Document généré le 18 juin 2025 17:50

Documentation et bibliothèques

L'indexation automatique : état de la question et perspectives d'avenir

Yves Courrier

Volume 23, numéro 2, juin 1977

URI : https://id.erudit.org/iderudit/1055247ar
DOI : https://doi.org/10.7202/1055247ar

Aller au sommaire du numéro

Éditeur(s)

Association pour l'avancement des sciences et des techniques de la documentation (ASTED)

ISSN

0315-2340 (imprimé)
2291-8949 (numérique)

Découvrir la revue

Citer cet article

Courrier, Y. (1977). L'indexation automatique : état de la question et perspectives d'avenir. *Documentation et bibliothèques*, *23*(2), 59–72. https://doi.org/10.7202/1055247ar

Résumé de l'article

L'auteur décrit d'abord le principe de l'édition automatique d'index, en particulier les index permutés et le système PRECIS, et souligne les tendances actuelles. Il expose ensuite les diverses méthodes utilisées pour extraire automatiquement les mots-clés : les méthodes statistiques et les méthodes linguistiques. Pour ces dernières, il explicite les méthodes syntaxiques et les méthodes sémantiques. Il montre enfin que la recherche en intelligence artificielle est pertinente pour l'indexation automatique mais qu'elle n'est pas encore assez développée pour apporter des résultats substantiels.

Tous droits réservés © Association pour l'avancement des sciences et des techniques de la documentation (ASTED), 1977

Ce document est protégé par la loi sur le droit d'auteur. L'utilisation des services d'Érudit (y compris la reproduction) est assujettie à sa politique d'utilisation que vous pouvez consulter en ligne.

https://apropos.erudit.org/fr/usagers/politique-dutilisation/

Cet article est diffusé et préservé par Érudit.

Érudit est un consortium interuniversitaire sans but lucratif composé de l'Université de Montréal, l'Université Laval et l'Université du Québec à Montréal. Il a pour mission la promotion et la valorisation de la recherche.

https://www.erudit.org/fr/

L'indexation automatique: état de la question et perspectives d'avenir

Yves Courrier
École de bibliothéconomie
Université de Montréal

L'auteur décrit d'abord le principe de l'édition automatique d'index, en particulier les index permutés et le système PRECIS, et souligne les tendances actuelles. Il expose ensuite les diverses méthodes utilisées pour extraire automatiquement les mots-clés: les méthodes statistiques et les méthodes linguistiques. Pour ces dernières, il explicite les méthodes syntaxiques et les méthodes sémantiques. Il montre enfin que la recherche en intelligence artificielle est pertinente pour l'indexation automatique mais qu'elle n'est pas encore assez développée pour apporter des résultats substantiels.

The principle of the automatic production of indexes is first described, more particularly rotated index and the PRECIS system. Then the various methods used to extract automatically key-words are described: statistical methods and linguistic methods. For the later, syntactic methods and semantic methods are more precisely described. Finally, research in artificial intelligence is shown to be pertinent for automatic indexing, but not developped enough yet to bring substantial results.

Se trata en este artículo de la producción automática de índice, en particular de los índices permutados y del sistema PRECIS así que de las tendencias actuales en este asunto. El autor presenta los diversos métodos empleados para extraer automáticamente las palabras claves de los textos: los métodos estadísticos y los métodos linguísticos. El autor desarrolla estos últimos exponiendo los métodos sintácticos así que los métodos semánticos. En fin demuestra que las investigaciones hechas en el campo de la inteligencia artificial se revelan pertinentes con respeto al indizado automatizado pero no todavía bastante desarrolla das para llevar unos resultados substanciales.

Au début des années soixante-dix, paraissaient quelques livres et articles de périodiques traitant de la linguistique et des sciences de l'information. Parmi les ouvrages, on peut citer Coyaud[1], Sparck-Jones

1. Maurice Coyaud, *Linguistique et documentation*, Paris, Larousse, 1972.

et Kay[2], ainsi que Bély[3]. Parmi les articles les plus importants, mentionnons Gardin[4] et Montgomery[5].

En fait, cette documentation traite en grande partie d'un problème plus étroit que les rapports de la linguistique et de la documentation: le sujet qu'on y aborde vraiment se limite plutôt à l'utilisation des connaissances linguistiques actuelles dans l'indexation automatique. On peut se demander si le temps n'est pas venu de synthétiser, de clarifier, de faire le point sur une documentation à la fois abondante et difficile. Le recul du temps devrait également permettre une appréciation plus sereine.

Peut-on, comme on semble souvent le croire, confier aux ordinateurs les tâches d'indexation? En quoi l'indexation automatique peut-elle changer le travail des documentalistes? C'est pour répondre à ces questions que nous allons examiner brièvement quelques travaux récents en ce domaine, tant du côté des réalisations pratiques que du côté de la recherche.

