Documentation et bibliothèques

Nouveaux horizons en indexation automatique de monographies
New Horizons in the Automatic Indexing of Monographs
Nuevos horizontes en la indexación automática de monografías

Lyne Da Sylva

Volume 48, Number 4, October–December 2002

URI: https://id.erudit.org/iderudit/1030353ar
DOI: https://doi.org/10.7202/1030353ar

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 article

Da Sylva, L. (2002). Nouveaux horizons en indexation automatique de monographies. *Documentation et bibliothèques*, *48*(4), 155–167. https://doi.org/10.7202/1030353ar

Article abstract

What is the state of automatic indexing of monographs? The first attempts at automatic indexing in the early 1960s did not always produce, according to professional indexers, satisfactory systems.

However, it is worth one's while to re-examine the opportunities offered by automatic indexing given the growing number of numeric documents for which it could be possible to provide an index similar to the more familiar the back-of-the-book indexes. Several important innovations in the field of the automatic processing of language have been developed over the last fifteen years and new applications could be used for automatic indexing of monographs. This article outlines the issues and identifies solutions to improve the current systems of automatic indexing of monographs.

Nouveaux horizons en indexation automatique de monographies

Lyne Da Sylva
Professeure adjointe
EBSI, Université de Montréal
lyne.da.sylva@umontreal.ca

Quel est l'état de la question en indexation automatique de monographies ? Bien que les premières tentatives d'indexation automatique datent du début des années 1960, elles n'ont toujours pas abouti à des systèmes satisfaisants du point de vue des indexeurs professionnels.
Pourtant il y a lieu de s'interroger sur les possibilités actuelles d'indexation automatique, compte tenu du nombre croissant de documents numériques pour lesquels il serait intéressant de fournir un index comme celui qu'on trouve à la fin d'un livre (back-of-the-book index). *En outre, les quinze dernières années ont vu des innovations importantes dans le domaine du traitement automatique des langues (TAL), qui pourraient avoir des applications avantageuses pour l'indexation automatique de monographies. Cet article propose de définir la problématique et d'identifier les nouvelles pistes de solutions à explorer afin de dépasser les performances des systèmes actuellement offerts pour l'indexation automatique de monographies.*

New Horizons in the Automatic Indexing of Monographs
What is the state of automatic indexing of monographs ? The first attempts at automatic indexing in the early 1960s did not always produce, according to professional indexers, satisfactory systems.
However, it is worth one's while to re-examine the opportunities offered by automatic indexing given the growing number of numeric documents for which it could be possible to provide an index similar to the more familiar the back-of-the-book indexes. Several important innovations in the field of the automatic processing of language have been developed over the last fifteen years and new applications could be used for automatic indexing of monographs. This article outlines the issues and identifies solutions to improve the current systems of automatic indexing of monographs.

Nuevos horizontes en la indexación automática de monografías
¿Cuál es la situación actual de la indexación automática de las monografías ? Si bien las primeras tentativas de indexación automática datan de comienzos de la década de los años 60, no siempre resultaron en sistemas satisfactorios, según los especialistas en este campo.
Sin embargo, cabe interrogarse sobre las posibilidades actuales de la indexación automática, dado el número crecientes de documentos digitales para los cuales podría ser interesante proveer un índice al final de un libro (back-of-the-book index). Además, los últimos quince años fueron testigos de innovaciones importantes en el campo del tratamiento automático de los idiomas (TAL), lo que podría tener aplicaciones interesantes en la indexación automática de las monografías. Este artículo propone definir la problemática y determinar nuevas pistas de soluciones que se deben explorar para superar el nivel de desempeño de los sistemas actuales destinados a la indexación automática de monografías.

Quelques définitions

L'introduction de quelques précisions sur des termes clés utilisés dans cet article servira à brosser un tableau rapide des recherches antérieures dans ce domaine.

Indexation de monographies ou indexation microscopique

L'indexation de monographies présente une problématique assez différente de l'indexation dite « pour les bases de données ». Dans le deuxième cas, que nous appellerons indexation macroscopique, la finalité de l'indexation est de produire une liste plutôt courte de descripteurs (ou termes d'indexation) qui saisissent l'essentiel des sujets traités dans le document, dans le but de permettre le repérage du document dans une collection.

L'indexation de monographies (ou indexation microscopique[1]), par contre, vise à produire un index plutôt exhaustif contenant la majorité des sujets importants pour lesquels il existe un passage informatif dans l'ouvrage. Les notions de « passage informatif » et de « sujet important » sont difficiles à définir avec précision, mais elles sont primordiales : l'occurrence d'un terme à une page donnée de l'ouvrage n'implique pas qu'une entrée à l'index fera référence à cette page, même si ce terme représente un sujet important dans le document. Le passage peut être jugé comme n'étant « pas suffisamment informatif ». Cette distinction représente un défi pour l'indexation automatique qui ne peut tenir compte du contexte que de façon limitée pour déterminer quelles occurrences de termes « importants » mériteront une référence dans l'index.

L'orientation des deux types d'indexation est très différente, et d'ailleurs les professionnels se spécialisent habituellement dans un type à l'exclusion de l'autre. Parmi les différences principales, on note la fonction de l'index (repérer un document à l'intérieur d'une collection plutôt qu'un passage à l'intérieur d'un document), le nombre d'entrées d'index générées (normalement en deçà d'une vingtaine pour l'indexation d'un document d'une collection contre plusieurs centaines pour une monographie), le niveau d'exhaustivité différent qui en découle, le vocabulaire utilisé (vocabulaire contrôlé dans le premier cas, libre dans le second) et la forme des entrées (bien que celle-ci soit largement tributaire du langage documentaire, dans le premier cas). Il est utile de mettre l'accent sur cette dernière différence de laquelle découlent

1. Les préfixes micro- et macro-, pour distinguer les deux types d'indexation, sont utilisés par d'autres auteurs préoccupés par la même problématique, dont Lanteigne (1994b). Voir aussi Klement (2001).

des techniques assez différentes pour produire automatiquement les deux types d'index.

Un index à la fin d'un livre est caractérisé par la structure de ses entrées : chacune est constituée d'une vedette principale, éventuellement accompagnée de sous-vedettes ; les entrées peuvent être reliées par des renvois.

Regardons d'abord l'organisation des entrées : celle-ci permet l'expression de relations sémantiques entre la vedette et les sous-vedettes, telle que l'hyperonymie illustrée dans l'exemple (1) ci-dessous.

