Communication et langages

L'analyse automatique des textes : information - analyse - action

Pierre Raynaud

Résumé

Depuis plus de quinze ans, Pierre Raynaud est conseil en communication. Il est l'inventeur de la méthode dite de communication directive. Il a publié en 1977 L'art de manipuler*, ou éléments de communication directive. Cette méthode permettant de modifier autrui repose entre autres sur une technique d'analyse de langage. On a déjà lu dans* Communication et langages *de mars 1977 une analyse du livre* Démocratie française *de Valéry Giscard d'Estaing. Il nous présente aujourd'hui la dernière version de son système d'analyse automatique, logiciel expert nommé anaexpert, au travers d'un exemple réel établi au Cetelem.*

L'ANALYSE AUTOMATIQUE DES TEXTES : INFORMATION — ANALYSE — ACTION

par Pierre Raynaud

Depuis plus de quinze ans, Pierre Raynaud est conseil en communication. Il est l'inventeur de la méthode dite de communication directive. Il a publié en 1977 *L'art de manipuler, ou éléments de communication directive*. Cette méthode permettant de modifier autrui repose entre autres sur une technique d'analyse de langage. On a déjà lu dans *Communication et langages* de mars 1977 une analyse du livre *Démocratie française* de Valéry Giscard d'Estaing. Il nous présente aujourd'hui la dernière version de son système d'analyse automatique, logiciel expert nommé ANAEXPERT, au travers d'un exemple réel établi au Cetelem.

LES PROBLÈMES DE LA COMPRÉHENSION DU LANGAGE NATUREL

C'est un des sujets principaux de l'Intelligence Artificielle (IA). Depuis vingt ans les chercheurs dans ce domaine rencontrent toujours les mêmes difficultés, et celles-ci sont amenées par la façon même dont le problème est posé. En effet la proposition : « compréhension du langage naturel » contient en elle-même les problèmes qu'elle prétend résoudre.

Que signifie : « langage naturel » ? On sait, depuis le début de ce type de recherches (voir Chomsky), que le temps n'est pas venu où l'homme sera capable d'engendrer un logiciel expert pouvant comprendre *tout*. Tout ce qui se dit et qui peut se dire ; la langue et la parole. Alors l'alternative est simple : ou bien tenter de créer un système général qui « comprendra » un peu de tout ; ou limiter strictement le sujet et créer un système comprenant presque tout de peu. Les tentatives de la première catégorie, vastes programmes d'Etat ou d'universités n'ont pas abouti, comme on pouvait s'y attendre. La seconde solution a déjà abouti à quelques réalisations commerciales.

Il est évident que nous avons choisi la deuxième solution en étudiant des sujets précis, comme nous le verrons.
De notre point de vue, il s'agit d'une fausse alternative, et, cela pour deux raisons :
— premièrement, il n'est pas de logiciel ni de système-expert sans utilisateur. Et l'utilisateur ne se pose pas de problèmes généraux. La bonne démarche sera toujours en deux temps : étudier le besoin précis de l'utilisateur (donc les limites du langage à étudier), puis construire un système à son usage ;
— deuxièmement, la compréhension générale du langage naturel ne pourra survenir que plus tard, qu'après avoir atteint la maîtrise de systèmes plus modestes, par une sorte de généralisation-synthèse de tous les problèmes rencontrés lors de la construction de ces systèmes. A quel horizon ? Certainement pas avant l'an 2000.

Que signifie : « compréhension » ? Comment comprenons-nous les textes que nous lisons ? Comment, c'est-à-dire : par quels processus comprenons-nous, et que comprenons-nous ? Le premier constat est que (dans la mesure où nous connaissons la langue utilisée et le sujet) tout se passe comme si nous comprenions un texte « instantanément » et globalement sans nous préoccuper avant de relever les mots, les expressions, les thèmes, ni la construction de la phrase. C'est-à-dire que nous comprenons ce que nous lisons sans, apparemment, faire appel aux analyses (lexicales, thématiques, syntaxiques) ; or c'est justement cela, et cela seul, que l'ordinateur sait faire.
A ce stade de notre exposé, on ne peut s'empêcher de penser que, peut-être, le problème a été posé à l'envers. Car enfin ce sont bien ces mêmes hommes qui ont inventé l'ordinateur et qui ne peuvent lui faire faire ce qu'ils font eux-mêmes spontanément. Et si l'IA n'était pas une matière pour les informaticiens, mais pour les épistémologues ? Rêvons un instant à quelque individu pour qui les prémisses du problème seraient dans notre façon de comprendre et qui saurait ensuite transcrire cette façon en termes de machine.

