Documentation et bibliothèques

DOCUMENTATION BIBLIOTHÈQUES

La recherche textuelle : un choix technique et administratif

Yves Hudon

Volume 37, Number 2, April–June 1991

URI: https://id.erudit.org/iderudit/1028450ar
DOI: https://doi.org/10.7202/1028450ar

See table of contents

Publisher(s)

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

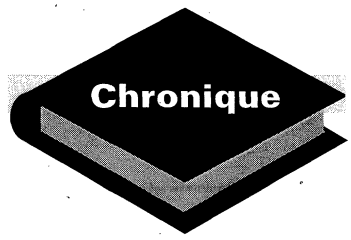
ISSN

0315-2340 (print)
2291-8949 (digital)

Explore this journal

Cite this document

Hudon, Y. (1991). La recherche textuelle : un choix technique et administratif. *Documentation et bibliothèques*, *37*(2), 73–77. https://doi.org/10.7202/1028450ar



La recherche textuelle : un choix technique et administratif

Depuis les premières tentatives égyptiennes de gérer les 700 000 documents de la bibliothèque d'Alexandrie, les préoccupations en regard des écrits étaient sensiblement demeurées les mêmes : le classement, le repérage et la restitution de documents. L'avènement de la technologie moderne a fait exploser le champ d'action de la gestion documentaire en facilitant la gestion du document et surtout en rendant possible l'accès direct à son contenu. Dans les organisations privées et publiques, en particulier, se manifeste un intérêt renouvelé pour la gestion documentaire à l'aide de nouveaux outils. Toutefois, il n'y a que de très rares études comparatives dans le domaine et les évaluations techniques de ce type de logiciels sont actuellement quasi inexistantes.

Un domaine nouveau

Le professionnel du texte (documentaliste, linguiste, etc.) s'attaque maintenant à la comparaison de textes, à leur enrichissement (annotations), à leur analyse, voire à la recherche sur le contenu des textes de l'organisation[1].

Bien que les logiciels disponibles permettent de réaliser des travaux textuels utiles et fort rentables, on se rend vite à l'évidence que la recherche et le développement ont encore beaucoup à faire avant de fournir un produit fiable à 100 %.

Il n'y a pas actuellement de logiciels qui permettent de répondre à l'ensemble des besoins d'une organisation, et la tendance actuelle semble s'orienter plus vers la spécialisation des logiciels qu'à leur généralisation : ce n'est pas que les logiciels soient inefficaces, mais la complexité du domaine textuel ne semble pas avoir livré tous ses secrets.

Il faut faire des choix selon les besoins et les possibilités techniques particulières des logiciels sur le marché.

Les logiciels

Les logiciels de recherche textuelle peuvent être associés à un besoin spécifique de l'organisation. On peut empiriquement les regrouper selon les trois principales questions que se pose l'organisation en regard des textes qu'elle produit ou qu'elle est appelée à consulter.

Face à un texte, les requêtes d'information sont principalement les suivantes :

- ***Où est le document ?*** Cette approche peut être reconnue comme la gestion documentaire sous sa forme première, le repérage de documents. Les logiciels qui y répondent sont généralement très au point et permettent de gérer efficacement bibliothèques, dossiers et documents administratifs avec une grande efficacité.

- ***Où se trouve telle information dans le texte ?*** Il s'agit là du premier niveau de la recherche textuelle : le repérage textuel. C'est le besoin que la plupart des logiciels commercialisés visent à combler, car il répond généralement à 70 ou 80 % des besoins informationnels sur les textes d'une organisation.

La recherche s'effectue le plus souvent à l'aide de mots clés, d'expressions ou de termes complexes, identifiés parfois à l'aide de dictionnaires ou de thésaurus.

- ***De quoi parle-t-on dans le document ?*** Cette approche, celle de l'analyse textuelle, est fort différente du repérage de textes. Le logiciel de recherche ou d'analyse doit posséder des fonctionnalités capables de reconnaître et d'analyser divers aspects linguistiques du texte pour en résoudre, par exemple, les ambiguïtés de termes ou d'expressions. Il doit analyser les concepts, les notions, les actions sous-entendues, ainsi que leur structure comme actes de discours ou de communication.