(1) Mammifère
Félins
Canidés
Bovidés

Elle exprime en même temps la différenciation de divers aspects de la vedette principale par les sous-vedettes de l'entrée ; voir l'exemple (2) ci-dessous (tiré de l'index dans Abeillé 1993, 316).

(2) *Structure*
argumentale : 43, 74-75.
de traits : 11, 16-17, 22-24, 29-36, 152-158, 162-173, 190, 195, 302.
de constituants (LFG) : 43-47, 63-64, 89, 91.
fonctionnelle (LFG) : 24, 43-51, 60-64, 78, 84, 89, 91.
profonde : 10, 12, 26, 64, 235 (n. 118).
de surface : 10, 12, 14, 19-22, 41, 63, 75, 201, 235 (n. 118).

En l'absence de subdivisions de la vedette *structure*, un utilisateur serait bien embêté d'avoir à consulter chacune des multiples références de pages afin de déterminer celles qui parlent véritablement et uniquement du type de *structure* qui l'intéresse. Cet exemple illustre à lui seul l'utilité des subdivisions, à laquelle nous venons de faire référence, et le problème des longues listes de références non différenciées, illustré ici par la sous-vedette *de traits*.

De même, l'organisation des entrées peut regrouper sous la même vedette principale des sous-vedettes reliées qui seraient dispersées par l'ordre alphabétique. Par exemple, dans un ouvrage où l'on retrouverait des mentions des diverses institutions représentant un pays à l'étranger, on pourrait recenser les concepts « ambassades du Canada », « consulats du Canada », « missions du Canada », et les regrouper dans une seule entrée sous la vedette principale « Canada ».

Par ailleurs, les renvois permettent à l'utilisateur de repérer des concepts reliés à une entrée donnée, mais qui ont été placés à l'index dans une entrée différente. Certains renvois font des rapprochements entre les entrées sémantiquement proches, comme les renvois de type *voir sous*, par exemple (Abeillé 1993, 313) :

(3) *Optionnalité (voir aussi Complément)*

Les renvois de type *voir* relient, eux, des synonymes ou des expressions équivalentes (Abeillé 1993, 310) :

(4) *FUG* : voir *Grammaire fonctionnelle d'unification*

L'index incorpore ainsi de différentes façons les rapprochements et les divergences sémantiques entre des concepts reliés. Un rôle équivalent est joué par le langage documentaire, dans le cas d'indexation macroscopique avec langage contrôlé ; cependant, l'indexation d'une monographie est habituellement faite en vocabulaire libre, et les relations sémantiques doivent toutes être fournies par l'indexeur.

L'indexation microscopique et l'indexation macroscopique présentent ainsi plusieurs différences importantes. Les articles de Klement (2002) et de Wellisch (1994) présentent toutefois des discussions très intéressantes sur la distinction entre les deux types d'indexation ; alors que Klement fait une opposition systématique entre les deux, selon une quarantaine de caractéristiques, Wellisch défend plutôt, entre ces deux types, un continuum de produits et de tâches d'indexation selon le type de texte. Notre propre position quant aux méthodes efficaces pour aborder l'indexation microscopique s'inspirera, paradoxalement, des deux points de vue opposés présentés dans ces articles.

Comme le note Lanteigne (1994a, 43), l'indexation de monographies serait « le parent pauvre » de la recherche en indexation[2]. Comme on le verra plus loin, ce déséquilibre est présent également dans les travaux en indexation automatique.

Monographie

Il convient de préciser ce que nous entendons ici par monographie. La définition du Petit Robert, « *Étude complète et détaillée qui se propose d'épuiser un sujet précis relativement restreint (monographie d'une région, d'un personnage historique)* », met l'accent sur le contenu, le sujet. L'AFNOR, dans la norme Z 44-050, définit une monographie plutôt comme un « *ouvrage formant un tout, en un ou plusieurs volumes, soit qu'il paraisse en une seule fois, soit que sa publication s'étende sur une durée limitée selon un plan établi à l'avance* ». On y fait donc référence davantage au mode de publication. Dans ce deuxième cas, la notion d'ouvrage imprimé est implicite.

Est-il utile d'être plus précis ? Wellisch (1994), ayant étudié la question, arrive à la conclusion qu'il est difficile de tracer une frontière claire entre une monographie et d'autres types de documents (écrits), dans un continuum comprenant les monographies classiques, les ouvrages écrits par plusieurs auteurs, les monographies en plusieurs tomes, les encyclopédies, etc., jusqu'aux périodiques. Ce qui nous pousse à vouloir définir le terme davantage, c'est le besoin de caractériser le type d'ouvrage pour lequel il serait utile de produire automatiquement un index microscopique. C'est dans cette optique que nous redéfinirons le terme « monographie », auquel nous accorderons un sens plus large que ce qui est conventionnellement admis. Une monographie consiste, pour nous, en une œuvre textuelle écrite en un tout, sur une thématique plutôt homogène (bien que cet aspect puisse ne pas être exigé), indépendamment de son format ou médium. Bien sûr, l'œuvre visée doit être d'une « certaine » longueur, afin qu'il soit jugé utile d'avoir un outil d'aide au repérage

2. Outre les ouvrages pédagogiques, les articles scientifiques se concentrent généralement sur des critiques de logiciels ou des comptes rendus de livres. Une recherche dans la base de données *Library and Information Science Abstracts* (LISA) n'indique que deux publications récentes portant les descripteurs « index de livres » et « recherche ».

d'information, et, forcément, elle doit être disponible en format numérique.

Notre définition exclut les documents numériques que sont les périodiques électroniques, les ouvrages en cours (comme les œuvres de fiction collaboratives sur Internet, pour lesquelles la validité de tout index serait de courte durée) ainsi que les documents de petite taille qui peuvent être parcourus aisément, à l'écran ou sur papier. Et plus spécifiquement, un article plutôt long peut être considéré comme une monographie, aussi bien qu'un site Web en entier ou une page isolée du site Web, si sa longueur le justifie. C'est justement pour ces derniers types de documents que l'on juge pertinent de produire automatiquement un index. Pour l'instant, les outils d'aide au repérage pour ces documents sont généralement la fonction de recherche plein texte (si disponible) et l'indexation fournie par les moteurs de recherche sur Internet. Or, ces outils présentent des limitations importantes et n'offrent pas le coup d'œil rapide sur un texte que permet un index bien conçu (voir notamment à ce sujet Broccoli 1998).

