle approche de la recherche automatisée permet un saut quantique technologique conduisant à un fonctionnement plus humain d'un moteur de recherche. L'implémentation de cette approche sera nécessairement liée à une révolution technologique et donc sociale : on pourra parler à un ordinateur qui "comprendra". La seconde application plus lointaine, mais déjà partiellement présente est de donner à l'ordinateur une certaine autonomie tant visuelle que décisionnelle. Ceci est une simple conséquence de l'implémentation d'un fonctionnement plus humain de la machine.

Ainsi et pour conclure, l'ordinateur du futur sera bien plus proche de l'humain qu'il ne l'est maintenant et les prémisses fondamentaux technologiques sont déjà en cours de développement, dans certains Pôles de Compétitivité comme CapDital (projet Infom@gic notamment).

3. Détails sur les processus de Fusion et de Tranco·age

3.1. La Fusion « Données Textuelles/ Données Numériques Structurées »

A titre illustratif de ce processus global de fusion analytique, attardons nous quelque peu sur le cas de la « fusion analytique » données structurées numériques => données non structurées textuelles, qui est déjà à l'ordre du jour de recherches en cours aux Etats-Unis (IBM par exemple, NSA, DARPA etc..) et en France, et dont le caractère crucial n'est plus mis en doute.

Nous entrons de ce fait dans l'ère de l'Analyse de l'Information **de très grands ensembles de données** (tant structurées que non structurées), et qui plus est, nous souhaitons et souhaiterons pouvoir tirer **simultanément** de ces très gros ensembles une « quintessence » de connaissances facile à utiliser et apte à faciliter la prise de décision.

La fusion ou traitement conjoint de données de type structuré (données numériques) et de données non structurées (données textuelles) doit être comprise ici au sens de l'**apport analytique réciproque**.

En d'autres termes, comment e·traire de données purement textuelles (textes libres) des informations :

- Soit déjà numériques, valeurs contenues dans des textes, qu'il faudra extraire de leur contexte
- Soit à numériser des dimensions analysables (satisfaction, gravité d'une situation, rareté ou dangerosité d'un événement, sentiments exprimés, thématiques du discours, etc..) extraites du contenu textuel, afin de pouvoir corréler ces nouvelles données numérisées avec des données numériques déjà collectées
- Soit inversement comment interpréter des séquences de textes ou de messages en corroborant les informations contenues dans ces textes avec des informations collatérales sur les mêmes sujets mais exprimées en valeurs numériques.

Le problème est ardu si on le comprend dans son sens strict d' « analyse conjointe » et **non d'une simple juxtaposition de données** numériques et de données textuelles, ce que certains nomment également « fusion » et que l'on trouvera sans difficulté apparente, proposée par les éditeurs de bases de données les plus connus, par exemple Oracle 9i ou IBM DB2 Text Extender.

Le fait que ce problème soit difficile tient, en particulier, au caractère de non compatibilité (au sens analytique du terme) des données numériques et des données textuelles. C'est la raison pour laquelle, la **mise en compatibilité des données est le point crucial du processus de fusion analytique**. Cette mise en compatibilité, consiste à ramener toutes les informations (numériques et textuelles dans ce cas) à un format numérique commun (on parlera alors de « transcodage »).

Les applications de cette approche globale de fusion texte/data (dans le sens texte vers données numériques), couvrent, dans le civil, les domaines du « Business Intelligence », de la « Gestion de la Relation Clients » et bien entendu de la Veille Economique et Stratégique (le « Competitive Intelligence » des anglo-saxons). La fusion Analytique s'applique aussi, de façon évidente et **tout particulièrement** au domaine Militaire et à la Sécurité du Territoire ainsi qu'à la « Sécurité Civile » d'un point de vue gouvernemental. Le Renseignement et la Surveillance ainsi que tout item mis par les anglo-saxons dans le terme « Homeland Security » font partie intégrante de ce domaine d'application.

La distinction entre données structurées et données non structurées, si elle est nette aujourd'hui (les compétences permettant d'analyser les premières venant de la Statistique ou de l'Informatique Décisionnelle, celles permettant d'analyser les secondes venant de la Linguistique Computationnelle) sera de moins en moins évidente lorsque toutes les sources d'information associeront de plus en plus ces différentes formes et structures dans des bases uniques ou tout du moins unifiées. Au niveau entreprises, rares sont celles qui traitent les deux sujets à la fois, car

REVUE DE L'ELECTRICITE ET DE L'ELECTRONIQUE

N° et date de parution : 72 - 01/02/2007
Diffusion : 3000
Périodicité : Mensuel
LAREVUEDELE_72_81_310.pdf
Page : 89
Taille : 90 %

les compétences requises viennent de cursus très différenciés à l'origine. Ainsi des sociétés travaillant sur l'ingénierie textuelle comme [VERITY-AUTONOMY, INKTOMY, OPENTEXT, CONVERA, NSTEIN, etc.., (côté Anglo-saxon) ou ARISEM, SINEQUA, EXALEAD, PERTIMM, TEMIS, DATOPS, SYNAPSE etc.. (Côté français)] ne disposent pas, sauf exception, de cette double compétence et filiation « Linguistique » et « Mathématique », permettant d'aborder la double problématique en simultané.

A ce propos, des entreprises américaines fort connues, spécialisées en systèmes décisionnels (basés sur des données structurées) ont voulu jouer sur les deux tableaux (telles : SAS 2,6 B$ de revenu en 2004, un grand leader mondial du Décisionnel ou SPSS à une échelle moindre). Elles ont largement échoué dans leur tentative d'élargissement vers le domaine textuel, la première n'ayant pas les compétences pour vendre et promouvoir son SAS Text Miner, et SPSS qui a racheté Lexiquest en France avec ses réseaux tissés historiquement en France et en Europe a pris la mauvaise décision de rapatrier les développements aux USA, laissant l'entité pensante et compétente en France privée de ses racines ; bilan plutôt mitigé.