« Comprendre » pose encore un autre problème. C'est qu'il n'y a pas deux personnes pour comprendre de la même façon le même texte. Le texte lui-même n'est qu'un ensemble de caractères (signifiants) et chaque lecteur y perçoit un ensemble de signifiés légèrement différents. Dans ce cas, que faut-il apprendre à Monsieur l'ordinateur : la compréhension de Pierre ou celle de Paul ?

En ce qui nous concerne, ces problèmes nous sont inconnus dans la mesure où nous avons privilégié l'utilisateur et non le logiciel. La bonne compréhension de l'ordinateur est celle qui se rapproche le plus de celle de l'utilisateur. Et foin des polémiques !

LE LOGICIEL ANAEXPERT

La meilleure façon d'expliquer le système est de décrire une installation réelle et d'en conter l'histoire. Nous avons choisi de prendre comme exemple notre expérience au Cetelem[1].

Depuis longtemps déjà, le service Consommateurs du Cetelem, se préoccupe de connaître l'opinion des clients, et leur degré de satisfaction. Aussi, envoie-t-il à tous les nouveaux clients venant d'obtenir un crédit pour un achat dans un magasin, un mini-questionnaire, le CETELEGUIDE, comportant la possibilité pour le client d'exprimer librement ses observations.

L'analyse de ces observations, en continu, jusqu'alors impossible, a été prise en charge par le système ANAEXPERT, au cours d'une opération en trois étapes que nous appellerons :

— *Les mots et expressions,*
— *Les thèmes,*
— *Les règles.*

LEXIQUE ET ANTI-LEXIQUE

Dans un premier temps, le logiciel a pris connaissance du vocabulaire utilisé par les clients. A chaque fiche entrée au clavier, Anaexpert demande ce qu'il doit faire des mots et expressions nouvelles qu'il ne connaît pas encore.

Ainsi, petit à petit, se forment deux *lexiques* fondamentaux : le lexique de base du langage (ou mots considérés comme intéressants par l'utilisateur) et l'anti-lexique (ou mots n'ayant pas été pris en compte). Au fur et à mesure que l'on engrange des données sur un sujet précis, comme celui-ci, on s'aperçoit que les mots nouveaux diminuent jusqu'à tendre vers zéro. On peut donc raisonnablement penser qu'au bout d'un certain volume de textes, (ici dès le deuxième mois, dès le 500e client), nous sommes en possession de la quasi-totalité des mots et expressions pertinents au regard de l'utilisateur.

1. Je tiens particulièrement à remercier ici mon client et ami Jacques Lanoë et tout le Cetelem, pour m'avoir autorisé à publier les résultats *authentiques* du mois de décembre 1986.

Actuellement le Cetelem possède un lexique complet d'environ 1500 mots et 300 expressions, et les mots nouveaux qui apparaissent encore de temps à autre, ne sont que des dérivés et variantes de mots déjà connus. Naturellement, ces 1500 mots ne sont pas de la même importance quantitative. Si on les classe par ordre de fréquence décroissante, on obtient une vérification de la fameuse loi de Zipf[2] qui ressemble quelque peu à la pseudo-loi des 80/20.
Ici, dans l'exemple qui est le nôtre on trouve que, au mois de décembre 1986, sur 300 clients :
— 125 mots sur 1200 sont dits par 3 personnes au moins (1 %),
— 23 mots par 9 personnes ou plus (3 %),
— 5 mots seulement par 30 personnes (10 %).

