

CAD

Notes de lecture : langue naturelle et métalangue en analyse documentaire

Les cahiers de l'analyse des données, tome 20, n° 1 (1995),
p. 121-126

http://www.numdam.org/item?id=CAD_1995__20_1_121_0

© Les cahiers de l'analyse des données, Dunod, 1995, tous droits réservés.

L'accès aux archives de la revue « Les cahiers de l'analyse des données » implique l'accord avec les conditions générales d'utilisation (<http://www.numdam.org/legal.php>). Toute utilisation commerciale ou impression systématique est constitutive d'une infraction pénale. Toute copie ou impression de ce fichier doit contenir la présente mention de copyright.

NUMDAM

Article numérisé dans le cadre du programme
Numérisation de documents anciens mathématiques

<http://www.numdam.org/>

NOTES DE LECTURE: LANGUE NATURELLE ET MÉTALANGUE EN ANALYSE DOCUMENTAIRE

[MÉTALANGUE]

J.-P. BENZÉCRI

Apparus il y a moins de 50 ans, les calculateurs électroniques ont aujourd'hui une puissance qui naguères eût pu offrir matière à des spéculations utopiques plutôt qu'à des projets; et l'on conjecture que les matrices de processeurs permettront bientôt de traiter les images avec autant de volubilité qu'on traite déjà les suites de caractères.

Pendant, s'il s'agit d'embrasser le contenu d'un grand corpus de documents [et nous nous bornerons à considérer les textes, laissant les images en réserve], certains déplorent que l'outil électronique ne soit pas un alambic propre à extraire, par lui-même, de la masse, la seule fraction que demande la pensée.

Un spécialiste d'une de ces branches des humanités, qu'on a cru vanter en les dénommant *sciences humaines*, a écrit sur ce thème des pages qui ont retenu notre attention. Les considérations générales y faisant écho aux souvenirs de la propre carrière de l'auteur, philosophie critique et autocritique sont mêlées de telle sorte qu'on ne peut réfuter celle-là sans acquiescer à celle-ci, en quelque manière. Mieux vaut donc que nos notes de lecture citent non l'homme, mais ses propos, desquels on aura ôté toute référence spécifique. Voici!

1 L'utopie

“...on opposait ... il y a trente ans ... la compilation et le collationnement justiciables de l'ordinateur aux tâches nobles de l'explication et de l'interprétation qui lui échappaient.”

“J'ai moi-même avancé des vues semblables dans les années cinquante, à l'appui d'un projet qui devait aboutir à la création ... d'un nouveau laboratoire ... chargé du traitement automatique de l'information [dans un domaine particulier]. Des chercheurs éminents ... s'opposaient au projet; vingt ans après, les banques de

données ont acquis droit de cité et l'on étonnerait aujourd'hui ... le monde des sciences humaines, si l'on s'attardait à reprendre, pour les réfuter, les arguments de cette opposition passée.”

“[Reste] que malgré ce ralliement général ... les promoteurs actuels jugent encore nécessaire de dresser un bilan des bénéfices attendus, dans des termes qui étaient les nôtres il y a 35 ans, mais que nous nous garderions ... de reprendre aujourd'hui.”

“Ainsi, à l'époque de nos premiers fichiers ... sur cartes perforées, nous leur donnions pour usage ... " que quiconque y trouve ... sans ambiguïté, la réponse à n'importe quelle question d'ordre factuel, si particulière soit-elle"; en évoquant "l'exhaustivité et l'objectivité" des données, nécessaires à l'accomplissement du miracle.”

“L'expérience s'est chargée de m'apprendre la vanité d'un tel pari, et la réflexion sa naïveté.”

2 Recherche documentaire

L'auteur cité entreprend d'illustrer par un exemple, extérieur à sa propre spécialité, les malentendus qui, selon lui, s'attachent à l'expression de "traitement en langage naturel".

“[On annonce]... "une banque exhaustive de données se rapportant à tous les sujets de la civilisation gréco-latine".”

