René PATESSON
Bruxelles

LE TRAITEMENT INFORMATIQUE DES DONNEES EN SCIENCES HUMAINES : QUELQUES PRINCIPES.

Notre propos est de présenter ici quelques principes qui sont à la base de notre activité dans la construction de logiciels généraux de traitements de données applicables aux disciplines des sciences humaines. Au préalable nous situerons les conditions dans lesquelles le chercheur est amené à opter pour l'informatique.

D'une manière très générale, le choix informatique dans le traitement des données se fait en considérant trois critères de base : rapidité, fiabilité et exhaustivité. La rapidité est la propriété essentielle de l'ordinateur de pouvoir exécuter quasi instantanément des traitements complexes sur de grands ensembles de données, une fois que celles-ci ont été introduites dans le système informatique et que les programmes de traitement ont été écrits. Une fois les données enregistrées, vérifiées et corrigées, et les algorithmes testés, plus aucune erreur matérielle n'est possible lors des traitements ; la fiabilité informatique est totale et seules les erreurs de logique dans la programmation, indépendantes par essence même de l'ordinateur, sont possibles. Par ailleurs, à tout moment c'est l'entièreté du corpus qui est disponible pour le programme et la vérification des hypothèses peut se faire rigoureusement sur le fichier complet.

En corollaire, l'ordinateur débarasse le chercheur des tâches de manipulation et de collationnement de ses données et

lui permet de se consacrer plus longuement à l'exploitation des résultats. Les hypothèses qui lui viennent alors à l'esprit, dans la mesure où celles-ci peuvent être décrites dans le langage de programmation, seront éprouvées avec rapidité. En traitement manuel, les hésitations à entreprendre certains dépouillements dont la pertinence ne se justifie peut-être pas compte tenu des résultats esperés, conduisent souvent à abandonner les investigations.

Se pose alors la distinction entre d'une part l'informatique proprement dite, incluant le matériel, tel ordinateur, les langages, les mémoires, etc. qui sont autant d'outils à la disposition du chercheur, et d'autre part les choix méthodologiques, les programmes dont seule la théorie offre une justification.

Cette distinction est importante et disparaît bien souvent aux yeux du profane qui au-travers du mythe de la machine à l'intelligence anthropomorphe confond ces deux entités n'établissant pas la solution de continuité entre les causes et les moyens. Pourtant le choix de la méthode est la pièce maîtresse du traitement par ordinateur ; elle prend en considération les théories et les modèles et conduit à décrire un ensemble de règles à appliquer sur les données.
Les règles sont la démarche du chercheur dans son investigation du corpus et il les appliquerait en l'absence de toute machinerie. L'intervention de l'informatique est postérieure à ces choix ; elle aidera dans la possibilité de retranscrire les règles en algorithmes tenant compte des restrictions apportées par les langages, et mettra en disponibilité un outil d'exécution automatique des opérations. Par ailleurs, dans son application aveugle, l'ordinateur reproduira strictement ce que le programme contient, qu'il y ait des erreurs de logique ou non, sur les données enregistrées qu'elles contiennent des informations erronnées ou non.

Pour le choix des modalités de traitement plusieurs cas sont possibles dont nous exposons les extrêmes.

Ayant précisé sa démarche, le chercheur construit un logiciel de traitement ad hoc. L'investissement en programmation

est important mais il peut se justifier par la taille du corpus, par le degré de raffinement dans les exploitations. La maîtrise du système par l'utilisateur est complète ; il obtiendra exactement ce qu'il souhaite et dans la forme qui lui convient. Toute progression dans la recherche s'accompagnera d'une évolution parallèle des logiciels en y intégrant de nouvelles fonctions construites sur la base des résultats précédents.

A l'opposé et pour des raisons concernant l'économie dans l'investissement informatique essentiellement sur le plan de la programmation, une fois sa démarche décrite, le chercheur recourt à des programmes existants offrant des traitements généraux et standardisés. Par recoupements dans la multiplicité des résultats on parvient à extraire l'information pertinente, dont la présentation sera moins raffinée que pour la solution ad hoc.