Si l'on examine combien de clients ont prononcé au moins 1 des 5 mots clés principaux, on en trouve plus des 2/3 !
Quels sont les 23 mots clés fondamentaux des clients Cetelem (par ordre décroissant d'importance) : Je, crédit, Cetelem, carte, Aurore, achat, client, satisfaisant, toujours, vendeur, magasin, problème, maison, dossier, ans, content, accepté, beaucoup, contente, services, accueil, satisfaite, jamais.

L'ANALYSE THÉMATIQUE

Une grande originalité du système ANAEXPERT est de privilégier l'analyse thématique sur l'analyse dite syntaxique. C'est en effet une approche rare en intelligence artificielle. Mais qu'est-ce qu'un thème ?

Selon notre système, un mot en soi ne signifie rien en dehors de son utilisation, c'est-à-dire en dehors des autres mots qui se trouvent dans son contexte (co-occurence) et des mots pouvant le remplacer (commutativité), au sein de la même phrase.
Donnons un exemple. Soit la phrase d'un client Cetelem : « Je suis content du Cetelem depuis toujours ».
Le mot « content » peut être analysé de deux façons :
— soit comme lié à « je », à « Cetelem » et à la notion de temps « depuis toujours » = « ancien client » ;
— soit comme pouvant être remplacé par « satisfait », « mécontent », « décu »...

Nous disons que, d'une certaine façon, tous les mots et expressions pouvant commuter entre eux au sein des mêmes phrases

2. Voir G. Zipf, *La psychobiologie du langage*, Retz, 1974.

appartiennent au même thème et donc, par là même, le définissent. Bien qu'admettant des nuances et des exceptions, cette proposition est globalement acceptable.

C'est ainsi qu'une partie importante des mots et expressions du lexique, pourront être classés dans des tiroirs, ou boîtes thématiques, ou micro-thèmes.
Dans notre cas, l'utilisateur a choisi, avec nous-mêmes, après analyse des phrases des clients, de regrouper les mots en 66 thèmes fondamentaux dont voici les principaux, (les chiffres sont le nombre de clients sur 300 ayant évoqué ces thèmes) :

PRINCIPAUX THÈMES

Le Cetelem (mot et synonymes)	183
Je, moi	145
Satisfait (en général)	72
Ancien client	52
Les achats (mot et synonymes)	47
La carte Aurore	34
Demandes générales	33
Remerciements (grâce à...)	28
Les magasins (mot magasin et synonymes)	22
Les vendeurs (mot et synonymes)	21
Achats cités	21
Facile, simple	18
L'accord du crédit, obtention	16
Le dossier crédit Cetelem	15
Les magasins cités	15
Client Cetelem	15
Demande sur carte Aurore ou Cetelem	14
Le taux du crédit	12
Remboursements, mensualités	12
Sympathique	11
Rien à dire, aucune remarque	11
Les renseignements personnels	11
La famille, les amis	10

A partir du moment où le logiciel analyseur « comprend » ce qu'est un thème, il se comporte déjà comme un humain. Par exemple, il connaît 54 façons de se déclarer satisfait du Cetelem, 99 façons de se déclarer « ancien client », 20 façons de dire qu'il n'y a rien à signaler, 6 façons de parler du mauvais accueil dans le magasin...
A cette étape de la construction du système, nous pouvons, pour continuer l'apprentissage, donner au logiciel des textes nouveaux « à manger », et observer son comportement. Or qu'observons-nous ? Des oublis et des erreurs... Si l'analyse sémantique et thématique a été correctement effectuée, le logiciel à ce stade « comprend » 80 % du texte.

IMPORTANCE DES RÈGLES SE RAPPORTANT AU STYLE

Parmi les mots du lexique de base un nombre très important de mots n'ont pu être classés dans un thème unique. Ce sont les mots « ambigus » qui signifient différemment selon le contexte. Or, certains de ces mots sont fondamentaux pour notre étude. Comment ignorer par exemple le mot « client » ? Pour le Cetelem, ce mot signifie très différemment selon qu'il s'agit du *client Cetelem* ou *du client d'un magasin.*
Aussi faut-il donner au logiciel, dans sa base de connaissance, une série de règles de comportement du style :
« Si, dans une phrase, nous trouvons le mot client *et* soit *Cetelem* ou *votre maison,...* et que nous ne trouvons ni *magasin,* ni *vendeur,* ni *achat,* alors il s'agit de *client Cetelem* sinon... »
La combinaison des *et* et des *ou* peut très rapidement rendre ces règles complexes.