Seules de grandes entreprises comme IBM, Microsoft, ORACLE, Raytheon, EDS, XEROX etc... aux Etats-Unis, alliées bien entendu à des éditeurs ou sociétés spécialisées ou THALES, EADS, France Telecom en France et dans une moindre mesure des alliances solides entre sociétés complémentaires de type PME spécialisées, peuvent relever le défi, par leur capacité à pérenniser les investissements

Des analyses récentes, conduites dans le domaine du Marketing, montrent que la « valeur » (en d'autres termes, le pouvoir discriminant) d'informations provenant directement des clients (lettres de réclamation, appels téléphoniques, mails ou commentaires sur Internet) est nettement plus importante pour résoudre des problèmes d'attrition, de fidélisation et de satisfaction de la clientèle que les données structurées, traditionnellement stockées et historiées dans les bases de données marketing ou dans les Datamarts et Datawarehouses de l'entreprise. Ceci est dû au caractère intrinsèque de récence, de pouvoir explicatif sémantique fort, de liberté d'expression et de pertinence de l'information, brute, directe sans transformation et sans intermédiaire, caractéristique du support textuel.

Les données structurées, a contrario, étant souvent issues de modèles de données imposés a priori et de processus normatifs simplificateurs.

Ce besoin de fusion des sources en des bases de données unifiées et uniques où données structurées et non structurées s'associeront, donnera lieu à des processus d'analyse nouveaux et non encore réellement pratiqués aujourd'hui sur une grande échelle, donnant de ce fait un espace nouveau à l'analyse en profondeur.

« Ainsi de façon très concrète, une grande banque de la Région Parisienne a utilisé cette approche pour comprendre pourquoi un nombre important de clients était en train de la quitter : elle a donc rendu compatibles les données textuelles émanant de discours clients (mails, courriers etc...) avec les données numériques de sa base de données Marketing clientèle. Elle a ainsi, pu mettre en évidence, des relations de cause à effet entre données comportementales et données venant d'opinions exprimées, qui grâce à leur caractère hautement explicatif et leur récence, ont apporté un « plus » fondamental en matière d'interprétation. Ainsi a t-elle trouvé (entre autre) que les clients qui s'exprimaient le plus négativement vis à vis de cette banque étaient (en corrélation) plus âgés que la moyenne mais que ceux qui prenaient réellement des décisions de rétorsion (rupture de contrat, cessation de relation, etc...) avaient un âge compris entre 43 et 48 ans et que leur statut marital les positionnaient majoritairement comme des clients générant un revenu important pour la banque, et célibataires… bilan et conclusion du processus décisionnel déduit par la banque : mise en place d'un suivi exhaustif et direct de l'ensemble des clients générant un revenu important pour la banque, âgés de 40 à 50 ans et célibataires… bilan : diminution de 27 % des départs sur les 6 mois suivant la mise en place du processus… »

A titre d'exemples nous donnons, ci joint, une liste de nouvelles possibilités offertes par cette approche dans le domaine du « Business Intelligence » et de la « Gestion de la Relation Client » :

- Recherche de corrélation entre l'âge, le niveau de revenus, le niveau d'éducation etc... (données structurées) du client et le niveau de mécontentement ou d'alerte contenu dans son discours (données sémantiques extraites du texte : mécontentement, insatisfaction, jugements, inquiétudes etc....), (voir exemple ci-dessus). Il faut rappeler qu'un bon client perdu est générateur de coûts internes extrêmement importants pour le remplacer par un nouvel entrant du même type. (25 000 € en moyenne par client perdu)
- Recherche d'une **liaison forte** entre le comportement acheteur (ventes croisées) de clients et leur niveau de satisfaction exprimée,
- Recherche d'une **corrélation** entre l'extension potentielle de contrats d'assurance de clients et la multiplicité des questions qu'il pose au travers de mails sur les couvertures de certains produits,
- Recherche d'une **liaison** entre la diminution du chiffre d'affaire de certains secteurs d'une entreprise et le niveau de mécontentement des clients vis-à-vis de certaines agences ou certains représentants commerciaux,
- Recherche d'une **liaison** entre la citation systématique de certains concurrents dans les courriers ou mails clients et le fait que ces derniers sont en train d'étudier des processus de transfert (comptes, contrats etc..) vers le concurrent cité dans leurs mails ou lettres.
- Etc...

REVUE DE L'ELECTRICITE ET DE L'ELECTRONIQUE

N° et date de parution : 72 - 01/02/2007
Diffusion : 3000
Périodicité : Mensuel
LAREVUEDELE_72_81_310.pdf
Page : 90
Taille : 95 %

Toutes ces corrélations, ces liaisons, ces implications permettent de façon extrêmement efficace d'améliorer les prédictions et les scores de clients en **mode anticipatif** lors de problématiques d'attrition, de rupture de contrats, de propensions d'achats, de ventes croisées, de fidélisation, de « churn » en téléphonie, etc... (c'est-à-dire des « micro crises » en matière business)

On peut tout aussi bien étendre ce qui précède au cas de la gestion de la Relation client dans le contexte de l'Extension de Garanties (Incidentologie) chez un constructeur automobile, on pourra se servir de la fusion analytique Data/texte pour :

- Rechercher une **corrélation** entre certains modèles d'une gamme et la remontée d'informations sur la non fiabilité de pièces de ces modèles, indiquée dans des rapports écrits de services après vente.
- Rechercher une **corrélation** entre la marque d'un véhicule, sa puissance ou tout autre caractéristique mesurable et le niveau de satisfaction du client conducteur, exprimée au travers de courriers reçus par le service consommateurs.
- une **relation de cause à effet** (implication statistique) entre la cylindrée d'une voiture et les contenus des rapports et constats échangés entre une compagnie d'assurance et le service « pièces détachées » d'une marque.

En fait tous les secteurs économiques sont justiciables de cette problématique de la fusion Data/texte ainsi :

- La recherche de la corrélation entre la fidélité à une **compagnie aérienne** et les sentiments vis-à-vis de cette compagnie exprimés dans des mails de voyageurs ou des courriers est d'autant plus importante que la ligne desservie est très concurrentielle.
- La recherche de corrélation et d'implications entre propriétés de molécules et discours des consommateurs sur leurs sensations perçues lors d'essais en **cosmétologie** est totalement pertinente (les sensation perçues n'étant pas forcément corrélées aux propriétés physiologiques et chimiques des mélanges moléculaires, il importe de mesure les corrélations à implications positives)
- La recherche d'implications (efficacité de médicaments versus placebo) au travers de relevés et de rapports médicaux de description des sensations exprimées textuellement par des patients tests lors de protocoles et essais cliniques en **médecine ou en pharmacodynamique**, est une piste sur laquelle se lancent certains industriels américains (Pfizer, Monsanto, Eli Lilly…).

