Documentation et bibliothèques

La condensation et l'indexation : l'apport des approches de type textuel
Condensation and Indexation: The Contribution of Full Text Approaches
Condensacion e indexacion: aporte de los enfoques textuales

Luc Jodoin

Volume 38, numéro 2, avril–juin 1992

Analyse et gestion de l'information textuelle

URI : https://id.erudit.org/iderudit/1028611ar
DOI : https://doi.org/10.7202/1028611ar

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)

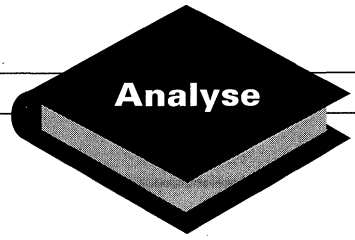
Découvrir la revue

Citer cet article

Jodoin, L. (1992). La condensation et l'indexation : l'apport des approches de type textuel. *Documentation et bibliothèques*, *38*(2), 71–74. https://doi.org/10.7202/1028611ar

Résumé de l'article

L'auteur analyse les processus de condensation et d'indexation à la lumière des théories sémiotiques. Il dresse un tableau sommaire de l'évolution de ces analyses de type textuel et montre en quoi elles permettent de rendre compte des processus d'extraction de concepts lors de l'analyse documentaire. Une approche est proposée dans le but de dépasser l'opposition entre les analyses centrées sur le texte et celles centrées sur le seul lecteur. Il est suggéré d'orienter les pratiques de traitement de l'information en intégrant la notion de « Lecteur modèle » de Umberto Eco et en tenant compte des contextes de communication propres aux différents domaines du savoir.



La condensation et l'indexation : l'apport des approches de type textuel

Luc Jodoin
Division du traitement documentaire
Bibliothèque municipale de Montréal

L'auteur analyse les processus de condensation et d'indexation à la lumière des théories sémiotiques. Il dresse un tableau sommaire de l'évolution de ces analyses de type textuel et montre en quoi elles permettent de rendre compte des processus d'extraction de concepts lors de l'analyse documentaire. Une approche est proposée dans le but de dépasser l'opposition entre les analyses centrées sur le texte et celles centrées sur le seul lecteur. Il est suggéré d'orienter les pratiques de traitement de l'information en intégrant la notion de « Lecteur modèle » de Umberto Eco et en tenant compte des contextes de communication propres aux différents domaines du savoir.

Condensation and Indexation: The Contribution of Full Text Approaches

Using the theories of semiotics as a backdrop, the author analyses the processes of condensation and indexation. He provides a thumbnail sketch of the progress of the textual theories and demonstrates how they extract concepts during subject analysis. A new approach is put forward as a compromise between text-based analyses and reader-based analyses. It is suggested that practices be oriented toward information analysis integrated with Umberto Eco's reader model. The communication contexts of each sphere of human knowledge must also be taken into consideration.

Condensacion e indexacion: aporte de los enfoques textuales

El autor analiza los procesos de condensación e indexación desde el punto de vista de las teorías semióticas. Hace un cuadro sinóptico de la evolución de esos análisis textuales y muestra de qué manera permiten dar cuenta del proceso de extracción de conceptos durante el análisis de un documento. Se propone un enfoque para eliminar la oposición que existe entre los análisis de textos y los centrados en el lector único. Se sugiere la orientación de prácticas de tratamiento de información, integrando la noción de « lector modelo » de Umberto Eco y tomando en consideración los contextos de comunicación propios de los diferentes campos del saber.

La représentation de l'information

Les étapes menant à la réduction et à la représentation du contenu des documents font l'unanimité en science de l'information. L'analyste, dans un premier temps, extrait les idées et les concepts principaux d'un document, puis il procède à leur traduction, en respectant le vocabulaire et la syntaxe d'un langage documentaire.