Pour bien comprendre ce qui va suivre, il est essentiel de réaliser que lorsqu'on parle d'indexation automatique, on peut faire allusion à deux opérations complètement différentes: soit l'édition d'index par imprimante d'ordinateur, soit le choix de mots-clés pour représenter le contenu des documents. Dans le premier cas, il s'agit d'une opération relativement simple, strictement mécanique, et donc facile à réaliser sur ordinateur. Dans le second cas, il s'agit de l'identification des concepts importants d'un document, donc d'une opération intellectuelle dont on connaît mal encore les mécanismes, et donc très difficile à simuler avec un ordinateur.

La distinction est capitale: d'une part travail de tabulation et d'édition, d'autre part opération d'analyse intellectuelle. Il est regrettable qu'on emploie le même terme d'indexation automatique pour ces deux tâches, d'autant plus que, comme on le verra, les techniques et les réalisations sont très différentes. Cette simplification n'est explicable que parce que, dans tous les cas, on aboutit à un index ou une classification éditée à l'aide de l'ordinateur. Mais lorsqu'on veut examiner de plus près l'ensemble des opérations effectuées, on doit absolument maintenir la distinction établie. On étudiera donc, dans une première partie, l'édition automatique d'index et, dans une seconde partie, l'analyse automatique.

2. Karen Sparck-Jones and Martin Kay, *Linguistics and Information Science*, New York, Academic Press, 1973.
3. N. Bély et al., *Procédures d'analyse sémantique appliquées à la documentation scientifique*, Paris, Gauthier-Villars, 1970.
4. Jean-Claude Gardin, «Document analysis and linguistic theory», *Journal of Documentation*, vol. 29 (June 1973), 137-168.
5. Christine-A. Montgomery, «Linguistics and information science», *Journal of the American Society for Information Science*, vol. 22, no. 2 (May-June 1972), 195-219.

L'édition automatique d'index

Description

Lorsqu'on veut décrire un document, on peut utiliser deux catégories d'éléments.

La première catégorie comprend les éléments signalétiques: auteur, titre, source, illustration, etc. Ces éléments, fait remarquer Vickery, réfèrent à l'origine du document. Ce sont des étiquettes qui concernent l'extérieur du document. La deuxième catégorie, au contraire, comprend les éléments analytiques: indices de classification, descripteurs, résumés, etc. Ces éléments réfèrent au contenu du document, aux sujets abordés. Ces étiquettes concernent l'intérieur du document. L'édition automatique d'index est tout simplement la manipulation de ces deux catégories d'éléments fournis au préalable à l'ordinateur.

Dans une bibliographie comme *RADAR,* par exemple, les textes des revues sont indexés par des spécialistes. On remplit ensuite un bordereau qui contient les données signalétiques et les données analytiques. On introduit toutes ces informations dans l'ordinateur. Ce dernier fait d'abord un travail de tri: il regroupe les références qui se rattachent au même mot-clé; puis il reste un travail d'édition: on édite l'index alphabétique par sujets, par matières, par auteurs, suivant le cas. Il est donc évident que l'ordinateur n'accomplit que des tâches strictement mécaniques. L'indexation est faite «manuellement», c'est-à-dire par des indexeurs qui lisent les textes et décident des mots-clés. L'ordinateur ne fait que fournir ses grandes capacités de manipulation et de tri qui viennent compléter le travail d'indexation fait auparavant.

L'ordinateur n'aide en rien à résoudre les problèmes traditionnels de l'indexation, notamment le choix des descripteurs, la profondeur de l'analyse et les renvois. Il ne fait qu'accélérer le processus d'édition et, dans certaines conditions, le rendre moins coûteux. Les réalisations dans ce domaine sont innombrables et il serait aussi hasardeux que futile d'en vouloir donner une liste exhaustive.

Il faut cependant citer deux cas assez intéressants d'applications du même type qui démontrent comment on peut utiliser au mieux les capacités de l'ordinateur.

Index permutés

Mis au point par H.P. Luhn en 1957, les index permutés constituent peut-être la méthode la plus employée de publication automatique d'index. Le principe consiste à utiliser les mots significatifs des titres des entrées bibliographiques. Ces mots sont classés par ordre alphabétique et accompagnés du titre entier de façon à placer le mot significatif dans son contexte. En regard de chacune des entrées ainsi créées, on mentionne un numéro d'ordre qui permet de retrouver la notice complète (auteur, titre, source) dans une liste généralement par ordre alphabétique d'auteurs. Le procédé est extrêmement simple et

peu coûteux, d'autant plus que présentement presque tous les fabricants d'ordinateur peuvent fournir un programme tout fait qui réalise ce genre d'index.

Les avantages sont donc évidents: facilité d'utilisation et coût très bas. Il faut cependant signaler certains inconvénients: du fait qu'un titre réapparaît autant de fois qu'il contient de mots significatifs, on aboutit assez rapidement à des index très lourds à manier. De plus, il n'est pas toujours facile de distinguer les termes significatifs de ceux qui ne le sont pas. Toutefois le système fonctionne bien si les titres sont effectivement significatifs; ce qui est généralement le cas dans les sciences exactes; mais il y a beaucoup d'exceptions dans les humanités et dans les sciences sociales. Pour ces différentes raisons, les index permutés sont surtout utilisés dans les petits centres de documentation desservant une population scientifique restreinte. Ils fournissent un service d'information courante rapidement et à peu de frais.