Tout ce qui vient d'être dit ici dans le cadre d'applications civiles est tout à fait « dual » de ce qui pourrait et devrait être fait dans le cadre des services de Renseignements (DST, Militaire : DRM, DGSE, GENDARMERIE), des services de surveillance policières (RG, Police Scientifique), de protection du patrimoine de convoitises extérieures (Veille Stratégique et Economique) etc…

Donnons ici deux exemples simples et faciles à comprendre :

- **Dans le domaine de crises militaires extérieures :** Supposons que nous disposions d'un côté d'une base de données « personnalités » dans un contexte de crise « militaire », (base semi-structurée, contenant le(s) nom(s) ou les alias de ces personnalités, plus des champs descriptifs remplis : âge, organisation à laquelle elles appartiennent, l'ethnie, la religion, lieu de séjour majoritaire ou paramètres biométriques si l'on en dispose etc... et que parallèlement l'on dispose d'un moyen d'extraire d'un flux de dépêches OSINT ou de messages plus ciblés et moins ouverts et de rapports terrains, une structuration liant ces personnalités à des d'événements décrits et cités dans les flux textuels ; alors l'association automatique entre les lieux des événements, les dates de ces derniers, les personnalités citées permet assez facilement d'obtenir les relations établies entre elles (remontées de réseaux) et leur impact sur les événements (influences et dangerosité). Le problème est donc bien de « corréler » des données « structurées sur les personnalités » et les données textuelles y afférant.
- **Dans le domaine de Sécurité Intérieure :** Supposons que l'on puisse croiser des bases de données type DGI, GSM, GIE CB (le problème CNIL étant supposé résolu), et que l'on dispose comme précédemment de fichiers personnes et organisations semi-structurés. Du croisement des premières bases citées dès lors qu'on a les processus algorithmiques de désambiguïsation des identifiants, on peut facilement extraire avec des outils puissants de Data Mining à « contraintes » des profils « hors normes, ou déviants ». On extrait alors l'ensemble total des profils labellisés « potentiellement dangereux » que l'on croise avec les fichiers personnes (avec alias, bios (textes) et caractéristiques de base associées). Le résultat de cette approche en « mode découverte » est l'obtention de profils comportementaux caractéristiques « hors normes » que l'on va associer à des agrégats organisationnels des clans ou mafias (dont on connaît par ailleurs certains des modèles d'action). Le regroupement ou la classification de ces types permet des recherches plus ciblées et donc beaucoup plus efficaces.

La Fusion « Données Paroles/ Données Structurées »

Bien entendu, ce que nous venons de mentionner quant à la capacité de fusionner analytiquement données structurées (numériques) et non structurées (textuelles) est l'une des configurations de fusion à laquelle il faudra s'intéresser en priorité. Mais selon ce même principe de

REVUE DE L'ELECTRICITE ET DE L'ELECTRONIQUE

N° et date de parution : 72 - 01/02/2007
Diffusion : 3000
Périodicité : Mensuel
LAREVUEDELE_72_81_310.pdf
Page : 91
Taille : 95 %

fusion analytique, qu'en est-il pour la « Parole » et les données structurées ? D'une façon évidente on peut considérer le traitement de la parole (« Speech Recognition » des anglo-saxons) comme un cas particulier du traitement du signal (APPROCHE I). Mais même si dans **l'identification ou l'authentification** d'un locuteur, on utilise les « formants du signal » : Pitch, Intensité, Prosodie etc.. comme caractéristiques de qualification et d'identification (pourvu que l'on dispose par ailleurs d'une base de « modèles de voix de locuteurs ») **on n'atteindra jamais le niveau « sémantique »** c'est-à- dire la capacité d'interpréter ce qui est dit uniquement avec des traitements Analyse du signal en tant que tels.

Une façon de **démultiplier la possibilité d'analyse** dans ce cas particulier est de recourir au processus « Speech to Text » (c'est-à-dire de transformer le signal en texte analysable, pour récupérer du pouvoir sémantique). Des approches existent, sur lesquelles il faut encore investir, mais avec lesquelles on peut déjà, à partir d'un ensemble d'interceptions téléphoniques ou d'écoutes, avoir une « translation » textuelle qui autorise le recours aux technologies textuelles (déjà mentionnées en 3.1) et dès lors s'attaquer au contenu sémantique de ce nouveau type d'information. Conséquence : on va considérer ce nouvel apport d'information comme des textes (certes courts) afin d' utiliser le principe de fusion analytique Texte / Data, exposé précédemment. On obtient dans ce cas le modèle :
Parole (« Speech ») => Texte => Données structurées numériques

Dans cette configuration le texte est le codage pivot qui permet de générer la corrélation « Speech to Data ».

3.3. La Fusion « Données Images et Cartographiques / Données Numériques Structurées »

Encore une fois, ce que l'on a décrit précédemment dans le cas de la fusion Data / Text et Data / Speech, se généralise, de fait, à des types de données de nature encore plus différente. Ainsi en est-il des données images et cartographie (SIG). Si l'on fait abstraction dans ce cas de la problématique de la taille du stockage des images qui croît très très vite dès lors qu'on les manipule en nombre conséquent. Une règle simple, certes très approximative peut être donnée qui illustre ce propos : 1 000 000 d'enregistrements numériques décrits par 100 variables équivaut au stockage de 100 000 textes de 10 lignes (environ 1 000 caractères) et de 100 images de 1024X1024 pixels. Néanmoins l'image est considérée aujourd'hui comme un support essentiel de qualification d'une information.

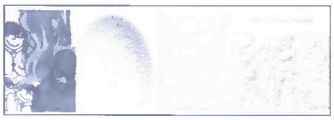
Une approche souvent utilisée pour gérer l'information image est de l'indexer textuellement (c'est en partie ce qui est fait dans le civil, agences de photos, centre d'iconographie etc..), mais le texte est ajouté à la main sur chaque photo et les processus de recherche d'information textuelle ne permettent seulement que de retrouver une image comme s'il s'agissait d'un texte court dans une collection. A ce propos, un autre challenge très important, est l'apport de l'image elle-même, comme moyen de désambiguïsation, lors de la recherche d'images via les annotations textuelles sous image ou photo.