Pour illustrer cette alternative nous utiliserons l'analogie des moyens de transport. Le conducteur, propriétaire de sa voiture peut aisément au départ de son domicile rejoindre n'importe quelle destination par l'itinéraire de son choix, optimisant certaines conditions comme la rapidité, les encombrements, l'agrément du paysage, etc. Par ailleurs le non conducteur qui veut se rendre en un lieu éloigné doit combiner son itinéraire au travers des moyens de transports à sa disposition : à pied jusqu'au bus ; le bus jusqu'à la gare de départ ; le train jusqu'à la gare d'arrivée ; à nouveau le bus ou le taxi jusqu'à sa destination.

Bien qu'imparfaite cette analogie met l'accent sur les différentes modalités du choix. S'ils arrivent tous deux à destination, le conducteur peut néanmoins se tromper de chemin en cours de route (erreurs de programmation) ; par contre l'utilisateur des transports en commun est dépendant des itinéraires (programmes existants) et des horaires (disponibilités de l'analyste).

Notre expérience dans le traitement des données nous a conduit à développer des logiciels généraux de traitement (packages) couvrant d'une part le dépouillement de matériel qui se présente sous la forme de textes et d'autre part sous

la forme de collections de renseignements.

Les principes de la construction de ces packages répondent à plusieurs objectifs :

a. offrir des dépouillements de base simples mais aux multiples possibilités et préparant éventuellement le traitement ultérieur par d'autres programmes ;

b. permettre le traitement sur des fichiers de données, eux-mêmes constitués en transformant le moins possible les sources documentaires ;

b. offrir des mises en page de résultats qui sont claires à lire et à interpreter et qui peuvent éventuellement être utilisées directement lors d'une publication ;

d. offrir un moyen simple de piloter le package par des instructions transparentes, des mots clés qui correspondent à une écriture en clair des opérations que l'utilisateur veut entreprendre.

Dans le domaine du traitement des textes deux packages ont été developpés. L'un réalise des concordances (CONCORDE) des textes soumis, l'autre (GITEX) permet l'analyse lexicale des textes à plusieurs niveaux : index complets ou partiels par catégorie de mot, par locuteur, par partie de texte ; relevés de statistiques lexicales ; études de co-occurrences ; analyses en grappe (co-occurences multiples); etc.

A titre d'exemple nous illustrons ici notre démarche dans la construction du package GITRI orienté,lui,vers le traitement des collections de renseignements.

Cette construction passe tout d'abord par le relevé des opérations courantes qui sont effectuées manuellement sur les fiches et dont le chercheur pourrait être débarassé, ensuite par le relevé des résultats auxquels il souhaite arriver lors de ces manipulations ; enfin par une formulation plus abstraite de ces opérations concrètes cadrant avec l'objectif de généralisation du logiciel.

Dans le relevé des operétions, la démarche est calquée sur celle de l'analyse du travail en psychologie industrielle par

laquelle on s'informe, à l'aide de techniques appropriées, auprès du chercheur des travaux qu'il exécute couramment. Nous avons pu ainsi relever les tâches suivantes :

- retranscrire les données à partir des documents originaux ;
- controler les informations retranscrites, les corriger éventuellement ;
- ajouter ou éliminer certaines informations ;
- établir des relevés et des listes ;
- faire des collationnements et des comparaisons ;
- établir des classements complets ou partiels ;
- rechercher une information ;
- réaliser des dénombrements, des comptages d'occurences ;
- retranscrire des résultats.

Sur la base de ces considérations nous en arrivons à formuler les points essentiels du traitement informatique des collections de renseignements auxquels doit répondre le logiciel.

En ce qui concerne les données, à chaque objet se rapporte une collection de renseignements qui sont autant d'entités variables en nombre, en nature, en longueur et en destination dans les traitements. Le logiciel doit adopter un mode de caractérisation de ces variables pour chacune de ces dimensions.

On distinguera trois classes d'opérations différentes selon qu'elles portent directement sur le corpus de base et répondent aux besoins de listage, contrôle ou transformation des données de base avant un traitement, selon qu'elles répondent aux opérations du traitement lui-même devant conduire aux résultats attendus comme les tris, les classements, les dénombrements et enfin selon qu'elles concernent l'édition, la présentation et la conservation des résultats.