PRECIS

PRECIS est lui aussi un système en deux étapes: indexation manuelle suivie de l'édition par ordinateur. Mais il s'agit d'un système d'une toute autre ampleur. Au lieu d'éditer — comme avec un index permuté — un index restreint dans un domaine très spécialisé, il s'agissait de produire l'index d'une bibliographie nationale (la *British National Bibliography*). On devait donc faire face à une documentation très abondante couvrant tous les domaines du savoir. Un index permuté était évidemment inutilisable, puisqu'il aurait généré un trop grand nombre d'entrées et que trop de titres n'étaient pas significatifs. Une indexation manuelle traditionnelle, améliorée d'une édition automatique, n'aurait pas non plus répondu aux besoins, car on voulait à la fois exprimer des sujets complexes et éviter — en entourant les descripteurs de leur contexte — des entrées incompréhensibles ou ambiguës. Ces deux raisons nécessitaient l'élaboration d'un système qui permette d'ajouter à chaque mot-clé d'autres mots-clés pour lui servir de contexte. Ainsi un ouvrage sur les migrations d'oiseaux en Chine aurait pour descripteur principal *migration* qui représente le sujet du livre, *oiseaux* et *Chine* étant les autres concepts utilisés pour donner un contexte.[6]

De plus, comme on peut aussi désirer accéder au document sous les aspects spécifiques du sujet, il fallait aussi que les descripteurs décrivant le contexte puissent être utilisés comme entrées principales: dans l'exemple cité, *oiseaux* et *Chine.* On devait donc considérer tour à tour chacun des descripteurs comme le concept principal, les autres termes restants servant alors de contexte. Dans notre exemple, on aurait:

CHINE. Oiseaux. Migration
OISEAUX. Chine. Migration

Ces permutations peuvent facilement entraîner deux problèmes principaux. Le premier tient à la multiplication des entrées. Une chaîne

6. Guy Dionne, «PRECIS I: Preserved Context Indexing System», *Documentation et bibliothèques*, vol. 21, no 1 (mars 1975), 9-21.

de trois descripteurs peut engendrer six entrées, une chaîne de quatre descripteurs, vingt-quatre; avec cinq descripteurs, on pourrait «générer» cent vingt combinaisons différentes. Il fallait donc éviter que toutes les permutations soient permises.

Le deuxième problème peut facilement être illustré: un ouvrage sur «l'administration des bibliothèques d'enseignement» risquait d'être interprété comme signifiant «l'enseignement de l'administration dans les bibliothèques» ou «les bibliothèques sur l'administration de l'enseignement». Il fallait donc trouver un système qui permette d'éliminer, parmi toutes les combinaisons possibles des descripteurs, celles qui n'ont pas de sens ou changent le sens du groupe de mots choisis.

Le système PRECIS obtient ce résultat de façon apparemment bien simple: chaque descripteur est précédé d'un code. Ce code interprète la situation sémantique du mot par rapport aux autres mots de son groupe et indique les combinaisons dans lesquelles ce mot peut se trouver. Il serait évidemment trop long d'expliquer tous ces codes (dans la dernière édition de PRECIS[7], on en trouve vingt-six) et les manipulations qu'ils permettent. Mais on voit bien qu'ici encore, il ne s'agit pas d'indexation automatique, mais de la production par ordinateur d'une indexation manuelle: l'indexeur choisit les mots-clés et les codes. L'ordinateur se contente d'éditer les entrées autorisées par les codes.

Les avantages du système sont nombreux: on peut pratiquement exprimer n'importe quel sujet, ce qui satisfait les exigences d'une bibliographie nationale, et cependant, on limite suffisamment le nombre d'entrées, ce qui évite une bibliographie trop volumineuse. D'où la popularité de PRECIS et les nombreuses adaptations — à l'essai — à des langues autres que l'anglais, pour lequel il a été originellement conçu.

Tendances actuelles

Qu'il s'agisse donc de PRECIS, des index permutés ou de tout autre système d'édition d'index, le travail se fait toujours en deux étapes: 1) indexation par un analyste; 2) enregistrement sur ordinateur, tri et édition. Les tendances actuelles reflètent cette dichotomie. L'évolution du matériel permet d'éditer des produits toujours plus nombreux et exhaustifs, mais le goulot d'étranglement se situe au stade de l'analyse.

La technologie des ordinateurs évolue en effet très vite: pour l'entrée, le stockage et la sortie des données, on trouve sur le marché des appareils qui fonctionnent toujours plus vite et mieux. Pour l'entrée des données, on peut enregistrer directement sur disques à l'aide de mini-programmes qui permettent un certain nombre de corrections. Pour le stockage, on propose des mémoires douées de capacités d'emmagasinage et d'accès considérablement améliorées. Pour la

7. Derek Austin, *PRECIS, a Manual of Concept Analysis and Subject Indexing*, London, BNB, 1974.

sortie, on a dépassé le stade de l'imprimante avec uniquement des majuscules: l'ordinateur produit une bande magnétique qui sera utilisée pour la composition automatique. Ces progrès permettent donc une manipulation des données plus rapide et moins coûteuse. Cela est si vrai qu'il existe maintenant un véritable marché des banques de données. Il comprend aussi bien les bibliographies nationales (MARC américain, britannique, canadien, Bibliographie nationale de la France, etc.) qu'un grand nombre de périodiques secondaires (*Inspec, Chemical Abstracts, Bulletin signalétique,* etc.). Le nombre de banques de données disponibles approche la centaine et les véritables problèmes ne sont plus de production, mais de normalisation et d'utilisation.