Cette approche permet de désambiguïser la recherche via le texte sous image, par sa conjonction avec une image type « pattern » ou « image modèle » donnant une esquisse de ce que l'on désire réellement. Cette approche est non seulement utile lors de la recherche d'images, mais également lors de l'extraction de connexions thématiques d'illustration. Ces approches sont étudiées à l'INRIA, au LIP6, et à l'ENST.

*Prenons un exemple, supposons que l'on recherche via « Google Images » une illustration par des **« courbes, unimodales simples »** pour un texte de vulgarisation scientifique que l'on est en train d'écrire. Si l'on rentre dans l'encart du moteur Gooogle Images **: « courbes unimodales »** la réponse est : « **pas de document trouvé »***

*Si l'on entre dans l'encarté du moteur de recherche, une question plus générale : **« Courbes »**, on va obtenir pour les premières entrées la liste d'images suivante :*

1 2 3 4

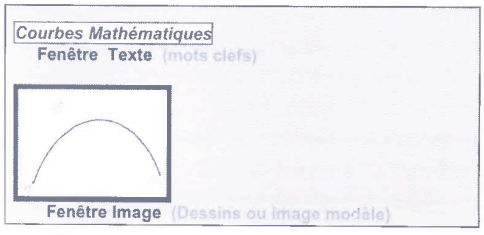


etc..

qui ne donne pas vraiment satisfaction

*Si l'on complexifie la question précédente par **« Courbes Mathématiques »**, les photos 1[4] (courbes du corps) et 4 (courbes de niveaux) disparaissent, mais on obtient un mélange de courbes, qui n'est nullement un filtrage de la liste précédente, mais qui nous fournit une nouvelle liste :*

Où nous sont proposées, dans l'ordre une « clothoïde », une courbe de Hilbert, une courbe de Peano, une courbe de la racine carrée complexe d'un nombre, une image d'oscilloscope, etc. Sauf à descendre profondément dans ces listes, on ne trouve aucune courbe digne de l'illustration cherchée. Il paraît évident que si l'on avait à notre disposition une double entrée avec d'un côté le texte attaquant l'annotation de l'image et d'un autre côté une fenêtre où l'on puisse « dessiner un contour » ou proposer une image « modèle ou témoin » de la façon suivante :

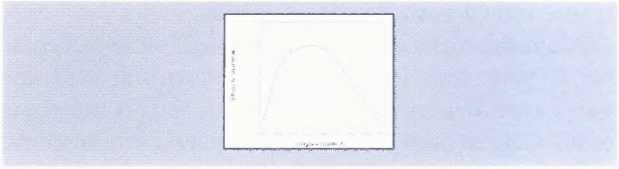


[4] La photo proposée réellemnent en N° 1 par Google Image n'est pas la photo donnée ici.

REVUE DE L'ELECTRICITE ET DE L'ELECTRONIQUE

N° et date de parution : 72 - 01/02/2007
Diffusion : 3000
Périodicité : Mensuel
LAREVUEDELE_72_81_310.pdf

On aurait davantage de chance de trouver par désambiguisation réciproque quelque chose comme l'image suivante :



Ceci suppose bien entendu que nous ayons parallèlement à un bon moteur de recherche par annotation textuelle, un bon algorithme d' « images matching » ou de calcul rapide de « distances entre images » ; c'est un point difficile sur lequel plusieurs laboratoires travaillent en ce moment.

Si le texte sert, comme dans le cas de la parole, de codage « Pivot » qui permet de gérer les images, il est rarement un facteur réel d'interprétation sémantique (car le texte associé à l'image est forcément court, sauf cas particulier) et l'image dans ce cas est considérée comme un support « illustratif » et non « analytique » par elle-même.

En revanche si l'on couple des algorithmes automatiques de recherche de contours, de formes caractéristiques etc.. dans l'image, à des générateurs de textes, on peut, moyennant une correspondance sémantique (un transcodage) faire correspondre à une image analysée un discours automatique généré (voir le concept de génération automatique de textes).

On obtient ainsi le processus d'interprétation de la fusion analytique sémantique Image/Data suivant le schéma :
Image –> Texte puis Texte –> Data. Et dès lors on se retrouve une fois de plus dans le processus déjà rencontré auparavant.

On peut également faire le même type de choses en cartographie via un codage pivot encore différent, pour donner à une carte un pouvoir informationnel direct sans recours à l'interprétation humaine (corrélation topographique qualitative versus données numériques) ce domaine porte déjà un nom il s'agit de la « Géo-intelligence ». Ceci veut dire qu'on est capable d'extraire automatiquement la « sémantique » corrélative d'une carte au même niveau qu'on le ferait pour un texte ou un enregistrement audio.

REVUE DE L'ELECTRICITE ET DE L'ELECTRONIQUE

N° et date de parution : 72 - 01/02/2007
Diffusion : 3000
Périodicité : Mensuel
LAREVUEDELE_72_81_310.pdf
Page : 93
Taille : 100 %

4. Le Problème de l'Hétérogénéité des Données « Intra Structure »

4.1. L'hétérogénéité propre aux données structurées

Le problème de l'hétérogénéité des données et la prise en compte de cette hétérogénéité dans les systèmes d'analyse est un point crucial qui se pose déjà de façon drastique au niveau des systèmes et des outils du marché, qui ne traitent pour l'instant que des données numériques.

En effet, que ce soient les outils anglo-saxons les plus connus ou issus des éditeurs les plus avancés du marché comme SAS Entreprise Miner, SPSS, IBM Intelligent Miner 4 Data, tous ne traitent que partiellement ce problème de l'hétérogénéité des données en simplifiant de façon parfois outrancière les données de nature différente pour les obliger à « **rentrer** » dans les modèles proposés, au détriment de la sémantique propre de ces données, n'oublions pas que toute base de données répond à une organisation dite « modèle de données » que les architectes imposent a priori aux « datawarehouses ».