La condensation, l'indexation et la classification sont similaires quant à l'extraction du contenu des documents[1], et ce n'est qu'au moment de la traduction de ce contenu qu'elles se distinguent. Toutes trois visent à dégager la principale macro-structure d'un document[2]. Leur spécificité, quant au produit final, réside dans le degré de profondeur de l'analyse. Ainsi, par exemple, une analyse classificatoire sera moins précise qu'une analyse par descripteurs.

Les publications qui s'intéressent à ce double aspect de la réduction du contenu des documents privilégient généralement le *second* moment du processus[3]. L'opération dite d'extraction est souvent prise à la légère. L'activité principale des analystes - la lecture - semble aller de soi, comme si la lecture pouvait se réduire à un banal transfert *passif* du contenu d'un document dans la mémoire d'un lecteur.

Il est courant de lire que l'opération d'extraction est simple, qu'il suffit de lire introduction et conclusion, de parcourir les intertitres et de survoler les textes pour déceler les mots qui semblent les plus riches de signification. Il ne s'agit pas de remettre en question la compétence des indexeurs et des rédacteurs de résumés. Au contraire, c'est justement parce que, dans le processus de la lecture, le récepteur investit ses propres connaissances d'un domaine (lui-même fortement structuré), qu'il peut faire l'économie de la lecture de la totalité d'un texte.

1. Brigitte Endres-Niggermeyer, « A Procedural Model of Abstracting, and Some Ideas for its Implementation », *TKE'90, Terminology and Knowledge Engineering: Applications*, INDEKS Velag Frankfurt, 1990, 220-243.
2. Sur la question voir Teun. A. Van Dijk, « Perspective Paper: Complex Semantic information Processing », in D.E. Walker et al., *Natural Language in Information Science: Perspectives and Directions for Research*, Stockholm, Skriptor, 1977, p. 127-163; Clare Begthol, « Bibliographic Classification Theory and Text Linguistics: Aboutness Analysis, Intertextuality and the Cognitive Act of Classifying Documents », *Journal of Documentation*, vol. 42, no. 2 (June 1986), 84-113.
3. Voir les commentaires de: W. J. Hutchins, « The Concept of Aboutness in Subject Indexing », *Aslib Proceedings*, vol. 30, no. 5 (May 1978), 172-181; B. Frohmann, « Rules of Indexing: A Critique of Mentalism in Information Retrieval Theory », *Journal of Documentation*, vol. 46, no. 2 (June 1990), 81-101; Brigitte Endres-Niggermeyer, « A Procedural Model of Abstracting... »

Pour rendre compte du processus d'extraction des concepts, un détour s'impose par des analyses de type sémiotique qui prennent en compte le rôle actif du lecteur dans l'actualisation de la signification.

Ces courants théoriques peuvent s'appliquer au domaine de la documentation, que ce soit sous l'angle de la représentation, du traitement ou du repérage de l'information.

Le système du discours

Pour Algirdas J. Greimas, le discours ne peut être réduit à un simple agrégat de termes-objets. De tels termes seraient en eux-mêmes dépourvus de signification, car « la signification présuppose l'existence de la relation: c'est l'apparition de la relation entre les termes qui est la condition nécessaire de la signification »[4].

Cette conception du discours, qui doit beaucoup à Saussure, s'impose de plus en plus en science de l'information. Les chercheurs l'ont pour ainsi dire faite leur, à la suite d'une démarche quelque peu empirique. Rappelons, par exemple, l'échec des bases de données textuelles à représenter le contenu des documents avec les seuls mots clés (des termes-objets) ou les piètres résultats de recherche que l'on obtient avec des bases de données plein texte qui utilisent les mêmes paramètres que les bases de données formatées. Ces systèmes ne donnent jamais accès qu'à des concepts isolés, sans tenir compte du réseau sémantique dans lequel ils s'inscrivent. Ainsi, par exemple, un usager qui fait une recherche sur l'influence de **A** sur **B** dans un contexte **C** récupérera aussi des documents portant sur l'effet de **C** sur **A** dans un contexte **B**. Quoique l'on puisse utiliser différentes techniques (opérateurs d'adjacence, ordre des mots, co-occurrence, etc.) lors de l'interrogation des bases de données plein texte, il est, à toutes fins utiles, impossible d'accéder à un concept en fonction d'une relation sémantique spécifique.