La normalisation des banques devra faire l'objet d'efforts soutenus de la part de tous ceux qui sont concernés: on sait que certaines bibliographies nationales sont normalisées, mais la situation est encore catastrophique pour les banques de données bibliographiques produites par diverses organisations privées. Quant à la facilité d'utilisation, il est évident que la normalisation permettra des économies considérables de traitement en ordinateur, mais — et nous revenons à l'indexation automatique — elle ne suffira pas à procurer des outils réellement pratiques. Les banques de données bibliographiques ne seront véritablement utilisables qu'accompagnées d'une approche par sujets. En effet, on retrouve après l'automatisation les mêmes problèmes qu'on avait avant elle: la production documentaire se multiplie à un rythme tel que l'utilisation en devient très difficile. L'introduction des techniques informatiques permet déjà un meilleur contrôle de la production, mais le problème de l'usager n'est pas résolu pour autant. C'est déjà beaucoup de disposer de bibliographies nationales ou de revues secondaires sur bandes magnétiques. Mais ces bibliographies sont tellement abondantes qu'aucun usager ne peut y trouver son compte. Les compilations nationales et internationales de millions de titres d'ouvrages ou d'articles de périodiques, considérablement facilitées par l'arrivée de l'ordinateur, exigent un nouveau pas en avant: l'accès par sujets. Mais, comme nous l'avons vu, la classification et l'indexation elles-mêmes sont encore réalisées de façon manuelle, ce qui pose des problèmes énormes de coût, de rapidité et de qualité.

Il est en effet extrêmement coûteux d'affecter des analystes à la rédaction de résumés, à l'indexation et à la classification des documents que l'on veut entrer dans une banque de données; l'opération est assez longue et retarde la publication des bibliographies; enfin, la qualité de l'indexation reste toujours assez faible, puisqu'on ne sait toujours pas comment obtenir une indexation uniforme. On se trouve ainsi placé devant un paradoxe: on peut produire sur ordinateur des bibliographies très complètes, au niveau national ou par spécialités, mais elles demeurent très difficiles à utiliser à cause d'une indexation incomplète et non uniforme. Les recherches par sujets restent très inadéquates à cause de leur manque de précision. L'usager nage dans un flot de références inutiles ou bien il passe à côté de documents très importants pour lui. L'édition automatique d'index est certes un progrès important, mais qui demeure insuffisant pour résoudre les problèmes bibliographiques d'une société submergée par l'information.

On en vient donc au deuxième volet de notre présentation: la recherche sur l'indexation automatique.

Indexation automatique

Il faut essayer d'extraire les concepts essentiels de documents, et non plus de faire réaliser par l'ordinateur des opérations de tri et d'édition. Or, comme l'a bien noté Jean-Claude Gardin, c'est là une opération éminemment intellectuelle pour laquelle on en est encore au stade de la recherche fondamentale. Il n'existe aucun système opérationnel quoique les travaux et les essais soient extrêmement abondants. Pour s'orienter dans cette production, on peut diviser les recherches en deux groupes: méthodes statistiques, méthodes linguistiques.

Méthodes statistiques

L'utilisation des méthodes statistiques en indexation automatique à débuté vers la fin des années 1950. Le principe en est fort simple et il se fonde sur deux phénomènes statistiques: les fréquences et les cooccurrences. Dans les deux cas, on utilise les possibilités les plus immédiates de l'ordinateur, à savoir sa grande capacité de calcul. La première étape du traitement statistique consiste donc à entrer le texte complet du document à indexer dans l'ordinateur, et à calculer les fréquences de tous les mots. On élimite les mots trop fréquents comme non significatifs: les articles, les prépositions, les verbes être et avoir, etc. et les mots qui ne permettent pas de distinguer les documents à l'intérieur du corpus; ainsi, dans un ensemble de textes concernant la bibliographie, il est probable que les mots bibliographe et bibliographie reviendront très souvent: ils sont trop fréquents pour aider à caractériser les textes d'un corpus, c'est-à-dire pour les distinguer les uns des autres. On élimine aussi les mots trop peu fréquents, car il est probable que les sujets auxquels ils réfèrent sont très peu développés: si le mot «abonné», par exemple, se retrouve une fois dans l'ensemble des textes sur la bibliographie, il est peu probable que ce soit un concept utile dans le domaine. Il reste donc les mots à fréquence moyenne qui sont considérés comme représentant les concepts importants du corpus et peuvent servir à indexer les documents où ils apparaissent le plus souvent.