Il est évident que des données « continues » (des poids, des tailles, des revenus, des mesures physiques, etc...) ne doivent pas être traitées dans le même environnement méthodologique que des données qualitatives comme la catégorie socioprofessionnelle, le lieu d'habitation, la présence ou l'absence d'une caractéristique ou d'un attribut, le sexe, etc.., ou si l'on désire le faire, il faut respecter un certain nombre de conditions qu'on désigne a par « **mise en compatibilité** ». Souvent cette mise en compatibilité revient soit à ramener les données qualitatives en des données continues par des processus (venant souvent des anglo-saxons) de « logistisation » avec quelques aberrations notoires (voir encarté) soit (et c'est plutôt l'approche de l'école francophone) à ramener les données numériques à des données qualitatives hiérarchiques par recours à des processus de seuillage ou de découpage non dénués de subjectivité (voir encartés suivants) :

Des palliatifs existent comme ceux qui ont été développés dans le cadre de la théorie de la « Similarité Régularisée » permettant de tenir davantage compte de la sémantique de chaque variable en travaillant sur des indices de similarité associés à une variable variable, corrigeant automatiquement les simplifications excessives faites par les analystes au moment de la sélection et de la collecte des informations et données à traiter.

Donnons quelques exemples simples de ce que recouvre cette approche de « Similarité régularisée » :

La notion de Similarité Régularisée pour un descripteur ou variable Vk *entre deux objets i et j est une fonction dépendant de la similarité intrinsèque* $S^k(i, j)$ *générée par le descripteur* k, *et d'une fonction* $\Pi^k(i, j)$, *mesurant la difficulté pour les individus* i *et* j *de partager la même valeur de* Vk *:*

$$SR^k(i,j) = S^k(i,j)(1 - \Pi^k(i,j))$$

Il existe un grand nombre de fonctions $\Pi^k(i, j)$ *certaines reposent sur les considérations* ***logiques***, *d'autres sur des considérations* ***probabilistes*** *(les plus simples), ainsi la fonction* $\Pi^k(i,j) = \frac{1}{p_k}$ *est-elle associée au cas très simple de variables ou descripteurs à* Pk *modalités et la similarité régularisée associée devient :*

$$SR^k(i,j) = S^k(i,j)\left(1 - \frac{1}{p_k}\right)$$

Dans ce cas on voit bien que plus Pk *est grand plus la similarité* $SR^k(i, j)$ *est forte. Mais ceci est un cas simpliste, il existe aussi des variantes de la fonction* $\Pi^k(i, j)$ *reposant sur des considérations Statistiques voire pour les plus complexes (structures densitaires), exprimables uniquement via le recours à des notations de la théorie de l'Analyse Relationnelle (représentation matricielle de chaque variable par des graphes de relations binaires) (voir par la suite).*

D'ailleurs on peut montrer mathématiquement que la similarité Régularisée globale est la somme (ou la moyenne arithmétique) des similarités régularisées de chaque variables selon la formule suivante, qui intègre en une seule formulation l'agrégation de similarités complexes calculées sur chacune des variables :

$$SR(i,j) = \frac{1}{m}\sum_{k=1}^{m} SR^k(i,j)$$

$$SR(i,j) = \frac{1}{m}\sum_{k=1}^{m} S^k(i,j)(1 - \Pi^k(i,j))$$

REVUE DE L'ELECTRICITE ET DE L'ELECTRONIQUE

N° et date de parution : 72 - 01/02/2007
Diffusion : 3000
Périodicité : Mensuel
LAREVUEDELE_72_81_310.pdf
Page : 94
Taille : 95 %

Donnons des exemples de fonction de similarité régularisée plus complexe : en effet si nous codons une variable sous forme relationnelle , soit les deux exemples suivants :

$V_k =$

a	1
b	1
c	3
d	2
e	2

Alors son codage relationnel s'écrit :

$V_k =$

a	1	1			
b	1	1			
c			1		
d				1	1
e				1	1

Sous cette forme V_k, qui est une variable qualitative, correspond à une matrice binaire relationnelle dont le codage s'exprime par :

$$C_{ij}^k = \begin{matrix} 1 & si & V_k(i) = V_k(j) \\ 0 & sinon & \end{matrix}$$

Dans le cas d'une variable quantitative le codage associé peut prendre plusieurs formes possibles, dont une assez simple, mais non optimale, décrite ci-dessous :

$$C_{ij}^k = \begin{matrix} 1 & si & V_k(i) - V_k(j) < s \\ 0 & sinon & \end{matrix}$$

s *étant un seuil de séparabilité*

Si nous prenons à titre d'exemple comme indice de similarité « sémantique » sur V_k (cas de la variable qualitative, décrite plus haut), un indice dit de « présence-rareté [5] défini par :

$$S^k(i,j) = \begin{cases} \dfrac{C_{ij}^k}{\sum_j C_{ij}^k} = \dfrac{1}{n(i)} & si\ V_k(i) = V_k(j) \\ 0 & sinon \end{cases}$$

où n(i) = Nombre [7] d'objets semblables à i

Si par ailleurs on définit une fonction $\Pi^k(i,j)$ de difficulté de « matching », fonction de k, i, j, *suivant la formulation suivante :*

$$\Pi^k(i,j) \quad \sum\sum \frac{C_{ij}^k}{N^2} \quad \frac{C_{..}^k}{N^2}$$

cette fonction représentant dans le cas général la « densité » de « 1 » dans la matrice relationnelle C^k de V_k , en particulier si toutes les modalités sont équiréparties dans la population (c'est-à-dire) chacune regroupe des effectifs d'objets de même taille, la fonction ci-dessus nous redonne le cas déjà décrit précédemment :

$$\Pi^k(i,j) = \frac{1}{p_k}$$

à ce propos , les fonctions de « difficulté de matching » données ci dessus, sont des constantes ne dépendant que de V_k, mais une fonction comme celle qui suit, correspondant à une configuration que nous ne développerons pas, est, elle, bien dépendante de i *et* j *:*

$$\Pi^k(i,j) = \frac{\sum_i C_{ij}^k + \sum_j C_{ij}^k}{2N} = \frac{C_{i.}^k + C_{.j}^k}{2N}$$