Indexation automatique

Il est utile de présenter aussi ce que l'on entend par « indexation automatique ». Ce concept touche à la fois les applications d'indexation microscopique et macroscopique. En effet, une partie importante des tâches effectuées est commune aux deux approches, notamment lorsqu'il s'agit de repérer dans le texte des expressions ou des termes susceptibles de constituer des entrées à l'index.

Par contre, telle que nous l'avons décrite ci-dessus, l'indexation microscopique a des particularités qui exigent certains types de traitements automatiques non pertinents pour l'indexation macroscopique. Si l'on décompose la tâche d'indexation microscopique, on peut identifier diverses étapes où l'intervention de l'ordinateur peut être bénéfique. Parallèlement, on constate une gradation dans les outils logiciels proposés pour l'indexation.

Indexation manuelle avec logiciels dédiés

Des logiciels dédiés pour effectuer l'indexation d'ouvrages imprimés gèrent les aspects mécaniques de la production d'index : la gestion des numéros de pages, le tri par ordre alphabétique, la présentation typographique sur la page, etc. Se trouvent dans cette classe, notamment, les logiciels CINDEX, Macrex et SkyIndex. Ces logiciels reposent sur l'intervention humaine pour tout le travail intellectuel d'indexation, c'est-à-dire le repérage de passages informatifs, le choix de termes utiles pour indexer ceux-ci, et le regroupement des vedettes en entrées structurées.

Bien sûr, ces tâches dites mécaniques devront aussi être effectuées par un système d'indexation automatique, mais elles sont triviales d'un point de vue computationnel.

Indexation assistée par ordinateur

L'indexation assistée par ordinateur propose d'alléger la tâche intellectuelle de l'indexation en offrant à l'analyste divers outils d'aide : accès en ligne au vocabulaire contrôlé, aux guides et manuels, à la politique d'indexation ; présentation à l'écran de « grilles d'indexation » contenant les différents champs à compléter (elles servent alors d'aide-mémoire en plus de minimiser le temps et l'effort de saisie) ; correction automatique des mots entrés au clavier ; etc. Cette approche est davantage présente, on l'aura compris, dans les contextes d'indexation macroscopique (Hodge et Milstead 1998, présentent un survol de l'informatisation des services documentaires aux États-Unis).

Pour l'indexation automatique microscopique, les applications pertinentes sont celles où les systèmes accomplissent une partie de l'identification des expressions présentes dans le texte susceptibles d'être utiles à l'indexation. Des termes candidats sont alors proposés mais doivent être validés par l'indexeur. Le logiciel SATO (voir notamment Bertrand-Gastaldy et Pagola, 1994) appartient à cette famille de logiciels. Un autre exemple encore plus représentatif est le logiciel Sonar Bookends Professional, de la firme Virginia Systems[3]. Celui-ci peut à la fois compiler une liste de mots et termes candidats, structurer (jusqu'à un certain point) les entrées en vedettes et sous-vedettes, définir des variantes à rechercher pour une entrée donnée, et présenter les termes candidats en contexte afin de permettre à l'utilisateur de conserver ou non une entrée à l'index pour cette occurrence du terme. On peut trouver dans Wright (2000) un aperçu et une critique du logiciel. Celui-ci peut être utilisé de façon complètement automatique, mais gagne à être utilisé de façon supervisée, et de ce fait représente un exemple d'indexation microscopique assistée. Il présente une brochette de fonctionnalités rarement offertes par un seul logiciel ; nous reviendrons toutefois ci-dessous sur ses limites.

Indexation complètement automatique

Dans l'indexation complètement automatique, l'objectif est d'exclure complètement l'intervention humaine de la chaîne de traitement. Le système prend en charge à la fois le repérage et la structuration de termes candidats ainsi que les aspects mécaniques d'édition, puis présente à l'utilisateur un index final qui, on l'espère, sera utile.

Nous l'avons déjà dit, le logiciel Sonar Bookends Professional représente une application d'indexation qui peut être utilisée de façon complètement automatique (mais dont on peut toutefois modifier les résultats).

L'indexation complètement automatique est mieux illustrée par l'indexation effectuée par les robots-collecteurs des moteurs de recherche comme Google, AltaVista, NorthernLight, etc. Elle consiste effectivement à recenser toute occurrence des mots des sites et pages Web, que l'on peut dès lors considérer comme des monographies plus ou moins longues. Elle permet ensuite l'accès non seulement à la page Web en question mais aussi au contexte d'occurrence du mot dans la page. C'est sur ce point qu'elle se rapproche de l'indexation microscopique. D'ailleurs, il est permis de penser qu'une des sources de frustration des utilisateurs des moteurs de recherche est le fait que la distinction entre indexation microscopique et macroscopique ne soit pas faite : l'utilisateur, par sa requête, espère davantage repérer un document sur un sujet, mais on lui présente un passage d'un document contenant les mots de sa requête, ce qui n'est pas nécessairement la même chose. On imagine le chaos qui

3. Disponible en version de démonstration à l'adresse suivante : <http://www.virginiasystems.com>.

suivrait une telle requête dans une bibliothèque...

Une indexation totalement automatique est effectuée par certains logiciels de numérisation : des systèmes de numérisation de textes avec reconnaissance optique des caractères, qui sont en très grand nombre, et des logiciels de reconnaissance vocale (comme ProDEX[4]). Enfin, les systèmes de gestion de base de données (SGBD) se servent également d'index produits automatiquement (parfois suite à une commande explicite de l'utilisateur). L'indexation automatique vise à permettre aux utilisateurs de faire des recherches dans les documents gérés par les systèmes. Le produit de ce type d'indexation est un index d'unitermes[5], qui ne peut être exploité efficacement qu'à l'aide de requêtes booléennes savamment conçues.

Ces dernières applications sont les seules pour lesquelles l'indexation microscopique entièrement automatique est réellement utilisée.

En conclusion, ce que l'on entend par indexation automatique est la prise en charge de tout le processus d'indexation, de l'analyse du texte jusqu'à la présentation des résultats à l'utilisateur, sans intervention humaine. On peut invoquer des applications où cette fonctionnalité serait justifiée et utile.

Avant de poursuivre, il convient de préciser, dans le cadre de l'indexation automatique, deux façons importantes de repérer dans un texte des termes ou expressions intéressantes : l'indexation par extraction et l'indexation par assignation.