“Exhaustive en quel sens? En extension ... avec quelque 3500000 références...; mais en compréhension aussi bien si l'on prend à la lettre la promesse d'un "accès à tous les sujets". Si l'utopie est devenue réalité, à quelle innovation le devons-nous ?”

Sont d'abord considérées la capacité et la célérité acquises par l'outil informatique. Mais c'est pour poser aussitôt, à juste titre, le problème de la consultation. Est-ce que pour "tout sujet", pour "n'importe quelle question":

“...l'infortuné consultant devra parcourir les 3500000 références ... afin de repérer dans cet océan de mots les îlots ... intéressant son propos.”

Finalement, les éditeurs du corpus reconnaissent qu'on n'accédera aux textes que "grâce à de très nombreux mots-clé":

“... l'automatisation porte sur ... la lecture ... le collationnement... Quant au traitement de l'information proprement dit ... il met en jeu une métalangue ... (code, thesaurus, langage documentaire, liste de descripteurs); ... d'où l'impropriété des références au "traitement en langage naturelle".”

À un tel traitement, l'auteur ne fait pas d'objection de principe: de la

compétence des analystes humains, il induit - sans toutefois entreprendre de démonstration philosophique - la possibilité d'un algorithme d'analyse automatique. Mais, sous le nom de "réflexivité", il demande, au préalable, une recherche introspective sur la pensée, évoquant ce qu'était la psychologie de l'intelligence, pour l'École de Wurtzburg, il y a un siècle.

"... loin de nous épargner toute recherche sur la représentation des connaissances dans nos domaines particuliers, les nouvelles technologies nous obligent à donner des réponses précises, pratiques, opératoires..."

Programme qui nous intéresse; mais dont nous n'acceptons pas qu'il soit proposé sans en attendre immédiatement de fruit; et comme si la faculté d'organisation de l'élaborateur électronique n'y avait pas le premier rôle.

En effet, sans disposer des corpus que certains laboratoires ont pris pour profession de rassembler, nous avons reconnu, dans les textes, un ordre des mots qui en exprime le contenu.

3 Analyse du vocabulaire et recherche du thème

D'abord, sans aucune construction de géométrie multidimensionnelle, on voit que les mots pleins les plus fréquents d'un texte donnent le contenu de celui-ci: ainsi d'après [CAD XII-XVII (2)], §3.1.1, (in CAD, Vol.XVIII, n°1; 1993), on citera un article intitulé "Compression des images polychromes et sensibilité au contraste chromatique", dont les mots pleins les plus fréquents (avec de 10 à 30 occurrences) sont:

{10<blanc, couleur, période, sin, ensemble, chromatique, stimuli, seuil, cos, fonction, luminance, couleurs, image=30}.

Or, tout en se prêtant à une élaboration progressive, la notion de mot outil est communément admise *a priori*: on peut donc, par soustraction, extraire automatiquement la liste ordonnée des mots pleins; ce qui constitue une indexation automatique.

Par là, on entreprendra d'interroger le corpus.

Une liste de mots pleins rangés d'après le nombre des textes où ils figurent parmi les dix plus fréquents, suggère à l'utilisateur, au "consultant", les termes qui peuvent être ceux de sa question. Mais, selon notre expérience, la question doit s'adresser à un espace organisé; non à un ensemble de textes, fussent-ils indicés.

Chacun admettra qu'une question particulière ne s'adresse qu'à une partie du corpus; même si celui-ci est renfermé dans un domaine tel que la "civilisation gréco-latine". Il faut donc segmenter le corpus. À cela, l'introspection, les hiérarchies logiques ne servent guère: car sans une

élaboration linguistique préalable, fondée sur l'analyse statistique multidimensionnelle, les cribles logiques qui viennent à l'esprit sont d'un si médiocre rendement que l'abondance des informations accable le lecteur plus qu'elle ne le sert (cf. [TAX. TEXTES], in *CAD*, Vol.XVII, n°3; 1993).