La deuxième étape — le calcul de cooccurrences — consiste à grouper les termes qui apparaissent ensemble dans les mêmes documents. Ainsi, dans un corpus concernant la bibliographie, si on trouve dix textes où les mots «entrée bibliographique», «notice bibliographique», «description bibliographique» y sont fréquemment employés, on peut considérer que ces trois mots réfèrent à un même concept. Cette deuxième étape permet surtout de regrouper les synonymes. On peut d'ailleurs utiliser le calcul des cooccurrences pour grouper les documents au lieu des termes (i.e. les documents où l'on retrouve les mêmes mots sont considérés comme faisant partie d'une même classe) et l'on fait ce qu'on appelle alors de la classification automatique.

Il y a eu entre 1960 et 1965 une quantité énorme de publications

— surtout aux États-Unis — concernant les méthodes statistiques. Aux principes qui ont été décrits, on a essayé d'ajouter des raffinements de plus en plus complexes, mais ces méthodes n'ont jamais abouti (sauf peut-être une exception) à des réalisations opérationnelles. En 1973, Sparck-Jones et Kay écrivaient: «Il n'y a pas eu de progrès intellectuel substantiel dans ce domaine dans les récentes années.»[8] Les raisons sont d'ailleurs évidentes: on ne voit pas pourquoi la fréquence d'un mot serait liée à l'importance d'un sujet, lorsqu'on connaît la diversité d'expression que le langage naturel peut offrir pour exprimer une même idée. Il est probable que la statistique linguistique constituera un apport substantiel à notre connaissance du langage, et donc à l'indexation automatique, à condition qu'elle ne soit pas considérée comme le raccourci idéal qui permet de négliger l'étude et la prise en considération de tous les autres aspects de la linguistique.

Méthodes linguistiques

Remarques historiques

Avant d'aborder les méthodes linguistiques, il importe de formuler deux remarques concernant l'évolution de la recherche dans ce domaine. Être capable de déterminer le sujet ou les concepts principaux d'un document par l'examen de la suite de caractères qui forment son texte suppose, d'une part, que l'on connaisse suffisamment les mécanismes du langage pour comprendre comment s'effectue le passage de la forme au sens et, d'autre part, que l'on soit capable de simuler ce processus avec un ordinateur. Or, ces deux conditions sont loin d'être remplies.

En ce qui concerne la première condition, comprendre les mécanismes qui permettront de passer de la forme au sens, il faut signaler que la linguistique ou l'étude scientifique du langage est une science très récente. Tout d'abord, les concepts qui servent de base à l'édifice actuel datent presque tous de ce siècle (structure, fonction, paradigme) et certains de sa deuxième moitié (transformation, chaîne, grammaire des cas, sémantique générative, etc.). Ensuite, les phénomènes linguistiques à peu près bien compris actuellement concernent le domaine des sons et de la grammaire, et non celui du sens: la multiplicité des théories sémantiques qui ont présentement cours est autant une preuve d'ignorance que de dynamisme. Enfin, il n'existe aucune théorie intégrée du langage qui couvre précisément l'ensemble des faits reconnus et attestés comme pertinents, y compris, et surtout, les mécanismes qui permettent de passer de la forme au sens et vice-versa. Lorsqu'on veut étudier l'état de la recherche en indexation automatique, il faut donc garder à l'esprit la diversité des théories linguistiques et les lacunes importantes dont elles souffrent.

Si l'on en vient à la deuxième condition, force nous est de constater qu'il est très difficile de simuler le processus qui permet de passer de la forme au sens avec un ordinateur, puisqu'on en ignore les principes fondamentaux. La recherche dans ce domaine est donc

8. Karen Sparck-Jones and Martin Kay, *Linguistics...*, 129.

extrêmement variée et parcellaire. Le passé nous donne d'ailleurs en ce domaine une leçon qu'il est intéressant de rappeler. On sait que les tout premiers ordinateurs virent le jour en 1946 et qu'ils servaient surtout à effectuer de longs calculs de balistique. Or, moins de cinq ans plus tard, on songeait à les appliquer à la traduction automatique[9]. La tâche semblait en effet idéale pour une machine automatique à grande capacité de mémoire: il aurait suffi d'entrer un dictionnaire bilingue et les règles de grammaire des deux langues. La suite des événements a démontré davantage la force de l'enthousiasme pour les ordinateurs que la perspicacité des pourvoyeurs de fonds. Il a fallu presque quinze ans avant de réaliser que le processus de traduction n'était pas réductible à l'utilisation combinée d'un dictionnaire et d'une grammaire. Bien plus, on s'est aperçu que la connaissance des mécanismes du langage était très sommaire et qu'il valait mieux investir dans la recherche fondamentale plutôt que de jeter des millions de dollars dans des projets dont on ne voyait pas l'aboutissement[10]. C'était en 1965.

Cet aperçu historique de la traduction automatique et de la linguistique aideront à comprendre pourquoi la véritable indexation automatique n'est encore qu'à l'état de projet.