Le package comportera alors des possibilités simples pour caractériser les éléments variables dans les données et comprendra une liste de mots clés avec leurs options qui seront autant d'opérations que le chercheur veut voir exécuter sur son corpus.

A titre d'exemple nous reproduisons ici les cartes de paramétrisation du package utilisé dans le dépouillement d'une

étude sur les registres des patentables à Bruxelles à la fin du XIXe siècle (1) ainsi que les résultats obtenus dans les conditions des traitements demandés par les constructions en questions.

Remarquons que le choix des codes qui peuvent paraître hermétiques au lecteur non concerné, a été fait par le chercheur uniquement dans un souci d'économie d'écriture dans la retranscription de ses données et qu'il aurait très bien pu adopter un mode d'écriture plus explicite, le programme n'imposant pas arbitrairement l'usage de tels codes.

---

(1) G. KURGAN-VAN HENTENRYK et G. VIRE. Les registres des patentables, sources de l'histoire de Bruxelles à la fin du XIXe siècle, *Histoire et méthodes*, IV, 367-415. Ed. de l'Université de Bruxelles, 1981.

Cartes de paramétrisation nécessaires pour établir la hiérarchie des patentes (cumulées) payées par les patentables

TRI PATENTES	titre du tri
10 ZONES	nombre de zones définies sur 1 carte
1 CARTES	nombre de cartes par enregistrement
ZONES	définition des zones :
01/S/SECT 1/01-1/02	— numéro de la section ;
02/S/IDENT 1/03-1/06	— identificateur au sein de la section ;
03/S/CARTE 1 1/07-1/57	— adresse, sexe, code professionnel, activités du patentable ;
04/S/1 1/58	— activité principale du patentable ;
05/S/CL IMP 1/62-1/64	— classe d'impôt (toutes les patentes payées par un patentable sont cumulées et indiquées dans la classe correspondante du tarif A) ;
06/S/CARTE 2 1/59-1/61	— autres éléments d'information relatifs au patentable ;
07/S/N STAT 1/65-1/67	— numéro statistique de la profession ;
08/S/CARTE 3 1/68-1/78	— autres éléments d'information relatifs au patentable ;
09/S/SA 1/79	— secteur d'activité ;
10/S/BA 1/80	— branche d'activité.
OPTIONS	caractéristiques du tri :
TRIS 10 9 7 5 2	— ordre dans lequel les zones sont triées ;
ÉDITIONS 1 2 3 4 6 5 7 8 9 10	— ordre dans lequel les zones sont éditées ;
MISE EN PAGE 1 3 7 58 59 62 65 68 79 80 50 LIGNES PAR PAGE	— position du début de chaque zone éditée sur le listing ;
4 ZONES DE COMPTAGE 5 7 10 9 80 90 100 110 5 5 5 5	— nombre et définition des zones qui font l'objet d'un comptage (numéros des zones, position sur le listing, nombre de chiffres prévu) ;
FILTRES 2 FILTRES 03 2 01 1 01 2 01 3	— nombre et définition des filtres (numéros des zones servant de filtres, type de filtre, nombre et liste des valeurs des filtres).
FIN	

Remarques :

Les mots en capitales représentent des mots-clé du programme de tri et doivent, dès lors, toujours être orthographiés de la même façon.

Les deux filtres ont pour but de retenir, lors du tri, les patentes payées par les patentables habitant la section 3 (second filtre), et d'éditer uniquement les données relatives à la patente payée pour l'activité exercée à titre principal (premier filtre).

Résultats obtenus à l'issue du tri des patentes (cumulées) payées par les patentables habitant la section 3.