Les règles aboutiront « nécessairement » à une des deux décisions fondamentales :
— ou bien coder un des thèmes possibles,
— ou bien ne rien coder (car l'ambiguïté subsiste par la présence ou l'absence de deux groupes de thèmes).

La nature et le nombre de règles que l'on peut imposer à un système d'analyse thématique sont quasi illimités. Elles n'ont jamais un statut universel, mais dépendent étroitement du sujet traité. Leurs buts peuvent être multiples :
— résoudre les ambiguïtés,
— éliminer les contradictions,
— établir des déductions logiques...

A ce stade, le logiciel comprend 95 % du texte qui lui est présenté.

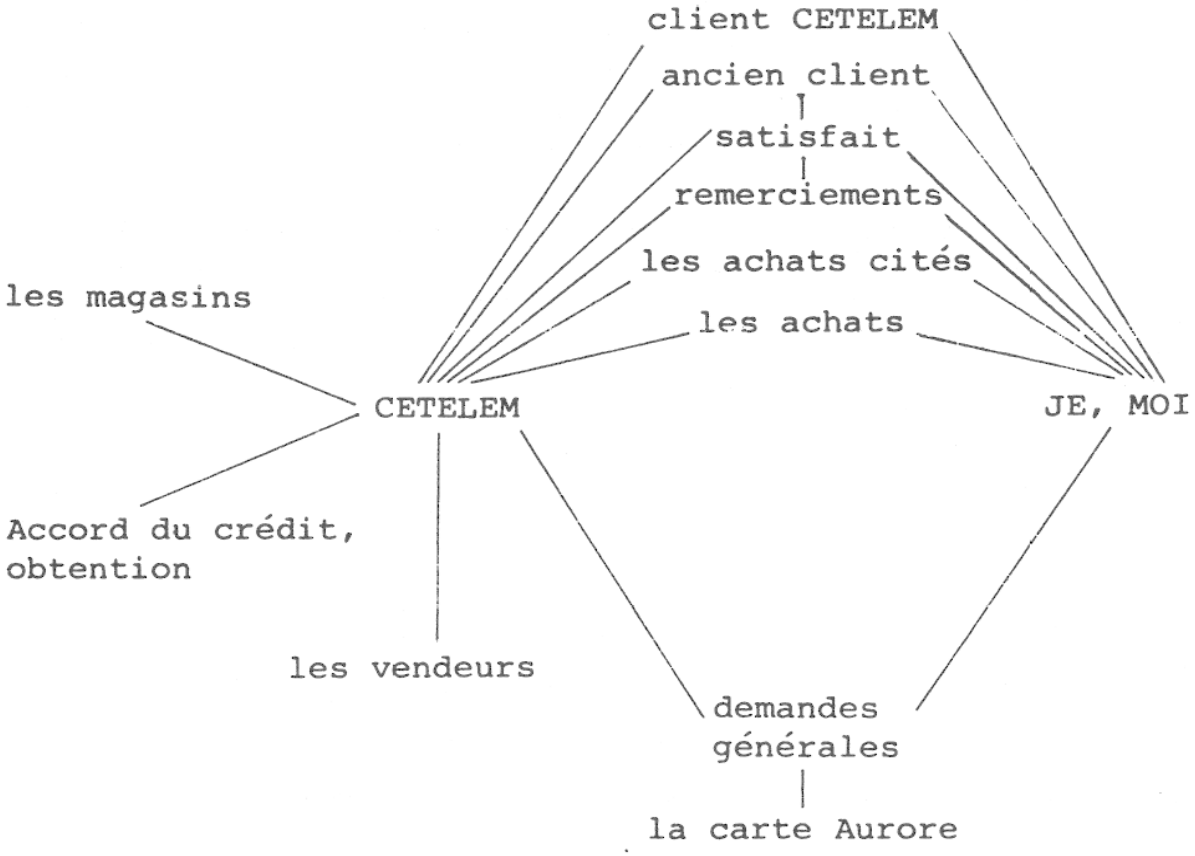
RÉSULTATS DE L'ANALYSE