Les recherches booléennes sur des fichiers inversés ne sont pas adaptées au plein texte du fait de la représentation statique du contenu des documents. Les descripteurs n'y sont pas pondérés en fonction de la macrostructure textuelle: un terme tiré d'un titre de section a la même valeur sémantique que tout autre mot extrait d'un paragraphe quelconque.

C'est d'ailleurs en s'attachant à représenter le contenu des documents à partir de la structure des textes (et non pas en fonction des termes-objets) que des améliorations sensibles se sont fait sentir au niveau du repérage de l'information dans certaines bases de données plein texte[5].

Le rôle du résumé consiste précisément à pallier ces insuffisances[6]. Il doit permettre une restitution des acceptions différentes des signes selon les rapports paradigmatiques et syntagmatiques qu'ils entretiennent avec les autres signes présents dans le texte. De plus, le sujet d'un document n'affleurant pas toujours à la surface du texte, il permet de rendre compte de nombreuses informations qui doivent être inférées à partir de la base explicite[7]. Enfin, on compte sur lui pour avoir accès à une représentation de la progression thématique[8]. Ces dimensions échappent complètement à une représentation des documents s'appuyant sur les seuls termes-objets.

La percée récente de la pragmatique a amené certains chercheurs à vouloir faire *tabula rasa* de l'intuition saussurienne. Nous pensons qu'il est possible d'intégrer les acquis de la pragmatique à l'intérieur de ce cadre qui demande à être réaménagé, certes, mais qui n'a sûrement pas épuisé toutes ses potentialités.

La pragmatique

Le principal reproche adressé à la sémiologie saussurienne concerne le peu de cas qu'elle fait du contexte dans lequel s'inscrit la langue. Saussure, dans sa tentative de fonder la sémiologie, priorisait le système plutôt que l'emploi de la langue. Il escamotait ainsi toute la dimension de la communicabilité, de la réalisation concrète de la langue.

C'est à Wittgenstein et à Peirce que l'on doit l'émergence du paradigme de la communicabilité.

Wittgenstein dégage le caractère central de la communicabilité. En critiquant la théorie subjectiviste et mentaliste de la signification, il pose que la pensée n'est pas d'abord quelque chose d'intérieur au sujet qu'il faudrait ensuite traduire en mots pour l'extérioriser. Chez lui, le langage n'est pas d'abord privé puis traduit dans un langage public, mais d'emblée public et déjà constitué d'un ensemble de règles. Wittgenstein ne s'intéresse pas tant à l'usage du mot dans la phrase qu'aux mots eux-mêmes, en tant que situations d'action. Ainsi, chez lui, « la visée première du langage n'est pas une visée de compréhension ou représentation, mais l'exercice d'une influence des uns sur les autres »[9].

Dans le cadre de cette substitution d'un paradigme de la communicabilité au paradigme de l'expressivité[10], l'apport de Peirce nous semble tout aussi fondamental.

Peirce rejoint néanmoins Saussure sur un point: le signe n'existe pas en soi. L'une des propriétés du signe consiste à toujours renvoyer à un autre signe.

> *Ainsi la pensée est elle-même un signe, qui renvoie à une autre pensée, laquelle est son signe interprétant. Ce dernier renvoie encore à une autre pensée qui l'interprète, en un processus continu et indéfini*[11].