Personne n'a encore rendu un ordinateur capable de lire un texte et d'en extraire les principaux mots-clés de façon à obtenir une indexation valable, étant bien entendu que le texte fait partie d'un corpus substantiel sur un sujet donné. Tous les travaux dont on peut parler sont donc des approches sous des angles parfois fort différents. On peut regrouper les recherches dans ce domaine de la façon suivante: indexation syntaxique, indexation sémantique, indexation mixte, recherche en intelligence artificielle.

Indexation syntaxique

L'indexation syntaxique a sans doute été l'approche qui a été explorée le plus tôt, à peu près en même temps que les méthodes statistiques. Ce fait est dû à des considérations historiques: après avoir formalisé la phonologie, les linguistes pouvaient s'attaquer à la formalisation, au moins sommaire, de certains aspects morphologiques et syntaxiques du langage. Il devenait donc tentant d'appliquer les théories linguistiques naissantes à l'analyse automatique des phrases. On distingue quatre écoles, selon leur ordre d'apparition: l'analyse par constituant immédiat, d'après les premiers travaux structuralistes; l'analyse par chaîne, en fonction des théories de Harris; l'analyse transformationnelle, d'après les travaux de Chomsky et, enfin, les autres, ce qui comprend des approches très diverses, basées parfois sur des théories linguistiques, parfois sur des approches originales plus motivées par les contraintes de l'ordinateur que par les théories linguistiques.

9. Voir par exemple dans Y. Bar-Hillel, *Language and Information*, Reading, Mass., Addison-Wesley, 1964.

10. Karen Sparck-Jones and Martin Kay, *Linguistics...*, 41.

Il n'est pas nécessaire d'aborder le détail des procédures utilisées, d'ailleurs très complexe. Ce qu'il faut retenir, c'est que toutes ces méthodes cherchent à analyser un texte phrase par phrase. Pour chaque phrase, on veut déterminer la structure syntaxique, c'est-à-dire établir les relations syntaxiques entre les unités lexicales constituant la phrase (sujet, verbe, complément d'objet, etc.). Les résultats ont en général été très décevants pour deux raisons: d'abord, dès qu'on veut analyser des phrases un peu complexes, l'ordinateur produit plusieurs analyses syntaxiques sans être capable d'identifier la bonne; ensuite, il faut au moins une minute, parfois quinze, pour analyser *une* phrase. On voit donc pourquoi on est loin de pouvoir analyser des textes entiers, et encore moins utiliser cette analyse purement syntaxique pour l'indexation. Les seules applications en repérage de l'information concernaient les systèmes de question-réponse, l'ordinateur étant capable d'analyser les courtes phrases interrogatives qu'un usager peut lui poser. Les expériences de Salton pour indexer avec des arbres syntaxiques privilégiés n'ont pas abouti aux résultats escomptés.

L'impasse de l'analyse syntaxique pure en indexation automatique a mené dans deux directions: recherche sémantique d'une part et méthode mixte d'autre part.

Puisque le principal problème de l'analyse syntaxique réside dans l'impossibilité de choisir entre plusieurs analyses possibles, on essaie d'introduire des données sémantiques. Ces dernières permettent de faire ce choix grâce à certaines restrictions qu'imposent les relations sémantiques. Cette nouvelle façon de poser le problème était d'ailleurs apparue presque en même temps en linguistique théorique. Mais avant de voir comment cette fusion de l'analyse syntaxique et de l'analyse sémantique s'est opérée, examinons les tentatives purement sémantiques d'indexation automatique.

Indexation sémantique

Dans l'indexation sémantique, il s'agit de reconnaître dans un texte les mots qui sont des descripteurs autorisés. La démarche se divise donc en trois étapes: 1) reconnaissance des mots significatifs d'un texte; 2) établissement des équivalences avec les descripteurs autorisés; 3) choix des descripteurs importants.

Au niveau de la première étape, le problème est le suivant: quelles sont les unités linguistiques utilisables pour l'indexation? Ce peut être, en effet, soit des mots, soit des racines, soit des groupes de mots. Pour les mots, en autant qu'il s'agit de substantifs, il y a peu de difficultés, sauf dans les cas d'homonymie. Mais on peut utiliser certains mots même s'ils ne sont pas des substantifs (ex.: oxyder, oxydable). On en est donc venu à chercher des racines (oxyd-) qui permettraient de retrouver les trois mots «oxyd-ation», «oxyd-er», «oxyd-able» et de les indexer par le même descripteur. Enfin, il faut souvent être capable de trouver des groupes de mots tels que «situation de stress» ou «ablation des lobes». Selon une approche purement sémantique, c'est une opération assez difficile puisqu'on ignore les césures naturelles de la phrase que seule une analyse syntaxique

pourrait révéler. Une indexation sémantique devrait donc être capable de repérer ces trois genres d'unités: mots, racines, groupes de mots.