Une telle approche linguistique de la recherche textuelle caractérise plus particulièrement les logiciels de recherche visant principalement à extraire et synthétiser la connaissance des textes. Ces logiciels fournissent une plus grande précision dans la recherche textuelle mais sont généralement plus difficiles d'accès en raison de la complexité du langage. De plus, leur utilisation est encore onéreuse sur une grande échelle. À titre d'exemple, on les utilisera efficacement pour créer des index sujets, ou pour étudier le contenu lexical du corpus de textes.

Bien que cette optique de la recherche occupe à peine 20 % du champ de la recherche textuelle, son rôle est vite pressenti par les professionnels du texte comme indispensable, particulièrement lorsque les textes touchent des domaines variés ou lorsque la quantité de textes est très grande.

L'utilisation d'index, par exemple, s'avérera très utile, voire indispensable, lorsque l'on désirera avoir une image complète des expressions et notions véhiculées à l'intérieur d'une organisation. D'une certaine manière, cette approche permettra de mettre en lumière les domaines de connaissance dans lesquels s'intègre l'activité de l'organisation.

La problématique du choix

Le choix peut paraître simple au départ, car les logiciels de repérage de textes, principalement les logiciels commercialisés, sont bien supportés ; formation, aide technique, etc. Les logiciels d'analyse, caractérisant plus spécifiquement les logiciels de recherche universitaire, sont par contre plus difficiles d'accès.

Mais alors ceci implique un autre choix qui ne vient pas faciliter les choses.

La grande majorité des logiciels de repérage, principalement les logiciels commercialisés, mettent l'emphase sur la rapidité d'accès à de grands volumes de données textuelles, mais souvent aux dépens de la précision et de la fiabilité des résultats de la

1. L'auteur s'est appuyé sur deux livres de références essentiels : I. Lancashire and W. McCarty, *The Humanities Computing Yearbook*, Oxford, Clarendon Press,1988 et celui de W. Saffady, *Text Storage and Retrieval Systems : a Technology Survey and Product*, Wesport, Meckler, 1989.

requête. Ce n'est pas que la recherche n'y est pas efficace, loin de là. Le logiciel trouvera tout ce qu'on aura précisément demandé. Toutefois, l'écriture est beaucoup plus complexe qu'il peut y paraître à première vue et les ambiguïtés contextuelles, facilement éliminées par la pensée humaine et en fait presque toujours présentes dans le texte, sont souvent ignorées par le logiciel de repérage des mots ou expressions. La connaissance incluse dans le texte est plus qu'une simple jonction de mots.

Un exemple simple d'ambiguïté du contenu informatif sont les pronoms personnels dans un texte (il, lui, on...) ou tout autre mot qui n'a aucun sens *hors contexte.*

La recherche textuelle supposera donc, dans la grande majorité des cas, un compromis entre la rapidité d'accès à de forts volumes de textes, la précision et la validité des résultats.

Cet aspect prendra d'autant plus d'importance que le corpus de textes est volumineux ou de contextes variés, ainsi que l'illustre le graphique de la figure 1.

La courbe de la figure 1 met en évidence une certaine forme d'incompatibilité entre le fait de travailler sur de gros corpus et l'exigence d'un niveau de précision de la recherche textuelle. On est particulièrement confronté avec cette réalité dans des textes à caractères légaux où la précision des résultats ne peut tolérer qu'une faible marge d'erreur.

Pour bien comprendre et évaluer la problématique de la précision en rapport avec les besoins de l'organisation, il peut être d'une certaine utilité d'avoir une connaissance de la terminologie du domaine. Il est à noter que les termes utilisés varient selon les auteurs mais que les notions demeurent sensiblement les mêmes.

La précision. C'est le rapport du nombre total de textes retenus comme pertinents par l'usager sur l'ensemble des textes repérés par le logiciel de recherche textuelle.