4. Algirdas Julien Greimas, *Sémantique structurale*, Paris, Presses universitaires de France, 1986, p. 19.

5. Voir entre autres: Ugo Hahn, « Topic Parsing: Accounting for Text Macrostructures in Full-Text Analysis », *Information Processing & Management*, vol. 26, no. 1 (1990), 135-170; Elizabeth D. Liddy, « Structure of Information in Full-Text Abstracts », in *RIAO 88*, [Paris], Centre des Hautes Études Internationales d'Informatique Documentaire, 1988, p. 183-195; Takashi Maeda, « An Approach Toward Functional Text Structure Analysis of Scientific and Technical Documents », *Information Processing & Management*, vol. 17, no. 6 (1981), 329-339; Jiri Janos, « Theory of Functional Sentence Perspective and its Application for the Purposes of Automatic Extracting », *Information Processing and Management*, vol. 15 (1979), 19-25.

6. [CREDO], Centre de recherches sur la documentation et l'information, *Bases de données Cultures et religions antiques: Introduction méthodologique*, Lille, Université de Lille III, 1987, 87 p.

7. Teun. A. Van Dijk, « Perspective Paper... »

8. Jiri Janos, « Theory of Functional... »; Takashi Maeda, « An Approach Toward Functional... »

9. Françoise Armengaud, *La pragmatique*, Paris, PUF, 1985, p. 27 (Que sais-je ? no 2230)

10. *Ibid.*, p. 28.

11. *Ibid.*, p. 19.

Peirce se démarque cependant de Saussure dans la mesure où chez lui le signe est toujours fonction de l'usage. C'est d'ailleurs une logique différente de celle de Saussure qui préside à la construction de sa sémiotique. Alors que chez Saussure il y a un rapport d'extériorité du signe à l'interprète, chez Peirce tout se joue à l'intérieur même du signe par le jeu dynamique entre trois constituantes de ce même signe (le fondement, l'objet et l'interprétant)[12].

C'est cette nature triadique du signe qui lui permet de penser la signification comme un processus sans fin. L'interprétant devient toujours le fondement d'un autre signe, de telle sorte que la « semiosis » devient infinie.

Dans cette perspective, le sens et la signification étant toujours fonction des connaissances collatérales du lecteur et de ses capacités de traitement de l'information, il devient pratiquement impossible, dans une optique de repérage de l'information, de procéder à une représentation de l'information qui pourrait satisfaire l'ensemble des requêtes de tous les usagers potentiels[13].

Si la sémiologie de Peirce réussit à garantir l'impossibilité de fonder une « vérité » dogmatique des textes, elle demeure impuissante face au danger inverse, celui d'une régression à l'infini, d'une impossibilité de fonder ou de penser la communication[14].

La science de l'information, dans un légitime souci de satisfaction des besoins documentaires des usagers, a parfois tendance à verser dans le monologisme et le subjectivisme en faisant porter tout le poids du sémantisme sur les seuls usagers. Les travaux de Grize[15] nous semblent infirmer l'autonomie du sujet (de l'usager) eu égard aux significations communicatives.

Nous pensons qu'une partie de la réponse à cette problématique des « besoins » réside dans les textes eux-mêmes. Ainsi que l'a montré Umberto Eco, les textes sont des machines qui supposent la coopération du lecteur tout en étant des constructions de ce même lecteur : « un texte est un produit dont le sort interprétatif doit faire partie de son propre mécanisme génératif »[16].

Une distinction s'impose entre *interprétation* et *utilisation* des textes[17]. Umberto Eco opère cette distinction en ce qui concerne les textes narratifs. Nous croyons pouvoir la reprendre à notre compte pour l'appliquer aux textes informatifs. Un système d'information doit tenir compte de l'*interprétation* et non de l'*utilisation* des textes. Aucun langage documentaire ne pourra jamais se substituer aux textes eux-mêmes. Aucun système ne peut, ni ne doit, anticiper la totalité des *utilisations* de sa collection. On risquerait alors, sous prétexte de la pluralité du sens, de transmuer la « signifiance » en « signifiose ». Si l'on peut dire (lire) n'importe quoi de n'importe quel texte ou de n'importe quelle image alors tous les textes, toutes les images deviennent synonymes. Paradoxalement, sous prétexte de représenter le pluriel des textes, on aboutirait à une approche monologique du sens[18].