Une fois ces unités reconnues, il faut choisir l'équivalent dans la liste des descripteurs autorisés. Ceci n'est possible que si on a, au préalable, construit manuellement un dictionnaire ressemblant aux renvois préférentiels d'un thésaurus, par exemple:

Congrès EM CONFERENCE
Séminaire EM CONFÉRENCE

Pour les deux premières étapes d'une indexation sémantique, reconnaissance des mots significatifs et établissement des équivalents, on sait donc quelles sont les opérations à effectuer. Mais on n'a peut-être pas saisi l'ampleur de la tâche, dès qu'on veut travailler avec un corpus ouvert: il faut prévoir toutes les formes du langage naturel qui expriment des concepts importants dans un domaine et construire un dictionnaire d'équivalence entre les formes du langage naturel et les descripteurs acceptés. Évidemment, très peu de projets de ce type ont été entrepris.

Quant à la troisième étape, le choix des descripteurs importants, la recherche est encore embryonnaire. Les deux premières étapes produisent une liste de descripteurs présents dans le document. Rien ne dit quels sont les concepts importants, ceux sur lesquels l'auteur insiste ou pour lesquels il donne de l'information suffisamment développée ou nouvelle pour l'usager éventuel.

Les indexeurs établissent instinctivement ce choix par la compréhension du texte et certains critères formels[11]. Mais quiconque a essayé d'enseigner l'indexation sait combien il est difficile d'expliciter et de transmettre les critères de ce choix. Ceci explique le manque d'uniformité de l'indexation humaine et l'insuffisance du critère statistique. L'automatisation du choix des descripteurs est donc encore loin: elle suppose la découverte de critères formels pour l'analyse du discours. Les rares tentatives existantes (ex. Grimes[12]) montrent tout le chemin encore à parcourir.

Il est vrai qu'au lieu de partir de textes complets, on pourrait partir de résumés. Ce pis-aller sera peut-être la solution adoptée dans les années à venir.

Cependant, même au niveau des résumés, le choix des concepts n'est pas facile comme nous allons le voir à propos de l'indexation mixte.

Indexation mixte

Nous avons vu jusqu'à présent l'indexation syntaxique et l'indexation sémantique; les méthodes mixtes cherchent à combiner ces deux

11. Lawrence H. Oliver, *An Investigation of the Basic Process Involved in the Manual Indexing of Scientific Documents,* Bethesda, Md., General Electric Corp., 1966.

12. J.E. Grimes, *The Thread of Discourse,* Paris, Mouton, 1975.

sortes d'information pour aboutir à une représentation formalisée des documents. En fait, cette catégorie ne comprend qu'un seul projet de recherche mais qui vaut la peine d'être mentionné à cause de son ampleur et de son importance théorique[13].

Le but poursuivi était double: d'abord, rendre un ordinateur capable de trouver les descripteurs valables, ce qui inclut les trois étapes de l'indexation sémantique énumérées plus haut; ensuite, exprimer dans un langage documentaire les relations entre les concepts, ce qui est le résultat de l'analyse syntaxique.

Là non plus, nous n'entrerons pas dans le détail de l'expérience. Sommairement, l'ordinateur procédait en deux étapes: indexation sémantique puis indexation syntaxique. À la première étape, on développe un certain nombre d'algorithmes afin de relever les homonymies et les polysémies de toutes sortes, et de trouver les groupes de mots valables. À la seconde étape, on emmagasine un certain nombre de microstructures syntaxiques dont on connaît l'interprétation dans le langage naturel. Il suffit que l'ordinateur les reconnaisse pour que les concepts présents dans ces structures soient liés en conséquence. Ainsi, deux phrases du langage naturel telles que «l'atropine agit sur le foie» et «le foie subit l'action de l'atropine» seront interprétées de la même façon, bien que les structures syntaxiques du langage naturel soient assez différentes.

Évidemment, dans cette expérience, on n'a pas fait de recherche à partir de textes entiers. Les difficultés linguistiques étaient déjà assez considérables avec des textes plus courts. Il convient de souligner que le corpus était relativement grand (presque huit cents résumés tirés du *Bulletin signalétique* du CNRS) et que les résultats furent assez encourageants, compte tenu du parti pris d'éviter une analyse syntaxique complète qui, on l'a vu plus haut, est encore loin d'être réalisable. La double leçon de l'expérience mérite d'être retenue. D'une part, il est possible d'obtenir une indexation satisfaisante, avec un ordinateur, sans faire une analyse grammaticale complète. D'autre part, si l'on veut rendre un ordinateur capable d'une telle tâche, il faut un investissement préalable considérable:

> «L'analyse sémantique d'un texte scientifique, fût-il déjà résumé, est une opération éminemment intelligente, qui exige une double compétence, sur le plan de la langue tout d'abord, mais aussi sur le plan de la pensée scientifique elle-même, puisqu'enfin l'on n'attend plus aujourd'hui d'un documentaliste omniscient qu'il soit capable de dégager indifféremment le sens d'un article de physique théorique ou de sociologie. La machine doit être instruite de la même manière dans les deux ordres de compétence.»[14]

Néanmoins, en indexant des résumés au lieu de textes complets, on évitait en grande partie la question non encore résolue du choix des descripteurs valables, puisque presque tous les descripteurs re-

13. N. Bély et al., *Procédures...*

14. *Ibid.*, xiv.

connus pouvaient servir à l'indexation. Le véritable problème sémantique reste toujours entier et c'est ce qui justifie les développements actuels de la recherche.