La pertinence. Cette notion est très relative ; la pertinence d'un document dépend du jugement de celui, requérant ou expert, qui sélectionne des documents repérés par le logiciel. C'est une notion importante, comme on le verra dans le cas de l'ambiguïté.

Bien que la pertinence laisse place à beaucoup d'interprétation en rapport avec la précision d'un logiciel de recherche textuelle, elle est toutefois primordiale pour évaluer la capacité d'un logiciel à répondre au besoin d'un usager. Le jugement de valeur sur la capacité de pertinence d'un logiciel sera généralement renforcé par celui de plusieurs spécialistes sur la pertinence des documents repérés en rapport avec un ensemble de requêtes types.

Les candidats (textes ou expressions). Il s'agit de tous les textes ou expressions susceptibles de repérage par le logiciel à la suite d'une requête.

Le bruit ou les rejets. Ce sont tous les textes ou expressions repérés et non pertinents à la requête. Ces impertinents, pourrait-on dire, sont souvent la source de problèmes. Ils croissent rapidement avec le niveau d'ambiguïté des textes au point de rendre parfois la requête non significative ou le travail de sélection des résultats de la requête fort ardu.

Les silences. Ce sont tous les candidats non repérés par le logiciel. On en parle généralement très peu, car le contrôle en est presque impossible, du moins dans un grand corpus de textes.

C'est à ce niveau que l'analyse et la création d'index, de lexiques de mots ou d'expressions complexes deviennent des outils utiles voire indispensables lorsque l'on juge que la précision d'un logiciel de repérage ne correspond pas aux exigences de la tâche à accomplir.

Les relations entre ces définitions sont illustrées au tableau 1.

Domaines d'application

Les principaux domaines d'application de la recherche textuelle automatisée sont : le repérage de textes ou de parties de textes, leur comparaison, l'aide à la traduction, la correction de fautes (orthographiques ou grammaticales), l'enrichissement du texte (ajout d'informations, les annotations, la synonymie, la précision d'un terme), l'analyse linguistique (syntaxique, morphologique...) et la recherche sur le contenu informatif.

En fait, tous les logiciels utilisés pour ces applications sont reliés au domaine du repérage. Ces outils spécifiques appliqués à la comparaison de textes jusqu'à l'hypertexte sont en réalité des outils complémentaires qui viennent raffiner le repérage.

Dans un premier temps, il est essentiel d'évaluer l'ensemble des textes pouvant faire l'objet de recherche textuelle en fonction de ces applications. Dès lors, on peut commencer à spécifier la recherche en fonction de types de documents.

Types de documents et recherche textuelle

En recherche textuelle, la principale difficulté que doivent surmonter les logiciels est l'ambiguïté linguistique des textes. Mais tous les textes n'ont pas le même niveau d'ambiguïté. C'est pour cette raison que l'on voit parfois des logiciels faire merveille dans une application et décevoir dans une autre.

Comment définir les textes a priori ? Cela dépend du contexte. Un texte peut ne pas être ambigu en droit et l'être dans un autre domaine, comme en politique ou en génie civil. Pour choisir un logiciel de repérage susceptible de rencontrer les attentes de l'organisation face aux textes, on pourrait les classer en fonction de leur niveau d'ambiguïté linguistique et en tenant compte de la tâche à faire, de l'objectif de communication et enfin du niveau de connaissance ou d'habileté de l'utilisateur ou du destinataire.

Bref, il faut porter attention à tous les éléments d'ambiguïté dans un contexte réel d'opération à l'intérieur de l'organisation.