La notion de lecteur modèle

On a cru longtemps que le sens était une catégorie du texte, immédiatement assimilable pour le lecteur. Le sens pouvait facilement s'expliquer : il ne s'agissait que d'une simple transposition des idées de l'auteur dans la mémoire du lecteur. Des chercheurs explorent aujourd'hui une hypothèse inverse, la lecture consiste en un processus actif où le lecteur crée lui-même le sens du texte. Pour ce faire, il mobilise à la fois le texte, ses propres connaissances et ses intentions de lecture[19].

On ne doit cependant pas en conclure à une polysémie polymorphe de tous les textes. Les auteurs utilisent certaines conventions pour écrire. Ils supposent que le lecteur connaît certains scripts ou scénarios préétablis. Tout texte, selon Eco, « présuppose la compétence de son Lecteur Modèle et en même temps il l'institue »[20]. Si cette supposition ne se vérifie pas, le message ne sera pas compris.

> *La thèse sous-jacente à l'ensemble de ces tendances [les modèles de lecture] consiste à poser que le fonctionnement d'un texte (...) ne peut s'expliquer que si l'on prend en considération, en plus ou à la place du moment de sa génération, le rôle joué par son destinataire du point de vue de sa compréhension, de son actualisation, de son interprétation, ainsi que la manière dont le texte lui-même prévoit de tels modes de participation*[21].

En ce sens, un texte scientifique prévoit dans sa propre actualisation un lecteur modèle précis. Sa structure, les catégories qu'il mobilise et, par exemple, le rôle éventuel que le texte jouera dans une communauté de chercheurs, tout cela limite en quelque sorte le jeu de la semiosis. Un texte informatif se prête moins à des lectures aberrantes (pour peu, évidemment, que le lecteur ait une connaissance préalable du domaine) qu'une oeuvre poétique.

Le processus de lecture est déjà inscrit dans tout texte et celui-ci comprend toujours le principe de sa propre transformation : un lecteur modèle. C'est ce principe qu'il faut arriver à dégager pour des catégories spécifiques du savoir.

Pour décrire la lecture sous l'angle de la communication, il ne suffit pas d'expliquer la circulation d'un message, via un code partagé, peu ou prou, entre un émetteur et un récepteur. La lecture ne consiste pas en un simple décodage mais en une traduction, un transcodage[22], une relation de

12. Pour une introduction à la pensée de Peirce, nous suggérons fortement la lecture du livre de Jean Fisette : *Introduction à la sémiotique de Peirce*, Montréal, XYZ, 1990, 86 p. (Collection « Études et documents »)
13. Suzanne Bertrand-Gastaldy, *Le recours à la sémiotique de Peirce en science de l'information* ; État de la question présenté dans le cadre d'un séminaire dirigé par Gérard Deledalle, Université du Québec à Montréal, automne 1986, [23 f.].
14. Umberto Eco, « Notes sur la sémiotique de la réception », *Actes Sémiotiques*, vol. 9, no 81 (1987), 27 p.
15. Jean-Blaise Grize, *Logique et langage*, France, OPHRYS, 1990, 153 p.
16. Umberto Eco, *Lector in fabula*, Paris, Grasset, 1985, p. 70.
17. *Ibid*, p. 76.
18. Catherine Kerbrat-Orecchioni, *La connotation*, Lyon, Presses universitaires de Lyon, 1977, p. 1-21.
19. Jocelyne Giasson, *La compréhension en lecture*, Boucherville, Gaëtan Morin, 1990, p. 19.
20. Umberto Eco, *Lector in fabula...*, p. 72.
21. Umberto Eco, « Notes sur la sémiotique... », p. 6.
22. Jean-Claude Choul, « Le traitement de l'information dans un modèle de lecture », *Canadian Journal of Information Science*, vol. 7 (1982), 57-68.

communication fondatrice qui nous oblige à saisir le lecteur comme un coénonciateur du texte original[23].