Indexation par extraction et par assignation

La distinction entre l'indexation par extraction et l'indexation par assignation est importante lorsque l'on parle d'indexation automatique. On peut se limiter à utiliser, pour indexer, uniquement des termes explicitement présents dans le texte ; c'est l'indexation par extraction. Ou l'on peut assigner des termes (habituellement d'un vocabulaire contrôlé) en se basant sur des indicateurs présents dans le texte et faire ainsi de l'indexation par assignation. Les moteurs de recherche sont basés sur l'indexation par extraction alors que des systèmes plus sophistiqués utilisent l'indexation par assignation.

Que dire de chacune relativement à l'indexation automatique ? D'abord, l'indexation par extraction est la plus facile à mettre en place, puisqu'elle se limite au repérage des mots présents dans le texte. Il faudra peut-être séparer les suites de mots intéressantes de celles qui ne le sont pas, mais il est très facile de concevoir un logiciel qui repère ces suites de mots dans le texte. Par contre, cette approche est très limitée et ses performances sont habituellement décevantes. Elle dépend lourdement de la phraséologie utilisée par l'auteur, sa terminologie exacte, ses tournures de phrases, etc.

Par cette méthode, on repérera uniquement les mots que l'auteur a choisi d'utiliser, et non pas tout autre synonyme pertinent qu'un humain ajouterait spontanément à l'index, par exemple. L'approche de Earl (1970) en est un exemple. L'évaluation de ce système confirme son potentiel limité.

L'indexation par assignation, elle, permet une plus grande richesse d'expression des concepts traités dans le document et un repérage plus flexible. Mais elle est beaucoup plus difficile à implémenter. Il faut fournir au logiciel non seulement le vocabulaire contrôlé mais aussi les variantes des termes et les façons de repérer ces variantes dans le texte. On parle généralement ici, ou bien d'un thésaurus élaboré spécialement pour couvrir le domaine de la monographie, ou plus généralement d'un système expert qui connaît les variantes multiples à prévoir et la forme désirée pour chacune dans l'index. L'élaboration de ces outils représente un travail considérable. Les travaux suivants s'inscrivent dans ce courant : Artandi (1963), que nous décrivons ci-dessous, Dillon (1982), Driscoll *et al.* (1991), Leung et Kan (1997), Jacquemin (2001), Jacquemin *et al.* (2002).

Ces deux approches (par extraction et par assignation) s'opposent et suggèrent des méthodes différentes pour l'indexation automatique. Il convient toutefois de mentionner qu'il ne s'agit pas de deux approches incompatibles : dans bon nombre de systèmes actuels, il est difficile de tracer une frontière entre les deux types d'indexation. Par exemple, si l'on peut identifier, à partir de l'expression « gouvernement du Québec et du Canada », les termes candidats « gouvernement du Québec » et « gouvernement du Canada », comment caractériser la méthode d'identification du deuxième terme ? Ce n'est plus de l'extraction simple de mots contigus, mais la reconstruction d'une expression à partir d'indicateurs dans le texte. La distinction n'est plus toujours très claire ni très utile.

Dans les traités d'indexation de livres, on souligne l'importance de « coller » à la terminologie utilisée par l'auteur. On n'utilise généralement pas de vocabulaire contrôlé. Toutefois, un indexeur veillera à inclure dans l'index des synonymes prévisibles pour aider l'utilisateur. On pourrait alors penser qu'en indexation automatique microscopique, on peut davantage avoir recours à l'indexation par extraction et limiter l'utilisation de l'indexation par assignation aux seuls cas de synonymes des termes du domaine. Cependant, une partie importante du travail d'indexation microscopique consiste à établir des liens entre les concepts traités dans le document, en utilisant par exemple des termes généraux qui chapeautent plusieurs concepts, même si le terme général n'est pas mentionné explicitement dans le document. Ceci représentera donc un défi de taille pour l'indexation automatique, qui ne pourra prétendre s'en tirer avec la simple extraction de suites de mots pris dans le texte. C'est justement une des lacunes importantes des logiciels d'indexation automatique.

Il faut ajouter, dans les approches d'indexation automatique par assignation, les méthodes statistiques comme celle de Leung et Kan (1997) ou de Deerwester *et al.* (1990). Bien que considérablement différentes, elles visent à assigner des termes d'indexation sur la base d'une analyse statistique préalable d'un corpus apparenté. On calculera la probabilité d'assigner un terme donné compte tenu des mots du texte à indexer. Ces approches relèvent du paradigme d'indexation macroscopique et on ne voit pas clairement à l'heure actuelle comment intégrer ces résultats à une approche microscopique.

4. Décrit à l'adresse suivante : <http://www.protext.com/protext.htm. (page consultée le 26 février 2003)>.
5. On indexe parfois, dans le cas des SGBD, des « multitermes » ou plutôt des occurrences complètes des champs de la base de données.

Indexation automatique de monographies

Nous arrivons donc à définir la problématique de l'indexation automatique de monographies. Celle-ci consiste à développer des systèmes qui, sans intervention humaine, construisent pour un texte numérique un index constitué d'entrées structurées ; ces entrées font référence à un passage déterminé dans le document où apparaît une discussion importante sur le concept exprimé par l'entrée.

Le type de système visé devra atteindre au minimum les cinq objectifs suivants :

1. l'identification des concepts-clés traités dans le document (par extraction, par assignation ou par une combinaison des deux),
2. la normalisation ou l'homogénéisation des termes glanés (nécessaire afin d'éviter de disperser inutilement de légères variantes lexicales),
3. l'identification de relations sémantiques clés entre les concepts,
4. la structuration des entrées en conséquence,
5. l'édition de l'index en résultant.

Quel est l'état des travaux sur ces différents plans ? Nous brosserons un tableau du domaine en abordant d'abord les travaux de recherche, puis l'offre en matière de logiciels commerciaux.

Les travaux de recherche en indexation automatique de monographies datent du début des années 1960. Artandi (1963) propose une approche d'indexation par assignation de termes selon un vocabulaire préétabli de termes jugés intéressants. L'expérience porte sur un chapitre isolé d'un manuel de chimie. Les résultats – prometteurs aux dires de l'auteur – sont toutefois confinés à ce domaine (étant donné le thésaurus spécialisé nécessaire). Aussi, comme le fait remarquer Lanteigne (1994a), les résultats sont faussés du fait qu'ils ne portent que sur un chapitre isolé : une partie importante des problèmes de l'indexation d'une monographie découle d'avoir à unifier, à regrouper et à ordonner des références à des sujets répétés mais sous des angles différents dans plusieurs parties d'un ouvrage. Dans la recherche d'Artandi, la tâche de structuration des entrées est donc limitée.