Dans cette optique, on peut suggérer que les textes de l'organisation soient regroupés de la façon suivante : 1) Les documents dont le lexique (termes utilisés) est normalisé ou restreint (textes prescriptifs, textes de loi, procédures) ; 2) Les documents dont le lexique n'est pas normalisé mais limité et dont le niveau d'ambiguïté textuelle peut être estimé de bas niveau ou de niveau moyen (contexte précis ou restreint) tels les textes informatifs, livres, rapports, documents spécialisés ; 3) Les textes à fort niveau d'ambiguïté comme les procès-verbaux, les comptes rendus verbaux

FIGURE 1

Les logiciels orientés/volume
VS
les logiciels orientés/précision (contenu informatif)

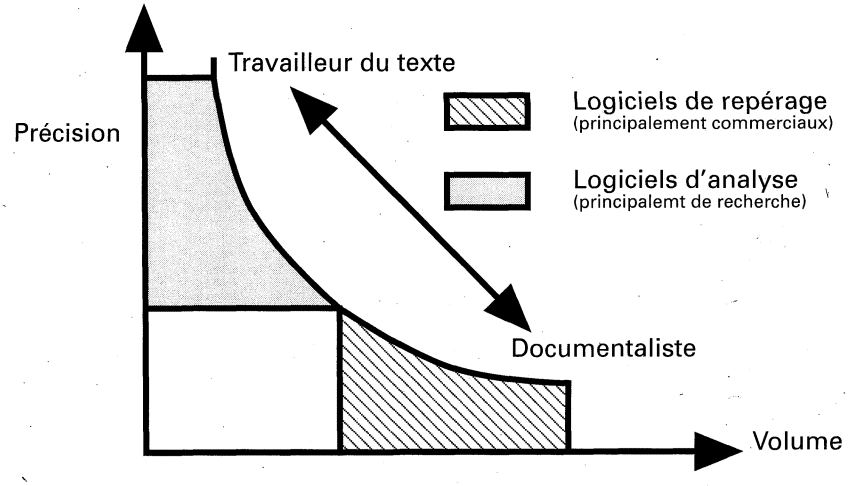
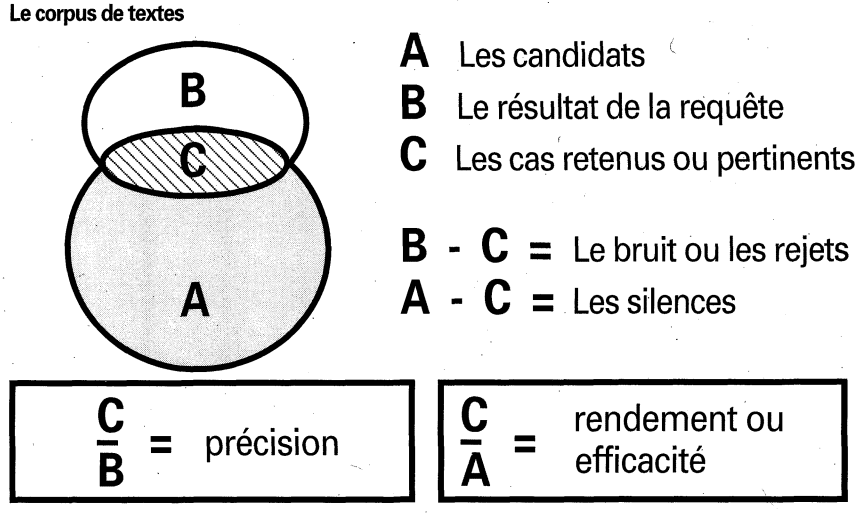


TABLEAU 1

Relation entre quelques définitions



Commentaire

Il serait de toute évidence plus logique de considérer **C/A** comme calcul de précision, mais cela ne semble valable que lorsqu'il est techniquement possible de calculer **A** ou de vérifier la totalité des candidats possibles. Ce qui fera dire à plus d'un spécialiste du domaine que la précision augmente plus le corpus de textes est petit. Car la possibilité d'évaluer **A** diminue avec l'accroissement du volume de textes. Voilà pourquoi, en pratique, on a tendance actuellement à calculer la précision sur les cas retenus, du moins sur de grands volumes de textes. Il serait intéressant de vérifier si le facteur **C/A** est toujours du même ordre que le facteur **C/B**. Cette question pourrait faire l'objet d'une recherche intéressante. Il est fort probable qu'un logiciel capable de donner des résultats avec un haut facteur de précision **C/B** est un logiciel qui a pu résoudre une part relativement importante des ambiguïtés textuelles et est donc susceptible d'avoir un bon rendement **C/A**.

ou encore les textes dont le contenu touche à des domaines multiples.