Ident.	Nom		Adresse	Code prof.	Act.	Cl. imp.	N° stat.	Br. et sect. d'activ.	Comptages		
3 311	VANDENVELDE	JMPL	NINOVE	4	ABAT P2211	6N315		2A			
									2	7	32
3 35	ANULLEN	EMR	REMP DES MOINES	84	FTBI P8811	7N391/393		2A			
									1	1	33
3 498	HENRI	LMR	FABRIQUES	9	CONF P8811	11N391		2A			
3 557	LANERES	LMR	FLANDRE	58	PATI P8811	11N391/392		2A			
3 66	PILAET	AMR	ANDERLECHT	63	COPA P8821	11N391/392		2A			
31459	RICHARD	EBD	ANSPACH	125	PATI P8821	11N391		2A			
31016	SIAEMS	EPL	ST GERY	12	CONF P8821	11N391		2A			
31162	VAN CRAENEN	VMR	VAN ARTEVELDE	15	PACO P8821	11N391/392		2A			
									6	7	39
3 663	BRAYARD	AMR	GRANDE ILE	1	PATI P8811	11N392/391		2A			
3 848	LANGHENDRIES	PMR	DEVAUX	21	PATI P8821	11N392		2A			
									2	2	41
3 769	BLOEK +	R	JERICHO	2 +	BOUL P8821	8N395		2A			
									1	1	42
31429	WERY	LFR	ANDERLECHT	70	BOUL P8821	10N395		2A			
3 303	WILLOCX	EMR	BORGVAL	5	BOUL P8821	10N395		2A			
									2	3	44
3 788	DEPSAM	CHR	MARCHE POULETS	11	BOUL P8811	11N395/302		2A			
31626	SMITS BAERT	LMR	FLANDRE	112	BOUL P8821	11N395		2A			
									2	5	46
3 56	BAEYENS	VER	ANDERLECHT	35 +	BOUL P8821	12N395		2A			
3 435	DE VULDER	VER	CHARTREUX	66	BOUL P8821	12N395		2A			
3 289	HELTZER	CMR	BODEGHEM	28	BOUL P8821	12N395		2A			
31139	PLOUM	JMR	SOIGNIES	8	BOUL P8821	12N395		2A			
31041	VAN BLADEL	EMR	STE CATHERINE	30	BOUL P8821	12N395		2A			
31916	VANLINTHOUT	EMR	SOIGNIES	28 +	BOUL P8821	12N395		2A			
31350	VERINCKX	BME	AUX GRAINS	27	BOUL P8821	12N395		2A			
									7	12	53
31055	BAEDIS	D R	SENNE	29	BOUL P8821	13N395		2A			
31277	BALISTER	HMR	VERDURE	23	BOUL P8821	13N395		2A			
31175	BRUGGEMAN	VER	VAN ARTEVELDE	85	BOUL P8821	13N395		2A			
31075	DECREE	PMR	SENNE	88	BOUL P8821	13N395		2A			
3 979	DEKOSTER	EMR	RICHES CLAIRES	4	BOUL P8821	13N395		2A			
31159	DELCAMPE	DMR	VAN ARTEVELDE	3	BOUL P8821	13N395		2A			
3 357	DELEEBEECK	LMR	CAMUSEL	59	BOUL P8821	13N395		2A			
31353	DEMEULEMEESTER	EME	AUX GRAINS	31	BOUL P8821	13N395		2A			
31970	DERRE	CMR	VAUTOUR	18	BOUL P8821	13N395		2A			
3 464	DESMET	JMR	CUILLER	12	BOUL P8821	13N395		2A			
31236	DESMET	HMR	VAUTOUR	7	BOUL P8821	13N395		2A			
31891	FERNON	PMR	SENNE	17	BOUL P8821	13N395		2A			
31239	HOEBRECHTS	GMR	VAUTOUR	17	BOUL P8821	13N395		2A			
3 272	MORTELMANS	PMR	BODEGHEM	51	BOUL P8821	13N395		2A			
3 150	SPILLEBANT	AMR	ANDERLECHT	146	BOUL P8821	13N395		2A			
3 60	VAN PEYER	VMR	ANDERLECHT	47	BOUL P8831	13N395		2A			
3 419	VANDENEEDE	FMR	CHARTREUX	8	BOUL P8821	13N395		2A			
31495	VANDENWYNGAERT	J R	BODEGHEM	35	BOUL P8821	13N395		2A			
3 348	VANNURVEL	FMR	CAMUSEL	3	BOUL P8821	13N395		2A			
									19	31	72