La similarité globale des deux individus i *et* j*, donnée par :*

$$SR^k(i,j) = S^k(i,j)(1 - \Pi^k(i,j))$$

prend alors la forme suivante dans le cas de la similarité de « présence-rareté » :

$$SR^k(i,j) - \frac{C_{ij}^k}{\sum_j C_{ij}^k}\left(1 - \frac{C_{..}^k}{N^2}\right)$$

qui s'écrit aussi :

$$SR^k(i,j) - \frac{C_{ij}^k}{C_{i.}^k}\left(1 - \frac{C_{..}^k}{N^2}\right)$$

On peut constater l'intérêt logique de cet indice quant à son pouvoir de « dépondération » de variables trop peu discriminantes. En effet si la variable V_k et sa représentation C^k relationnelle, représentent la variable « triviale » où tous les objets ont la même valeur, alors : $C_{..}^k = N^2$ et de fait : $SR^k(i,j) = 0\ \forall (i,j)$, ce qui montre que moins une variable a de modalités, plus l'indice de similarité régularisée associé est faible en valeur.

En conclusion de cette petite introduction aux cas simples d' indices de Similarité Régularisée, on a montré comment s'effectue la dépondération automatique, des variables ayant peu de modalités, au profit de celles en ayant beaucoup .

On calculera la similarité globale associée au cas de l'indice de « présence-rareté » précédent, en posant :

$$SR(i,j) = \frac{1}{m}\sum_{k=1}^{m}\frac{C_{ij}^k}{C_{i.}^k}\left(1 - \frac{C_{..}^k}{N^2}\right)$$

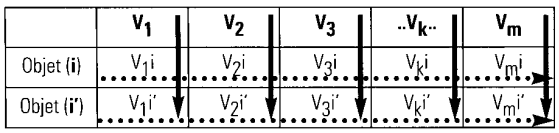
(Pour plus de détails sur cette approche voir référence [1]).

A titre de conclusion, et en complément à ce qui vient d'être dit, une autre façon de caractériser l'approche par « Similarité régularisée » peut être schématisée par le graphique suivant :

Un indice de similarité de « présence-rareté » entre deux objets i *et* j *est d'autant plus fort que* i *et* j *sont peu nombreux à partager la même valeur de V_k (d'où le concept de « rareté »). à titre d'exemple la similarité des objets (a et b), de l'exemple précédent est égale à :* Sk (a, b) =1/2, *puisque a et b possèdent la même valeur de V_k et parce qu'ils sont 2 à posséder cette même valeur.*

REVUE DE L'ELECTRICITE ET DE L'ELECTRONIQUE

N° et date de parution : 72 - 01/02/2007
Diffusion : 3000
Périodicité : Mensuel
LAREVUEDELE_72_81_310.pdf
Page : 95
Taille : 95 %



L'approche classique revient à calculer la similarité entre objets de façon longitudinale (flèches pointillées), l'approche par similarité régularisée commence par calculer la similarité entre individus, séparément variable par variable (en tenant compte de la sémantique propre à chaque variable) de façon verticale (flèches pleines), puis somme dans un deuxième temps les similarités calculées variable par variable, en une similarité agrégée globale.

Par ailleurs, on pourrait penser que l'approche décrite ci-dessus est totalement extérieure aux traitements statistiques usuels, tant les formules présentées s'éloignent des approches classiques, ceci n'est pas du tout le cas, comme on l'explique ci-dessous. En effet, cette approche théorique est reliée directement à des fondements statistiques de base, ici dans le cas de la similarité de « présence-rareté », il s'agit du Chi^2*, via l'égalité suivante (voir* [16]*) :*

$$\chi^2(V_k, V_{k'}) = N\left(\sum_{i=1}^{\lambda}\sum_{j=1}^{\lambda}\frac{C_{ij}^{k}}{C_{i.}^{k}}\frac{C_{ij}^{k'}}{C_{i.}^{k'}} - 1\right)$$

La statistique contingentielle du chi^2 *est donc, à quelques modifications de constantes près, le produit des indices de similarité de « présence-rareté » des deux variables* V_k *et* $V_{k'}$*.*

Pour illustrer les conséquences des recodages ou des simplifications a priori sur les données, présentons les deux cas connus suivants :

Exemple d'aberration possible du codage tout quantitatif à l'anglo-saxonne : la variable « Sexe » peut varier de 0 à 1 (ça ne gène personne d'extraire d'une régression logistique une valeur du sexe égale à 0,75 ???),

Exemple de la subjectivité du codage tout qualitatif à la francophone : si l'on découpe la variable « Age » en tranches [20 ans - 30 ans] puis [30 ans - 40 ans] etc.. pourquoi considérer comme différents une personne de 29 ans et quelqu'un ayant 30 ans ?? alors qu'il sont séparés dans le seuillage proposé, donc analysé différemment sur le concept « Age ».

Tout analyste, ayant déjà eu à traiter des fichiers de données comportementales sur des clients ou des fichiers de description de produits, de services ou des fichiers de descriptions biométriques etc... s'est souvent trouvé dans l'obligation dans une phase dite de pré-traitement de transformer, par exemple, des données sur l'âge des clients en « tranches d'age » au lieu de conserver les données brutes, les données sur les revenus étant transformées de la même façon en « tranches de revenu » etc...

Ce processus de pré-traitement n'est que la traduction évidente de l'incapacité des outils actuels, quels qu'ils soient, de traiter de façon directe, sans modification aucune, les données telles qu'elles se présentent dans la réalité.

En d'autres ter es, à quoi cela sert-il de conserver, de collecter, de structurer, des Giga-octets voire des Tera-octets (dans les datawarehouses les plus « monstrueux ») si l'analyse de ces données, faite a posteriori, ne respecte pas la nature des informations qui ont été stockées. Ces informations sont souvent, elles-mêmes, de mauvaise qualité : données aberrantes, erronées, manquantes, etc.. Ces simplifications peuvent dans certains cas détériorer l'information initiale de façon telle que toute conclusion obtenue ou toute corrélation ou implication trouvées pourraient n'être le fruit que de simplifications aléatoires, de seuillages ou de découpages a priori sans corroboration aucune avec une réalité terrain, une connaissance métier, ou bien une expérience de longue date de l'analyse des données ?.