Earl (1970) procède à l'indexation par extraction de termes. L'expérience ne porte elle aussi que sur un chapitre isolé. Les travaux de Salton (1988) proposent d'appliquer une analyse syntaxique (partielle) afin d'améliorer le repérage des expressions qui expriment les concepts traités. Par exemple, l'analyse d'une expression comme « antenne parabolique de révolution » révélerait que d'après la structure syntaxique, les mots-clés de l'expression sont « antenne » et « révolution » ; l'entrée proposée serait donc « antenne, révolution ». L'adjectif « parabolique », jugé accessoire, serait mis de côté. Ensuite, cette première expression et une deuxième, « antenne de révolution », seraient regroupées et représentées toutes les deux par la même entrée. Cela semble présenter une amélioration notable sur le repérage de multitermes complexes et difficiles à gérer, mais on voit vite que l'approche est davantage applicable à une langue comme l'anglais, où les termes sont généralement exprimés par la composition nominale (alors qu'en français, les prépositions comme « de » sont nécessaires). Encore ici, à part le regroupement d'entrées qui partagent les mêmes mots, la question de structuration des entrées est laissée de côté.

Gingras (1992) décrit un projet d'indexation automatique basé sur les capacités du logiciel SATO ; celui-ci offre diverses possibilités pour repérer des termes et constituer un vocabulaire contrôlé. Cependant, le projet relève davantage de l'indexation assistée que de l'indexation automatique, puisqu'il repose sur certaines interventions effectuées par un humain.

Plusieurs travaux de recherche en indexation macroscopique ainsi qu'en repérage d'information (*information retrieval*) sont pertinents lorsqu'il s'agit de repérer des termes ou expressions. Nous n'en relèverons pour l'instant qu'un échantillon : Smart (1992) décrit un système qui permet d'extraire des multitermes, une amélioration par rapport aux unitermes. Les travaux de Dillon et Gray (1983) augmentent la qualité de repérage des termes grâce à une analyse syntaxique. Turney (2000) s'inscrit dans le cadre de l'apprentissage machine : le système procède par algorithme génétique pour apprendre à reconnaître des expressions. Dans tous ces cas, la question de la structuration des entrées n'est évidemment pas abordée.

En termes de logiciels commerciaux, l'offre se fait plutôt rare. Le logiciel Indexicon, disponible au début des années 1990, a maintenant disparu et le logiciel Sonar Bookends est peu connu.

La rareté de l'offre découle naturellement du fait que ces logiciels s'adressent non pas à un public général, mais à des utilisateurs spécialisés : ceux qui désirent produire un index pour un livre destiné à être imprimé. Ces utilisateurs incluent les indexeurs professionnels. Or, ceux-ci mesurent l'intérêt du logiciel par rapport aux normes de qualité attendues d'un humain. Les logiciels disponibles ne sont tout simplement pas à la hauteur, comme le confirme chaque critique qui en est faite dans les revues professionnelles (voir entre autres Mulvany 1999 et Wright 2000). On demeure donc à l'état de prototypes de recherche dans la plupart des cas.

On peut s'étonner d'une offre si limitée, mais elle est bien légitime, étant donné la nature des défis auxquels les objectifs ci-dessus correspondent. Un examen plus approfondi de ces difficultés expliquera la source des lacunes observées et suggérera des pistes à suivre pour aller au-delà des limites actuelles.

Défis à relever

Reprenons les objectifs ci-dessus. Un système d'indexation automatique microscopique doit effectuer, au minimum, les tâches suivantes : l'identification des expressions intéressantes (les entrées candidates), l'homogénéisation des candidats, le repérage de relations (sémantiques) entre ces entrées candidates, le regroupement des entrées et l'édition de l'index. Qu'est-ce que chaque tâche implique ?

Identification des expressions intéressantes

L'identification des expressions intéressantes est faite lors d'une phase d'analyse du texte. Les approches utilisées peuvent être catégorisées selon divers paramètres, desquels nous retiendrons le type d'expressions recherchées et la façon d'identifier les expressions intéressantes ; nous discutons de chacun ci-dessous. La question de la fréquence

d'occurrence des termes est importante et nous l'aborderons ensuite, avant de parler des problèmes posés par la sémantique de la langue.

Les types d'expressions recherchées

L'index peut contenir des sujets (concepts traités dans le document) ou se présenter comme un index de noms propres (auteurs cités, personnages qui font l'objet d'une discussion, lieux géographiques, produits commerciaux, etc.). La façon de repérer automatiquement les deux types d'expressions n'est pas la même.

Les noms propres sont plus facilement identifiables dans le texte que les sujets, étant donné leur forme caractéristique, soit le fait qu'ils débutent par une majuscule. (Encore faut-il dire que cette facilité est due aux règles d'orthographe et de typographie française – et anglaise – mais ne vaut pas pour toutes les langues. Les noms propres constitués de plusieurs mots, comme « Les Entreprises culinaires Beaulac » ou « Françoise des Rosiers », demandent une stratégie étudiée.) Pour cela, un traitement basé sur des propriétés typographiques et lexicales limitées peut suffire (voir notamment les travaux d'identification de ce qu'on appelle les « entités nommées », dont Plamondon *et al.* 2002 et Mihalcea et Moldovan 2001). Les sujets, eux, sont généralement exprimés par des groupes nominaux. Les identifier implique d'avoir recours à des techniques plus ou moins sophistiquées d'analyse soit linguistique, soit statistique, que nous commentons ci-dessous.

D'autres types d'index peuvent aussi être proposés selon le genre : des listes de langues dans un traité de linguistique, des composés chimiques dans un traité de chimie, etc. Pour ce type d'index, il est préférable d'établir au départ une liste de termes à repérer, éventuellement accompagnée de règles de composition des termes simples (en chimie, par exemple). Des travaux récents en *extraction d'information* peuvent servir à cette fin (voir notamment les actes des colloques MUC-5 et MUC-7).

Voyons maintenant comment procéder à l'identification de sujets.

La méthode d'identification d'expressions intéressantes

Le repérage des expressions intéressantes peut se limiter à identifier toute forme dont la fréquence d'occurrence dépasse un certain seuil (sans se soucier de la forme de l'expression). C'était le cas des approches expérimentales de Earl (1970) et plus récemment de Smajda (1993). Toutefois, on ne peut faire totalement abstraction de la nature des mots de l'expression. En effet, les entrées résultantes seront constituées (presque exclusivement) de groupes nominaux ; il semble plus efficace alors de ne tenter de repérer, au départ, que des groupes nominaux[6]. De plus, il est préférable de normaliser les noms et les adjectifs à leur forme de base, par exemple le masculin singulier en français, pour comptabiliser ensemble toutes les formes morphologiques d'une expression variable.