Dans le premier cas, il sera en général facile de faire du repérage textuel et les divers domaines d'application apparaîtront comme des outils de raffinement pratiques mais non essentiels.

Toutefois, dans les deux autres contextes, le repérage exigera, selon le niveau de complexité des textes, des outils d'enrichissement ou d'analyse selon le cas, outils qui auront pour but de diminuer le facteur d'imprécision sur les candidats possibles à la requête. Dans cette situation, il pourra être utile de classer les divers types de logiciels selon la proposition illustrée au tableau 2.

Le niveau de complexité des textes joue un rôle majeur dans le choix d'un logiciel, car il est directement relié au niveau de bruit ou de rejet, à savoir la précision ou la pertinence des résultats de la requête textuelle. Le choix d'un logiciel implique donc une bonne évaluation des types de textes de l'organisation pour déterminer le niveau de précision utile, nécessaire ou essentiel selon le cas.

Il est à noter que le contexte doit aussi être pondéré en fonction du besoin réel de précision de l'organisation et ce, indépendamment de celui du texte. Ainsi les textes pour lesquels on exige des résultats d'un niveau minimal de précision seront considérés de type 1. Et ceux pour lesquels on exige une précision relativement importante pourront être considérés de type 2.

Toutefois, le facteur de précision n'est pas invoqué dans le dernier cas en raison du fort niveau d'ambiguïté de cette catégorie de textes. L'analyse ou l'enrichissement du corpus est alors nécessaire a priori, si l'on désire une précision significative.

Les principales fonctionnalités des logiciels

Il serait trop long d'analyser ici tous les éléments fonctionnels qui peuvent caractériser un logiciel. Du reste, là n'est pas le propos du présent article. Mais il peut être essentiel d'en avoir une idée pour bien aborder la

délicate tâche de choisir un logiciel de recherche textuelle.

Voici donc les principales caractéristiques des 40 logiciels pour lesquels nous avions des données suffisantes[2]. Dans le tableau 3, les caractéristiques sont notées selon leur fréquence relative d'apparition dans les produits qui ont fait l'objet de l'étude. Les caractéristiques plus rares, comme les analyseurs lexicométriques, les thésaurus, l'indexation assistée (enrichissement de textes), voire l'utilisation de l'intelligence artificielle, se retrouvent généralement dans les logiciels qui permettent une plus grande précision dans la recherche textuelle.

Implications financières

En matière de documentation, à ce stade-ci de l'évolution du besoin dans les organisations, le repérage textuel répond à la majorité des attentes à des coûts souvent avantageux par rapport aux logiciels d'analyse textuelle. Toutefois, les informations recherchées sont alors plus générales et l'impact administratif et financier des résultats se doit d'être généralement de moindre importance ou de porter moins à conséquence.

Le repérage basé sur l'analyse terminologique ou nécessitant un enrichissement du texte est plus coûteux (exige plus d'efforts), mais les économies en temps de recherche effectuée par les professionnels justifieront avantageusement ce type de technologie (figure 2).

En somme, le repérage se justifie surtout par son efficacité comme outil d'information. De son côté, l'analyse textuelle, compte tenu du fait que les efforts nécessaires augmentent rapidement au-dessus d'un certain seuil de précision, se rentabilise par l'économie de temps de travail, par la facilité de manipulation des textes et par la nécessité, parfois impérative, d'obtenir des données précises et surtout valides.

2. Parmi les logiciels ayant fait l'objet de recherche, on trouve: ADABASE, COBA, DARWIN, EDIBASE, NATUREL, PROTEXTE, SATO, STAIRS, TERMINO, TEXMATE, ZYINDEX. L'auteur sera heureux d'en fournir la liste complète aux lecteurs intéressés.