Ce retour sur les problématiques de la prise en compte des données numériques montre à l'évidence que le **respect de l'information** est un point majeur dans la qualité des analyses qui pourront être faites par la suite.

Fort de ce constat, qui au demeurant n'est pas encore compris à sa juste valeur, puisque tout le monde pratique, sans aucun recul, ce genre de pré-traitements sur les données numériques, comment pourrait-on alors imaginer traiter des données, qui bien au-delà de l'hétérogénéité intrinsèque **(intra)** aux données structurées va porter sur une hétérogénéité (inter) entre données structurées et données non structurées (textuelles, audio ou vidéo).

L'hétérogénéité, dans le cas des données textuelles, est plus directement fondée sur la nature du texte considéré : Comment comparer un « mail » de 15 lignes, plein de fautes d'orthographe, avec un « livre » de 120 pages d'un niveau de langue élevé ou un article scientifique de 15 pages avec formules et vocabulaire très spécialisé... on voit bien ici que tout ceci relèvera de la même problématique.

C'est également l'un des challenges de la « Fusion Analytique » des informations que de prouver que traiter simultanément et globalement l'information selon toutes ses formes tous ses types et toutes ses sources est largement plus performant que de la traiter et analyser séparément et localement. Ceci sans compter les **gains de productivité** liés aux chaînages d'outils, transparents pour l'utilisateur. Les outils inter-opèrent les uns avec les autres, les résultats des uns nourrissant les autres de façon directe en informations raffinées et plus riches. Cette démultiplication rajoute de la puissance au système (intégration des différents outils). Outils, qui s'ils avaient été utilisés séparément n'auraient donné que des résultats très parcellaires, peu manipulables et surtout peu synthétisables.

REVUE DE L'ELECTRICITE ET DE L'ELECTRONIQUE

N° et date de parution : 72 - 01/02/2007
Diffusion : 3000
Périodicité : Mensuel
LAREVUEDELE_72_81_310.pdf
Page : 96
Taille : 95 %

6. Conclusion

Ne nous trompons pas de cibles, même s'il faut être vigilants sur l'aspect technologique de l'ensemble des processus mis en e›ergue dans les 6 niveaux, appartenant aux deux Approches citées au début de ce dossier, c'est dans la maîtrise **des différentiateurs « intelligents »** et à grande valeur ajoutée que résidera le pouvoir de ésistance aux colossaux investissements que sont en train de déployer les anglo-saxons et en particulier les Américain: dans ces domaines.

N'oublions pas que nous avons perdu la **guerre de « l'Informatique »**. Car s'ils ont à craindre aujourd'hui des pays comme la Chine ou l'Inde à qui ils confient d'ailleurs, avec une certaine légèreté, des développements technologiques partiels, les Etats-Unis et leurs fleurons techniques et technologiques comme Intel, Microsoft, IBM, Oracle, Texas Instrument, Hewlett Packard, Dell, Google etc..... ne nous considèrent plus, nous les Européens, comme des concurrents en Informatique. D'ailleurs s'ils condescendent à jeter un œil sur SAP (Allemagne) ou Dassault Systèmes (France), ils ont déjà envisagé de s'approprier le premier et contrôlent le second (au travers d'IBM) en jouant pour eux le rôle de vecteur de ventes. Seuls Fujitsu et NEC conservent au Japon un petit pouvoir de contre-balancement de cette hégémonie. Ce n'est d'ailleurs pas la valeur de nos produits qui était en cause mais bien l'ampleur des investissements associés (Bull du temps d'Honeywell-Bull n'avait-elle pas trouvé, la première, les plans architecturaux des futurs PC ?).

Une conclusion stratégique s'impose : la France qui ne peut rivaliser au niveau investissements avec les Anglo-saxons dans l'ensemble de la chaîne « Analyse de l'Information » (« Information Dominance »), doit **concentrer au maximum son effort dans les technologies à « Effet de Levier »**. C'est-à-dire les technologies qui démultiplieront cet effort, en s'appuyant en particulier sur les domaines où nous sommes encore parmi les « leaders » mondiaux au niveau « conceptuel ».

Il nous semble opportun de tirer parti de cette situation en travaillant sur l'intégration de méthodologies et technologies à **« effet de levier »** dans les chaînes technologiques d'Analyse de l'Information.

En effet ces technologies fondées sur le concept de « savoir cumulatif », seront très originales, très pourvoyeuses de « valeur ajoutée » et surtout **très peu copiables** car issues d'un long proces‹us de maturation propre à os avancées françaises. L'effet de levier inhérent à ces technologies innovantes, réside dans sa faculté de démultiplier la puissance (tant au niveau des performances que de la qualité) des outils amont et aval de la zone à « effet de levier », c'est une garantie d'innovation, assurée pour l'ensemble des chaînes d'intégration tant systèmes que technologiques.

En conclusion nous pouvons tirer les recommandations suivantes :

1 La maîtrise de l'information aujourd'hui est un challenge du même niveau que celui de l'informatique il y a 30 ans.

2 L'assurance que nous soyons une Nation Cadre indépendante pour ses ressources et pour ses capteurs technologiques informationnels est fondée sur un investissement **concentré au maximum** sur nos « savoirs cumulatifs » (i.e. là où la France est toujours forte tant au niveau conceptuel que technologique) et sur notre capacité à les intégrer.

3 Le domaine entier de la « fusion de l'information » fait partie de ces secteurs clef, où nous ne devons pas nous faire distancer et où il existe des niches potentiellement fortes de « savoir cumulatifs »

4 La valeur ajoutée de toute la chaîne de traitement, liée à la maîtrise de l'information, est fondée sur les couches à « intelligence » de cette chaîne, c'est-à-dire celles ou les technologies à « effet de levier » se situent essentiellement, deux exemples clairs :
   - Au niveau capteurs sur la couche dite : « Multisensors / Multiplatforms fusion »
   - Au niveau Analyse sur la couche dite « Semantic correlation ».

Ceci ne veut pas dire qu'il faut délaisser le reste de l'édifice, mais que nous devons optimiser les affectations de nos investissements et les cibler pour un réel et visible « ROI » pour gagner le challenge tout à fait atteignable de la « Maîtrise de l'Information pour l'Indépendance de nos Décisions ».