Les résultats d'une analyse linguistique paraissent souvent sybillins à l'utilisateur, c'est-à-dire à l'esprit humain qui raisonne en termes de synthèse et d'interprétation et non en termes d'analyse scientifique.
Bien qu'ayant essayé d'apprendre au logiciel à « raisonner » comme un humain, les résultats restent suffisamment étranges, voire étrangers, pour faire réfléchir l'utilisateur d'une façon nouvelle.

Montrons par exemple ce que nous appellons un *graphe thématique.* Ce graphe est constitué de l'ensemble des thèmes qui

apparaissent ensemble chez plus de X % des clients. Il est à proprement parler le résumé des textes, au plan thématique.

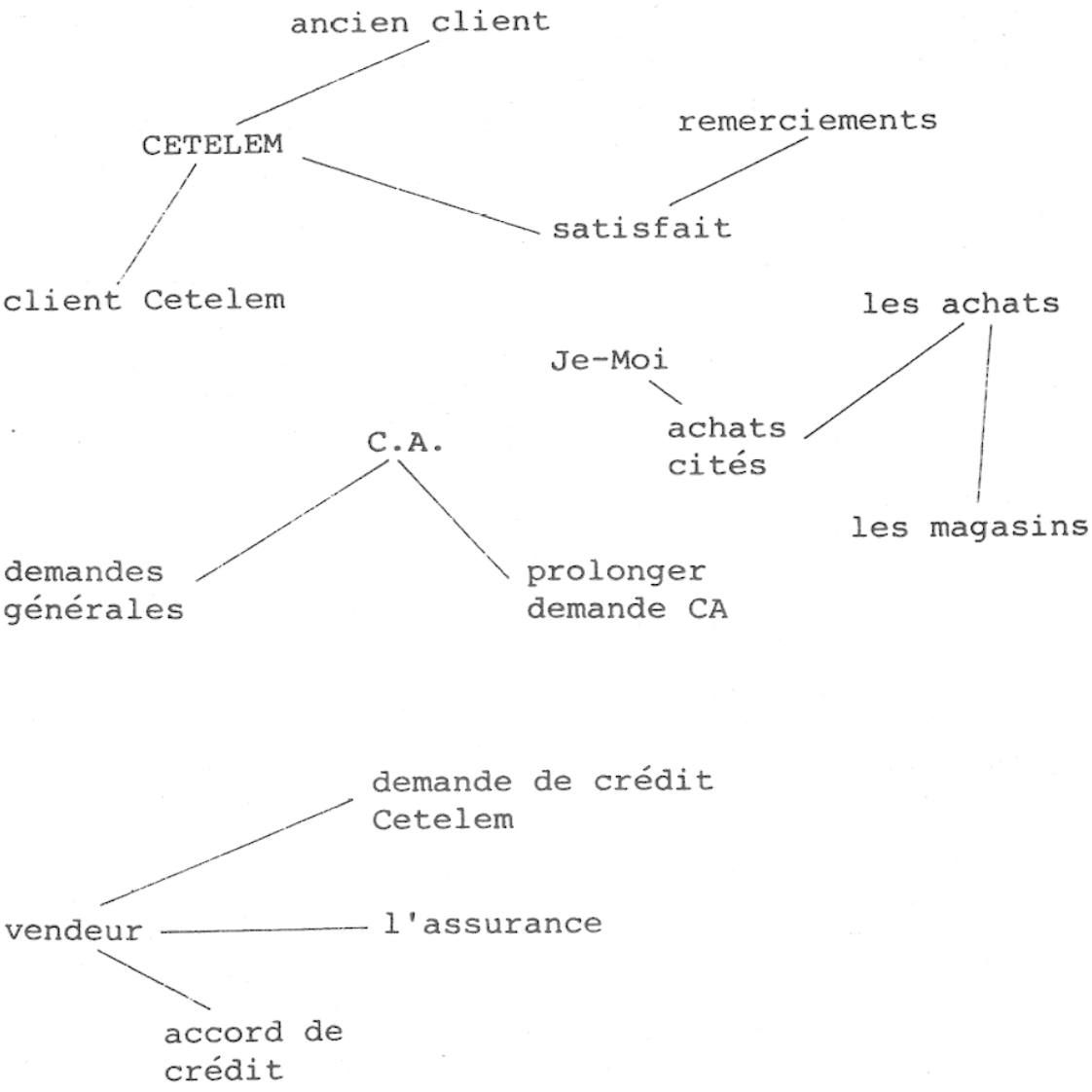
Voici le graphe des clients Cetelem en décembre 1986.