TABLEAU 2

Proposition de classification des logiciels selon le contexte

Critères	LOGICIELS — niveau de recherche — Repérage	Enrichissement	Analyse — Linguistique	Contenu
VOLUME	****	*	*	**
PRÉCISION	*	***	**	***
$ d'opération	$	$$$	$$	$$
	*** niveau d'efficacité vs champ d'application			
Textes				
1 - Faible niveau d'ambiguïté (textes normalisés) (vocabulaire restreint)	très efficace	peu utile	peu utile	utile
2 - niveau d'ambiguïté-faible niveau d'ambiguïté-moyen	efficacité restreinte faible rendement	utile très utile	utile utile	utile très utile
3 - haut niveau d'ambiguïté diversité de domaines	non recommandé (sans enrichissement ni analyse de contenu)	INDISPENSABLE	utile	très utile

TABLEAU 3

Les principales fonctionnalités des logiciels de recherche textuelle

F - très fréquent C - courant R - rare

La requête		**L'indexation** des fichiers	
- par mots clés	F	- automatique à l'entrée du texte	C
- par mots libres	C	- automatique sur demande	R
- en langage naturel (phrases libres)	R	- automatique en différé	C
- par ordre strict de mots (au choix de l'usager)	C	- semi-automatique ou assistée	R
- **simple** (une requête possible)	F	**Lexiques** (utilisation de répertoires)	F
- **multiple** (suite de requêtes ou sous-questions)	C	- des mots (dictionnaire inclus)	C
- **mémorisation** d'un processus de questions (au choix de l'usager)	R	- des mots bloqués (ou complexes)	C
		- des mots cachés (anti-dictionnaire)	C
Utilisation d'opérateurs		- de synonymes (dictionnaire)	R
- logiques	C	**Thésaurus** (capacité de gérer un thésaurus)	R
- booléens	C	**Analyseurs lexicométriques** (compilation sur les mots)	R
- numériques	R	- de fréquences	
Troncature à droite	F	- de co-occurences	
Troncature à gauche	F	- de pertinence	
Troncature à gauche et à droite	C	- de proximité	
Troncature au centre	C	**Sur divers alphabets**	R
Troncature sur plusieurs mots	R	Capacité de recherche*	
Utilisation de principes d'**I.A.** (Intelligence artificielle)	R		

* Pour ce qui est de la capacité de travailler sur divers alphabets ou types de caractères, on comprendra la nécessité de choisir un logiciel adapté aux caractères français.

Conclusion

Le gestionnaire devrait-il, dès le départ, envisager la possibilité d'utiliser plus d'un logiciel? À la lumière de l'étude, il nous apparaît prématuré d'envisager une solution unique à la recherche textuelle avant quelques années. L'usage de logiciels d'analyse ou d'enrichissement, relativement dispendieux, sera ordinairement réservé aux professionnels dont le travail est lié aux textes de l'organisation ou encore à un usage spécialisé.

Les logiciels de repérage conviendront, pour leur part, à un usage étendu sur des requêtes d'intérêt plus général. Enfin, faire le choix d'un logiciel de recherche textuelle exige une bonne connaissance de la nature des documents de l'organisation et du niveau de complexité des questions ou requêtes utiles, voire nécessaires à l'organisation.

Bien sûr, avec l'évolution du domaine et des besoins des organisations, les éléments de comparaison comme ceux abordés deviendront eux aussi plus complexes. C'est pourquoi nous nous en sommes tenus à ce qui semble être les critères élémentaires de décision pour le choix d'un logiciel dans l'état actuel de l'art.

Avec l'intégration récente de l'image et même de la voix au textuel, on imagine d'ores et déjà la place croissante qu'il faudra consacrer, au cours de la prochaine décennie, à l'approfondissement de la gestion documentaire et des nouvelles techniques afférentes.

Yves Hudon
Direction des logiciels d'application
Ministère des Communications
du Québec
Québec

FIGURE 2

L'effort relatif au besoin de repérage et d'analyse