Au contraire, la distance du χ^2 donne à l'ensemble une structure d'espace; qu'on explore par l'analyse de correspondance, puis la CAH. Dès lors, on peut, sans s'égarer, aller aux textes les plus proches d'un texte donné.

[Sans négliger la recherche de simples paragraphes, qui, sur un thème qui n'est pas celui du texte entier, jettent parfois des clartés qu'on chercherait en vain dans les textes propres à ce thème-là.]

D'une expérience portant sur 200 articles publiés dans *CAD*, on se bornera à rappeler la conclusion (cf. [*CAD* XII-XVII (3)], §7, in *CAD*, Vol.XVIII, n°1; 1993):

Deux cents articles n'offrent pas une base suffisante pour définir la typologie sémantique d'un domaine de la science: mais nous ne voyons pas d'obstacle qui empêche de poursuivre dans la même voie pour élaborer un corpus d'un volume supérieur.

Nous savons, d'autre part, que la seule distribution des mots outil révèle le genre d'un texte, sinon son contenu.

L'expérience la plus étendue porte sur un corpus de 889 fragments en langue grecque; composé pour moitié de textes classiques et pour moitié de textes de la Bible, dont la totalité du Nouveau Testament.

La structure taxinomique obtenue, expliquée en détail dans trois articles, est rappelée succinctement, en conclusion, dans le §6 de [TEXTES GRECS 3.D] (in *CAD*, Vol.XIX, n°2; 1994).

Et l'on décrit, dans [DISCR. CAH LING.], l'adjonction, à ce corpus de base, de 178 de fragments supplémentaires dont la place est bien vue d'après l'analyse discriminante (i.e. par affectation de tout fragment ou œuvre supplémentaire au fragment ou à l'œuvre du corpus de base qui en est le plus proche, au sens de la distance distributionnelle): cf. *infra*, §5.

Ainsi, point n'est besoin d'une métalangue.

En suivant la structure naturelle de la langue, on accède au contenu bien plus sûrement que par une schématisation artificielle de la pensée et de son expression. La pensée réfléchit sur soi-même; mais par ce miroir d'objets que lui offre la synthèse mathématique; et se réalise ce que d'aucuns, en se riant, donnaient pour jeu à l'introspection: se mettre à la fenêtre pour se regarder passer dans la rue.

4 Analyse du contenu

“... faire apparaître la signification des textes ... par une analyse mécanisée de leurs constituants.”

Telle était l'ambition des années soixante ... Aujourd'hui:

“... l'analyse automatique n'est plus censée engendrer des résultats mais plutôt des hypothèses ... des embryons d'hypothèses qu'il faut encore élaborer en recourant à des données et à des opérations qui ne font pas partie de l'appareil formel.”

“La question doit être posée non seulement pour l'analyse de contenu mais pour toute application de l'ordinateur fondée sur le même projet heuristique... c'est-à-dire la recherche de structures ou d'arrangements variés ... dans des tableaux ou des matrices où sont rassemblés les données que l'on a retenues...”

“Méthodes mathématiques et algorithmes ne se comptent plus; un bon livre ... fait suivre le panorama des démarches par une évaluation critique des résultats dans une discipline particulière... Or l'ouvrage s'achève par un appel à des recherches sur ... les raisonnements des spécialistes dans la perspective de l'informatique mais sans grand chose de commun avec l'analyse du contenu.”

Nous approuvons l'auteur d'entendre "analyse de contenu" au sens le plus large, comme une analyse mathématique des données. Mais nous ne louerions pas sans réserve le livre qu'il cite. Il y a là, certes, des analyses; mais non des chaînes d'analyses. Or il est rare que l'analyse du premier tableau proposé montre plus que l'aptitude de la méthode à retrouver sur les premiers axes une structure déjà connue. L'analyse doit guider le choix des données; délimiter les domaines entre lesquels on cherchera des relations.