Globalement, sans entrer dans le détail, nous observons que :
— les thèmes « sujets » : Cetelem et Je, sont liés à la *satisfaction*, à la notion d'*ancien client* et aux *achats* ; mais que d'autres thèmes semblent avoir une vie autonome : les vendeurs et les magasins d'un côté, la carte Aurore (carte de crédit) de l'autre.

Un autre type d'analyse thématique consiste à observer si certains thèmes n'ont pas tendance, soit à s'attirer, soit à se repousser.

Ainsi, en comparant la fréquence des co-apparitions des thèmes, à ce que serait la fréquence issue d'un tirage aléatoire, nous obtenons ce que nous appelons le *graphe d'attraction thématique* suivant.



Ce graphe fait apparaître de façon évidente que le langage des clients Cetelem (en décembre 1986) peut se subdiviser en quatre sous-langages indépendants qui peuvent ainsi se décrire :

1. Je suis un ancien client du Cetelem satisfait ; je vous remercie.
2. J'ai fait des achats dans des magasins.
3. Les vendeurs au moment de la rédaction du dossier de crédit.
4. J'ai des demandes au niveau de la carte Aurore.

On touche ici, une des plus grandes utilités de ce type d'analyse :

— savoir dans quelle mesure un langage est homogène et donc...
— établir une partition sémantiquement juste du langage parent.

UTILITÉS DE L'ANALYSE AUTOMATIQUE DES TEXTES

Nous pourrions continuer cet article en décrivant plus longuement les cas et exemples rencontrés en analyse thématique. Nous pourrions également évoquer les prochains développements que nous entrevoyons pour ce genre de logiciels-experts. Mais nous préférons montrer, en conclusion, quelques utilisations possibles de ce type de logiciels, arbitrairement classées en quatre catégories :

1. Les utilisations commerciales viennent spontanément à l'esprit, car, toutes les entreprises ont intérêt, comme le Cetelem, à connaître l'opinion de leurs clients, et de leurs employés, afin de mieux communiquer avec leur environnement et leurs structures ;

2. Ce type de logiciel permet également d'analyser des textes d'auteurs différents, ou d'époques différentes, et donc peut être une aide précieuse aux recherches littéraires ou stylistiques ;

3. Le but poursuivi par l'utilisateur d'un logiciel-expert en analyse de textes, peut être de nature « policière ». Par exemple : reconnaître l'auteur (ou le criminel) qui se cache derrière un fragile anonymat ; ou bien rechercher, au-delà du dit conscient d'un orateur ce qu'il a vraiment dit en croyant le cacher...

4. On peut également envisager une utilisation personnelle de ce type d'analyse pour apprendre à se mieux connaître, pour mieux cerner sa propre vision du monde...

D'une façon générale, l'analyse automatique des textes est une discipline rarement utilisée seulement dans le but de mieux connaître un sujet, mais surtout pour obtenir une modification de ce sujet. Connaître pour mieux maîtriser.
Le système ici présenté s'inscrit comme un élément indispensable dans la chaîne : *Information-analyse-action.*
Ce système pourrait être le premier représentant d'une nouvelle sorte d'outils pour l'action, que nous appellerons provisoirement des *analyseurs de faits*. Outils qui deviendront bientôt indispensables à tout homme d'action qu'il soit chef d'entreprise, homme politique, journaliste, enseignant...

P. Raynaud