Une autre approche consiste à fonder l'analyse sur des critères linguistiques (syntaxiques et morphologiques) pour identifier seulement les suites de mots susceptibles de constituer des expressions intéressantes (par exemple, Dillon et Gray 1983, Salton 1988, ou Smart 1992). Avec cette approche, la fréquence d'occurrence peut jouer un rôle secondaire puisque l'on considère que la forme des expressions est davantage garante de leur intérêt pour l'indexation. Mais on se servira quand même de cette fréquence pour faire le tri parmi les multiples expressions repérées.

Le défi ici est de développer des algorithmes robustes et efficaces pour identifier les groupes de mots. Jacquemin (2001) présente une approche intéressante inspirée de nombreux travaux en traitement automatique de la langue. Elle s'attaque à la tâche linguistique d'identification de variantes des termes à repérer.

Enfin, une approche plus directive consiste à compiler à l'avance une liste de termes candidats, avec leurs formes à rechercher dans le texte (y compris leurs synonymes potentiels et autres paraphrases utiles, ce qui dépasse alors l'envergure d'un thésaurus classique). Seules ces expressions seront utilisées pour proposer des termes candidats, et ce, sans égard à la fréquence d'occurrence dans le document. C'est l'approche notamment de Artandi (1963) et de Driscoll *et al.* (1991).

L'utilisation d'un thésaurus est indéniablement utile. Bien sûr, la difficulté réside alors dans la quête du thésaurus approprié à la monographie à l'étude. Les thésaurus spécialisés sont coûteux à développer et ne sont utiles que pour un domaine donné. Les thésaurus généraux ou linguistiques (tels que WordNet, par exemple) peuvent avoir une applicabilité limitée. Mais en l'absence d'un thésaurus spécialisé, ils peuvent représenter les seuls outils disponibles. Par contre, comme nous l'avons mentionné, l'extraction de la terminologie effectivement utilisée dans l'ouvrage (par méthodes statistiques ou linguistiques) respecte davantage l'esprit de l'indexation de livres. Il faudra donc faire montre de discrimination dans le choix et dans l'utilisation automatique d'un thésaurus éventuel.

Parmi les méthodes disponibles jusqu'à présent, l'approche hybride semble la plus prometteuse : il s'agit d'identifier des expressions non pas sur le seul critère de la fréquence, mais en tenant compte de leur forme et de leur construction lexico-syntaxique, tout en se servant de la fréquence dans un deuxième temps pour faire un choix parmi les candidats proposés.

Cette tâche est assez difficile. Notons toutefois que de nombreux travaux en repérage d'information et autres domaines connexes s'attaquent à cette même problématique et proposent des algorithmes de plus en plus efficaces et performants. Ce n'est donc pas là l'aspect le plus difficile du problème.

Les limites de la fréquence

Les approches qui se limitent à retenir les termes les plus fréquents se butent à plusieurs problèmes. D'abord, il n'y a qu'une fourchette de fréquence intéressante. Les mots les plus rares ne sont pas très utiles, puisqu'ils auront été mentionnés une seule fois dans le document, et sont vraisemblablement des accidents de la langue ou des concepts totalement subordonnés dans la discussion. Inversement, les mots les plus fréquents ne représentent souvent aucun intérêt puisqu'ils correspondent à la thématique

6. Cependant, des informations utiles pourraient être extraites de groupes de mots récurrents contenant des verbes, adjectifs ou adverbes, à la condition de les reformuler par la suite.

globale du document et donc apparaissent à peu près dans chaque passage[7].

Que peut-on dire des termes se situant entre ces deux extrêmes ? Parmi les plus fréquents, certains appartiennent véritablement à la langue de spécialité du document en question[8]. D'autres sont à toutes fins pratiques vides de sens : « exemple », « théorie », « structure », « cas », « élaboration », etc. Waller (1999) fait allusion à un niveau de vocabulaire nommé « vocabulaire scientifique de base » contenant de tels termes, peu intéressants comme termes d'indexation puisque leur sens est trop général. Ils devraient donc être mis de côté. Mais ils seront difficiles à distinguer, par la fréquence, d'autres termes potentiellement intéressants appartenant à la terminologie du document. Et, finalement, la fréquence d'occurrence d'une expression dans le document dépend non seulement de l'importance de la discussion sur ce sujet, mais aussi des exigences de la langue et du texte ; dans certains cas, un auteur peut procéder à une énumération qui répète accidentellement certains termes, alors qu'à d'autres moments il utilise des expressions anaphoriques (des pronoms, des démonstratifs comme « celui-ci », « ce dernier », etc.) qui, au contraire, diminuent d'autant la fréquence d'occurrence d'une expression-clé. Enfin, l'observation de Lanteigne (1994a, 43) est éloquente :

> [...] *même lorsqu'un terme d'indexation est sélectionné, il n'est pas dit que toutes les occurrences de celui-ci doivent être indexées. En effet, doit-on indexer les concepts niés, mentionnés en passant, figurant dans les exemples ou dont il est dit qu'on reparlera plus loin ?*

La fréquence d'occurrence est donc une donnée souvent trompeuse.

Le nombre d'entrées d'index générées pour un index microscopique diffère largement, nous l'avons dit, du nombre de descripteurs générés pour une indexation macroscopique. L'utilisation de la fréquence peut être davantage défendue dans ce dernier cas où effectivement on peut imaginer que les cinq ou dix expressions les plus fréquentes ont une corrélation avec les concepts clés du document. Cependant, pour l'indexation microscopique, étant donné le nombre élevé d'entrée désirées, on se trouvera à conserver un grand nombre de candidats à fréquence moyenne pour lesquels la pertinence dans le document ne peut être garantie (à cause à la fois des interférences du vocabulaire scientifique de base et de la difficulté de distinguer l'essentiel de l'accessoire pour une occurrence donnée par la simple fréquence).