« aboutness » et « meaning »

Nous estimons qu'il serait nécessaire de repenser la dichotomie classique en science de l'information entre « aboutness » et « meaning »[24]. Le premier terme témoigne d'une approche immanente: tentative d'une saisie de la stabilité sémantique d'un texte nonobstant l'usage (compétence pragmatique du lecteur, contexte, etc.). Le « meaning » quant à lui renvoie à l'inéluctable dérive du sens dont plusieurs rendent compte en invoquant le jeu de la semiosis illimitée.

La première approche correspond au règne absolu du texte. La seconde approche, qui pose l'investissement du lecteur dans le texte, rend caduque l'idée même de sens. Pour les spécialistes de l'indexation dont la tâche consiste précisément à représenter de façon cohérente le contenu des documents, une telle position apparaîtra plutôt problématique.

Il est néanmoins possible de penser à une position intermédiaire (du moins dans l'optique de l'indexation et de la rédaction de résumés) qui réconcilierait « l'objectivisme » des uns et le « subjectivisme » des autres. Il s'agit de cerner la stabilité du contenu d'un texte, non pas en vertu d'un principe immanent (« aboutness »), mais du fait de l'inscription des corpus textuels dans des contextes de communication.

Pour ce faire, on dégagera le sens d'un texte non plus sous l'angle de l'expressivité (celle de l'intention de l'auteur ou encore celle de l'ensemble des lecteurs aux milliers de besoins à jamais indiscernables) mais sous l'angle de la configuration discursive des champs du savoir (superstructure et catégories de discours précis). Dans cette perspective, l'analyse du discours juridique de Diane Poirier[25] et l'exploration du domaine de l'ingénierie de Rodecu Superceanu[26] font figure d'avant-garde.

Il ne s'agit plus de saisir les discours sous l'angle de la prolifération (explosion documentaire) mais par le biais de la mise à jour des contraintes propres à chaque domaine du savoir (les conditions de possibilité du savoir ainsi que les règles de production d'énoncés jugés valides dans ces différents champs)[27].

De cette façon, il semble que l'on pourra éviter l'écueil de la sempiternelle satisfaction des besoins de l'usager. Nous ne suggérons pas de produire des résumés qui feraient abstraction du lecteur, mais plutôt de procéder à une étude des catégories propres à chaque domaine du savoir, de manière à permettre, justement, de déceler le lecteur modèle en construction dans ces catégories de corpus textuels et d'ajuster en conséquence la rédaction des résumés et l'indexation des documents.

23. Umberto Eco, *Lector in fabula...* et « Notes sur la sémiotique... » ; Gilles Thérien, « Le discours littéraire » (chap. 2), in *Sémiologies*, Montréal, Les cahiers du département d'études littéraires, Université du Québec à Montréal, 1985, p. 65-97.
24. Pour une discussion sur ces notions en science de l'information, on consultera C. Begthol, « Bibliographic Classification Theory and Text Linguistics... »
25. Diane Poirier, *Des résumés adéquats pour la jurisprudence québécoise*, Montréal, Université de Montréal, École de bibliothéconomie et des sciences de l'information, 1985. Mémoire de maîtrise.
26. Rodecu Superceanu, « Super-Structural Categories in Scientific Abstracts », *Revue roumaine de linguistique*, vol. 31, no 5 (1986), 403-411.
27. Pour bien saisir la perspective de l'analyse du discours, on consultera : Dominique Maingueneau, *Nouvelles tendances en analyse du discours*, Paris, Hachette, 1987, 144 p. Voir aussi J.-C. Gardin, *Systèmes experts et sciences humaines*, Paris, Eyrolles, 1987, 269 p. et R. Superceanu, « Super-Structural Categories... »