Recherche en intelligence artificielle

On a vu, à propos de l'analyse syntaxique, d'une part qu'il était difficile d'analyser des phrases complexes, d'autre part qu'il fallait, probablement, introduire de l'information sémantique. Grâce à l'analyse sémantique, on a pu constater la nécessité d'explorer dans un domaine restreint les problèmes de structure sémantique. L'expérience de méthode mixte n'a abouti qu'à des résultats incomplets: il faudra un jour que l'ordinateur puisse «comprendre» le langage naturel, ou au moins faire comme si... Il était donc naturel que la recherche en linguistique automatique commence à explorer le domaine de l'intelligence artificielle. En délimitant un domaine conceptuel très restreint, on essaie d'aboutir à une formalisation sémantique complète. Les phrases concernant un tel domaine, quoique en théorie leur nombre puisse être infini, resteront syntaxiquement assez simples. En liant l'analyse linguistique à un système qui doit répondre ou agir en conséquence, on s'oblige à tenir compte du sens des phrases et donc à combiner l'information syntaxique et sémantique.

Les travaux en intelligence artificielle représentent la tendance la plus vivante des recherches en linguistique automatique. Là non plus, il n'est pas nécessaire d'expliquer la démarche ou même de citer les noms principaux. Les résultats, par contre, indiquent ce qu'on peut attendre du point de vue de l'indexation automatique..

D'après l'état de la question de Wilks[15], il semble possible de tirer les conclusions suivantes: quoique certaines réalisations (celle de Winograd en particulier) semblent impressionnantes pour le profane, on peut dire qu'il y a eu peu de progrès substantiel, au point que Wilks définit la situation actuelle par un choix stratégique:

> «... on sait que tout système ou structure sera finalement insuffisant face au langage, si bien qu'une seule question demeure: quand devra-t-on l'abandonner? La seule question intéressante (...) est de savoir jusqu'où il vaut la peine de pousser une approche structurale quelconque avant de recommencer à zéro.»[16]

En fait, il reste deux problèmes essentiels. Tout d'abord, on ne dispose pas d'un modèle satisfaisant pour emmagasiner l'information sémantique. Certains auteurs, par exemple Montgomery[17], ont souligné les similitudes des différents modèles; mais ces analogies sont insuffisantes. Un bon modèle de mémoire sémantique devrait permettre, entre autres, de choisir le niveau de représentations (entre les mots même du langage naturel et des marqueurs sémantiques primitifs), de

15. Yorick Wilks, «Natural language understanding systems within the A.I. paradigm: a survey», *American Journal of Computational Linguistics*, no. 1 (1976). Microfiche no. 40.

16. *Ibid.*, 40.

17. Christine A. Montgomery, «Linguistics...».

choisir l'information pertinente sans aller trop loin dans les implications et les précisions (ce que Wilks appelle le niveau phénoménologique).

Le deuxième problème est sans doute la contrepartie théorique du premier. On n'a pas de modèle de mémoire, comme on n'a pas de théorie sémantique. Quelle théorie permettrait de traiter à la fois l'information très générale (valable pour presque toutes les situations linguistiques) et l'information spécifique (propre à une situation très particulière)? Quelle théorie permettrait d'aborder autant ce qui tient du sens commun que ce qui appartient à la connaissance scientifique spécialisée?

. Il ne s'agit pas de dénigrer un domaine de recherche qui vient tout juste de naître. Pourquoi ne pas lui laisser le temps de faire ses preuves? Il n'en reste pas moins que l'applicabilité pour l'indexation automatique se réduit à presque rien.

Conclusion

Nous avons tenté de montrer que l'expression indexation automatique recouvrait deux réalités bien différentes. D'une part, on trouve l'édition automatique d'index, technique bien au point et pouvant se flatter d'un grand nombre de réalisations opérationnelles, d'autre part, la véritable indexation automatique, l'extraction du sens des documents, domaine de recherche encore loin des réalisations.

Dans le premier cas, le progrès de la technologie justifie l'attente de produits documentaires de meilleure qualité, par exemple au niveau du graphisme ou de la périodicité des refontes. L'automatisation des bibliographies nationales va continuer et gagner d'autres pays. Les banques de données spécialisées vont se développer et rejoindre de plus en plus de chercheurs. L'utilisation de l'ordinateur pour le contrôle bibliographique, que ce soit les données signalétiques ou les données analytiques, continuera d'apporter des changements profonds dans le travail bibliographique.

Dans le second cas, il faut au contraire être extrêmement prudent. Malgré certaines précautions, dues sans doute au climat de concurrence qui règne dans le monde de la recherche, les réalisations concrètes sont extrêmement rares. La recherche progresse lentement, parce que, finalement, on a encore peu de données sur le langage et aussi parce qu'on se refuse à vouloir faire l'investissement intellectuel nécessaire. La présente tendance vers la centralisation de l'entrée et du traitement des données, que l'on constate dans les grands systèmes publics et privés de documentation, devrait bientôt permettre cet investissement. Mais il semble que, pendant de nombreuses années encore, il faudra recourir aux analystes et aux indexeurs.