Tableau 1

Synonymes	*Reproduction in vitro*, clonage, transgénèse et immortalité... Premiers *clonages* (ou reproduction in vitro) ... La fécondation in vitro (*FIV*) ... ...la *fertilisation in-vitro*... ...la *création in vitro* d'un embryon...
Modifications linguistiques locales	...la fécondation artificielle in vivo et in vitro... ...échecs de fécondation (in vivo ou in vitro) ... ...créer un embryon in vitro... ...embryon a pu être obtenu in vitro... ...l'utilisation d'embryons in vitro... ...la première expérience de clonage et de gestation in vitro d'un embryon humain... De l'embryon in vitro au clonage...
Paraphrase étendue	Malgré ces efforts, des retards de développement apparaissent toujours in vitro ... Personnage aux multiples facettes, l'embryon apparaît tour à tour comme un concept, une entité, une potentialité humaine, une manifestation de la vie, une étape du continuum vital ; « multifonctionnel », in vivo, in vitro, objet de convoitises de la science, mais ayant droit au respect...

Dans tous les cas, d'ailleurs, il sera difficile de déterminer le seuil de fréquence minimal et maximal pertinent pour les termes à retenir[9].

Par ailleurs, la fréquence globale dans le texte est beaucoup moins utile que la fréquence associée aux sous-parties du texte. Ainsi, un terme très fréquent dans un chapitre ou dans une section mais absent du reste du texte devrait engendrer une contribution différente de celle d'un terme assez fréquent tout au long du texte. Or, cette information n'est généralement pas utilisée par les systèmes.

La sémantique de la langue : polysémie, synonymie et paraphrase

Il est bien connu qu'un concept peut apparaître dans un texte sous la forme de différents synonymes ; cette constatation est à la base des systèmes qui utilisent un thésaurus pour améliorer la qualité du repérage des termes. Toutefois, la richesse de paraphrase d'une langue dépasse souvent la simple utilisation de synonymes. Elle est présente dans tout le système langagier : un concept comme « fécondation in vitro » peut être exprimé par un nombre de synonymes mais aussi par différentes paraphrases, tel que l'illustre le tableau 1 (ces expressions ont été retrouvées à l'aide du moteur d'indexation Google).

Devant cet éventail de paraphrases, on conçoit la tâche titanesque de repérer toutes les expressions faisant référence à « fécondation in vitro » – et seulement celles-là.

L'identification des termes importants doit aussi tenir compte de la notion de polysémie : la caractéristique qu'ont les mots comme « voile » de posséder plus d'un sens. Avant de prétendre avoir repéré un mot, et de lui proposer un synonyme, encore faut-il savoir quel sens il a dans le texte. Il serait faux de proposer comme synonyme de « voile » le mot « coiffure » dans le contexte de la phrase suivante : « Le marin légèrement vêtu, avant de hisser les voiles, ajusta ses chaussures et enfonça sa casquette sur sa tête ». Les systèmes automatiques supposent que la redondance de la langue arrive à contourner ces difficultés. Toutefois, le courant de recherche en désambiguïsation lexicale en contexte (*word sense disambiguation* ; voir par exemple Kilgarriff et Palmer 2000) est plutôt issu de la constatation contraire.

7. Par contre, un indexeur juge souvent utile d'incorporer une vedette principale représentant le sujet principal de l'œuvre, ensuite précisée par certaines vedettes secondaires spécifiques qui ne trouvent leur place qu'à cet endroit. Ceci présente de nombreuses difficultés pour un traitement automatique.
8. Peu importe le domaine, même un ouvrage général pour le grand public sera caractérisé par un vocabulaire distinctif.
9. Il est intéressant de noter le comportement du logiciel Sonar Bookends : l'utilisateur fixe le nombre d'occurrences maximales pour un terme destiné à apparaître à l'index. On coupe les termes les plus fréquents, jugés inutiles.

Homogénéisation des entrées

Une fois les entrées candidates repérées, il faut procéder à une certaine homogénéisation ou normalisation des formes. Au strict minimum, les formes pluriel et singulier d'une même expression doivent être ramenées à une formulation commune. Les noms propres doivent être présentés sous une seule variante (normalement le nom de famille et le prénom, au long, inversés, accompagnés d'un titre s'il y a lieu). Si des expressions contenant des verbes ou des adjectifs ont été repérées, elles doivent être nominalisées (ce qui n'est cependant pas une mince tâche et, par conséquent, généralement omis).

Le logiciel peut aussi tenter de faire certains regroupements sur la base de mots (initiaux ou terminaux) contenus dans les expressions ; par exemple, les trois entrées en (5) peuvent être plutôt présentées comme en (6), si le logiciel reconnaît qu'elles débutent toutes par le mot « grammaire ».

(5) *Grammaire de réécriture :* 6, 7.
Grammaire de type 3 : 8.
Grammaire syntagmatique : 10, 20.

(6) *Grammaire*
de réécriture : 6, 7.
de type 3 : 8.
Syntagmatique : 10, 20.

Le logiciel peut aussi générer des entrées inversées à partir des entrées en (6) :

(7) *Type 3*
grammaire : 8.

Ces possibilités n'entraînent pas de grands coûts d'implémentation, mais elles ne sont paradoxalement pas souvent exploitées par les systèmes (exception faite, au minimum, de ce qui est décrit dans Bertrand-Gastaldy 1992, et de ce qui est offert par le logiciel Sonar Bookends Professional).

Repérage des relations sémantiques et regroupement des entrées

Étant donné un ensemble d'expressions extraites et normalisées, il est clair que la simple production d'une liste alphabétique de ces expressions ne peut être considérée comme un index acceptable selon les normes de production d'index de monographies. Idéalement, le logiciel devrait capter les relations sémantiques importantes existant entre les entrées. Ces relations seront manifestées dans la structuration des entrées en vedettes et sous-vedettes ainsi que dans l'utilisation de renvois.

L'utilisation d'un thésaurus peut soutenir la création de renvois de type *voir* entre synonymes. La structuration des entrées alors possible représente une véritable valeur ajoutée pour un système d'indexation si le document est destiné à être accessible sous format numérique. En effet, le logiciel utilisé pour visualiser le document contiendra vraisemblablement une fonction de recherche plein texte. Si celle-ci permet la recherche par troncature, elle représentera un outil de navigation appréciable qui peut remplacer une bonne partie de la fonctionnalité d'un index à la fin d'un livre. Il faut cependant reconnaître que l'index représente un outil utile d'appréhension du contenu d'un texte et qu'un index bien conçu peut apporter un complément très intéressant à la fonction de recherche. Mais la génération de l'index se doit alors de dépasser l'extraction d'expressions apparaissant directement dans le texte. Elle doit se pencher sur le problème de structuration des entrées.