Références

[1] H. BENHADDA & J.F. MARCOTORCHINO, *"Introduction à la similarité régularisée en analyse relationnelle"*, Revue de Statistique Appliquée, Vol.46,N°1,pp45-69,1998.

[2] Y. BENNANI, *"Apprentissage connexionniste"*, Editions « Hermès Science Publications », Paris, 2006 .

[3] I. BLOCH &T H. MAÎTRE, *"Fusion d'information en traitement du signal et de l'image"*, Livre, Editions « Hermes Sciences Publications », Paris, 2003.

[4] P. DEHEUVELS (Paris VI et Académie des Sciences) & J. F. MARCOTORCHINO, (Thales et Paris VI), *"Statistique et informatique, la nouvelle convergence"*, Revue RST de l'Académie des Sciences n° 8, TECDOC Editeur, Paris, juillet 2000.

[5] L. DENOYER & P. GALLINARI, *"Bayesian Network Model for Semi-structured Document Classification"*, Information Processing and Management, 2004.

[6] J.J. HOPFIELD, *"Neural Networks and Physical systems*

REVUE DE L'ELECTRICITE ET DE L'ELECTRONIQUE

N° et date de parution : 72 - 01/02/2007
Diffusion : 3000
Périodicité : Mensuel
LAREVUEDELE_72_81_310.pdf
Page : 97
Taille : 100 %

with emergent Collective Computational Abilities", Proceedings of the National Academy of Sciences, USA, 1979, pp. 2554-2558.

[7] N. HOWARD, *"Least Squares Classification and Principal Component in Qualitative Analysis in the Social Sciences"*, Dogan M. et Rokkam S. Eds, MIT Press Vol.13, N° 2, pp107-128, 1979.

[8] L.J. HUBERT & R.G. GOLLEDGE, *"Matrix Reorganization and Dynamic Programming: Applications to Paired Comparisons and Unidimensional Seriation"*, Psychometrika, Vol 46, N° 4, 1981.

[9] A. KUSIAK, A. VANELLI & R. K. KUMAR, *"Clustering Analysis: Models and Algorithms"*, Control & Cybernetics, Journal of Polish Academy of Sciences, Systems Research Institute, Vol.15,N° 2, 1986.

[12] L. LEBART & A. SALEM, *"Statistique textuelle"*, Livre, Dunod, Paris 1994.

13] A. LELU, *"Modèles neuronaux pour l'analyse documentaire et Textuelle"*, Thèse Université Paris VI, 1993.

[14] H. MAÎTRE, *"Le Traitement des images"*, Livre, Editions « Hermes Sciences Publications », 2004.

[15] C. MALIS, *"L' apport des technologies et des solutions avancées du traitement du langage, appliquées à la gestion des risques et des crises en contexte dual"*, Supports de la Conférence : Rencontres : ICC 2006 , Paris.

[16] J. F. MARCOTORCHINO, *"Maximal Association Theory as a Tool for Classification and Clustering"*, in the book: Classification as a Tool for Research, (Gaul W. et Schader M. Editors) North Holland , Amsterdam, 1986.

[17] J. F. MARCOTORCHINO, *"Les Technologies avancées de l'analyse de l'information: Text Mining, Data Mining et Fusion Data Mining-Text Mining"*, Dossier « la Gestion de la Connaissance », Revue de la SEE, REE n° 8, juillet 2001, EDP Sciences, Paris.

[18] W. H. RAND, *"Objective Criteria for the Evaluation of Clustering Methods"*, JASA, Vol 66, 1971.

[19] Rapport Infom@gic : SP3.21/ R1 : J. F. MARCOTORCHINO. *"Réduction de Dimensionnalité, Approches Relationnelles"*, avril 2006.

[20] M. RIFQI, V. BERGER & B. BOUCHON-MEUNIER, *"Discrimination Power of Measures of Comparison"*, Fuzzy Sets and Systems, Vol 110(2) pp 189-196, 2000.

[21] G. SAPORTA, *"About Maximal Association Criteria in Linear Analysis and in Clustering"*, in: Classification and Related Methods of Data Analysis, book edited by H. Bock, Elsevier Science Publishers, North Holland, 1988.

Les auteurs

Jean-François Marcotorchino est Directeur Scientifique de Thales Land & Joint Systems depuis 2002. Il est Docteur ès Sciences Mathématiques, et a été nommé « Professeur des Universités » en 1994. Il est actuellement PAST à l'Université de Marne la Vallée. Il est également Directeur de Recherche au Laboratoire LSTA (Laboratoire de Statistique Théoriques et Applications de l'Université Paris VI). Avant de rejoindre Thales, J.F. Marcotorchino a fait une grande partie de sa carrière chez IBM France et IBM EMEA, en tant que spécialiste des domaines de l'Aide à la décision de la Statistique Mathématique et de l'Analyse des données. Après avoir été Directeur du Centre Scientifique IBM de Paris pendant 12 ans, puis Directeur Scientifique, il a quitté IBM en 2001. J.F. Marcotorchino est par ailleurs Président de la Commission Thématique et Technique : « Ingénierie de la Connaissance » du Pôle de Compétitivité : CAP DIGITAL de la Région Parisienne. Il a été co-éditeur de la Revue Américaine « Applied Stochastic Models and Data Analysis » de Wiley, ainsi que d'un certain nombre de livres de la collection « Advanced Series in Management » de l'Editeur North Holland. Il est actuellement membre du comité de publications de la Revue REE de la SEE.

Christian Malis : Ancien Elève de l'Ecole Normale Supérieure de la Rue d'ULM, Christian Malis est actuellement Directeur de l'Innovation au sein de la Direction R&T de Thales Land & Joint Systems

Patrick Constant : Ancien élève de l'Ecole Nationale des Télécoms ENST, et Docteur en Linguistique Computationnelle, Patrick Constant, co-fondateur de Pertimm SAS, en est le Président Directeur Général aujourd'hui.

Henri Maitre : Docteur ès Sciences Physiques de l'Université Paris VI, ancien élève de l'Ecole Centrale de Lyon, Henri Maitre est professeur au Département TSI de l'ENST. Il est Directeur du Laboratoire de Communication et Traitement de l'Information, Unité Mixte de Recherche du CNRS et du GET/Télécom.