Afin de montrer ce que nous entendons par une telle stratégie de recherche, nous prenons pour exemples quatre études récentes qui ont en commun d'être toutes issues de la Grèce, patrie des humanités classiques.

5 Étapes dans la recherche en analyse des données

Dans [TECHN. ENS. GRÈCE] (in *CAD*, Vol.XIX, n°3), on analyse les réponses de 650 enseignants grecs interrogés au début d'une session de formation à l'informatique et aux nouvelles technologies. Le questionnaire peut être divisé en trois parties principales: A) Signalement: âge-sexe, Spécialité, Université d'origine, Carrière...; B) Connaissance des nouvelles technologies; C) Rôle potentiel attribué à l'informatique. La partie la plus solide étant B), on l'analyse d'abord. On vérifie ensuite que le niveau des connaissances est en rapport étroit avec le signalement. Quant aux perspectives, il apparaît que les stagiaires, à quelques exceptions près, adoptent des suggestions générales, qui ne sont pas étroitement liées à leur expérience personnelle, généralement très modeste.

Dans [LAMPES SAMOS] (in *CAD*, Vol.XIX, n°3), on considère 170 objets, décrits suivant 22 variables discrètes. Après diverses analyses globales, l'analyse d'un sous-tableau rectangulaire du tableau de BURT, montre, entre les modalités de décor et les autres modalités descriptives, une corrélation assez forte pour que les classifications des objets fondées sur ces deux groupes de modalités s'accordent à reconnaître les mêmes quatre types principaux.

Un corpus C de 889 fragments de textes grecs classiques a fait l'objet d'une série d'analyses stylistiques fondées, principalement, sur les fréquences d'emploi des mots outil (cf. *supra*, §3). Dans [DISCR. CAH LING.] (in *CAD*, Vol.XX, n°1), on entreprend de rapporter à ce corpus le traité des "Catégories"; œuvre qui est considérée comme la 1-ère partie de l'Organon - ou logique - d'ARISTOTE; Organon dont un autre texte, le *de Interpretatione*, figure dans le corpus C. Il apparaît que les fragments des "Catégories" vont avec les textes philosophiques de C, notamment ceux d'ARISTOTE; mais ils n'ont pas d'affinité nette avec le *de Interpretatione*. Résultat qui s'accorde avec l'hypothèse, que le style des Catégories n'est pas de la plus grande rigueur formelle.

Dans une étude à paraître, Th. BEHRAKIS et St. PAPASTAMOU, comparent, aux résultats des élections du 12-6-1994 pour le Parlement Européen, les réponses de 3000 électeurs grecs à un sondage ayant précédé le scrutin de quelques semaines. Après avoir recodé les notes, données sur des échelles d'évaluation, en estimant l'équation personnelle de chaque sujet, les auteurs soumettent les réponses à l'analyse des correspondances et à la CAH; et obtiennent finalement, par analyse discriminante, une estimation vraisemblable des tendances de ceux des électeurs qui, lors du sondage, se sont déclarés indécis.

6 Envoi

Blasé, l'auteur de l'article, objet de la présente note, a dit que "méthodes et algorithmes ne se comptent plus". Pour arbitrer entre ceux-ci, sans doute faudrait-il les avoir pratiqués tous, en s'appliquant à tirer de chacun le meilleur parti possible. Ce que personne n'a jamais fait. Reconnaître la structure d'un tableau 12×8 , qui exprime une simple antithèse ou une gradation manifeste d'un pôle à un autre, n'est pas un exploit qui suffise à attester l'intérêt d'une méthode: or c'est à de tels exemples que se bornent ordinairement les exposés pédagogiques destinés aux praticiens des disciplines qui collationnent d'amples données.

Mais s'il est vrai que l'analyse multidimensionnelle ne se fait connaître qu'en soumettant les données à une suite d'opérations qui sont l'analogue spatial d'un raisonnement logique, c'est d'après de telles suites qu'on doit apprécier les méthodes.

À ce prix seulement, le géomètre tient les promesses des autres.