Pour l'insertion de renvois, le système doit repérer des expressions qui deviennent des entrées candidates, identifier les synonymes parmi ces dernières, et unifier les références de chaque occurrence. Par exemple, deux entrées candidates en (8) doivent être identifiées et leurs références de pages réunies ; puis un renvoi établit le lien entre les deux, ce qui est illustré en (9)[10].

(8) *Bicyclette* : 12.
Vélo : 39.

(9) *Bicyclette* : 12, 39.
Vélo voir *Bicyclette.*

Tristement, parmi les relations sémantiques utiles et même cruciales pour l'utilisateur, seules certaines peuvent être détectées automatiquement par les systèmes. L'article de Bertrand-Gastaldy (1992) fait état d'un nombre de relations susceptibles d'être repérées automatiquement par le logiciel SATO. C'est indéniablement un point de départ important de notre réflexion. D'autres relations méritent de s'y ajouter.

En l'absence d'un thésaurus, on peut proposer certains types d'analyses automatiques qui suggèrent des rapprochements sémantiques entre termes. Observons l'entrée en (10) :

(10) *Gestion de projets*
formation : 75-76.
intervenants : 39-50.
méthodologie : 35-38, 54-75.
outils logiciels : 54, 58-72.

Celle-ci proviendrait, on le suppose, de l'analyse de divers passages portant sur la gestion de projets où l'on en mentionne différents aspects. Comment pourrait-on générer ce genre d'entrée automatiquement ? On voit mal comment dériver automatiquement la relation entre « gestion de projets » et « outils logiciels » sur la base de la forme des deux termes. On peut par contre soupçonner entre les deux dans le texte une cooccurrence notable. En règle générale, on s'attend à retrouver plus souvent ensemble des concepts reliés que des concepts non reliés. Un calcul de cooccurrence devrait permettre de distinguer les associations fortuites des associations réelles.

Le logiciel Sonar Bookends prétend produire des entrées structurées à plusieurs niveaux comme celle en (10). Cependant, il faut lui avoir fourni au préalable la structure même de l'entrée. Son travail se limite ensuite à rechercher des pages où à la fois la vedette principale et la sous-vedette sont présentes. Le travail de structuration, de regroupement de concepts reliés, doit être fait par l'humain.

Dans l'objectif de structurer les entrées, ajoutons aussi le problème de différencier de multiples références à la même vedette. La différenciation ne peut être faite *a posteriori.* Ayant identifié

10. On pourrait aussi faire du double postage et proposer les deux entrées « Bicyclette, 12, 39 » et « Vélo, 12, 39 », mais il faudra de toutes façons avoir repéré les synonymes et unifié les références de pages correspondantes.

une entrée avec un grand nombre de références, il n'est généralement pas possible de retourner dans le texte pour que le logiciel établisse dans un deuxième temps les bases de la distinction entre chacune. Celle-ci ne sera possible que si, dès la phase d'analyse et d'identification des termes-clés, le système a été doté de la possibilité de relier les termes correspondant à la vedette principale et ceux liés aux sous-vedettes permettant la différenciation.

Or, repérer les relations implique de conserver tous les termes repérés, à la fois les plus fréquents dans le texte comme « outils » ou « formation » dans l'exemple ci-dessus et d'autres relativement rares, comme peut-être « méthodologie ». Ces mots sont souvent utiles comme sous-vedette pour préciser la portée d'une référence.

La tâche de repérage des relations sémantiques, et du regroupement subséquent, est précisément le défi majeur rencontré par les systèmes automatiques. Elle fait appel à la compréhension de la sémantique des langues et représente un terrain de recherche important ; les solutions proposées devront faire preuve de beaucoup d'ingéniosité.

Édition de l'index

Il y a, en fait, peu à dire sur l'édition de l'index que tout système d'indexation automatique devra néanmoins effectuer. Une chose est digne de mention toutefois : la fluidité des textes disponibles sous format numérique peut rendre problématique la question des localisateurs ou numéros de pages. Notamment, si le texte est destiné à être consulté en ligne dans un format comme HTML où les numéros de pages n'ont aucun sens, il faudra recourir à d'autres types de localisateurs. Une question différente se pose pour les fichiers des logiciels de traitement de texte où les sauts de pages ne sont souvent connus que lors de l'impression puisqu'ils sont largement déterminés par le pilote de l'imprimante utilisée.

Un type de localisateur utile serait de l'ordre de l'hyperlien qui amène l'utilisateur directement dans le texte au passage pertinent, c'est-à-dire à une cible dans le texte. Cela présuppose qu'il est possible pour ces hyperliens de pointer directement vers tout endroit jugé pertinent par l'indexeur. Alors, ou l'indexeur doit avoir accès au fichier afin d'insérer les cibles nécessaires, ou il faut envisager un environnement d'édition et d'affichage qui permette d'insérer des cibles à n'importe quel endroit dans le texte. On peut imaginer par exemple des interfaces où une couche est superposée au texte original afin d'insérer ces cibles.

Comme exemple de limitations que peuvent imposer les systèmes, le logiciel Sonar Bookends indexe des documents déjà paginés et ne permet pas la gestion de numéros de pages en chiffres romains. Son module d'indexation de sites Web fournit comme localisateur uniquement le début du fichier HTML, ce qui est d'une utilité limitée.

Constat actuel

Quel constat peut-on faire alors ? Le manuscrit de Lanteigne (1994b) brosse un tableau éclairé des approches précédentes en indexation automatique microscopique. Par ailleurs, Lanteigne (1994a) émet des mises en garde sur le succès limité que peuvent laisser espérer de tels systèmes, étant donné les difficultés inhérentes à la tâche de compréhension de texte et de synthèse des sujets traités dans un document.

Si l'on en juge par les comptes rendus publiés par divers experts dans le domaine de l'indexation (humaine), les logiciels actuels ne répondent pas aux exigences. En effet, l'identification d'entrées candidates est d'une qualité très variable, le travail de révision est très important et les capacités de regroupement des concepts sont extrêmement limitées (voir notamment Wright 2000).

Bref, les travaux antérieurs en indexation de monographies n'atteignent pas un seuil de performance adéquat pour produire des index utiles. La frontière à franchir, soit la compréhension de la sémantique des textes, est au-delà des capacités des systèmes actuels de traitement de la langue. Que peut-on alors proposer de nouveau ? Avant de tenter de répondre à cette question, nous présentons les hypothèses de travail qui ont prévalu jusqu'à présent et nous suggérons de nouvelles hypothèses qui pourraient mener à des pistes de solutions.

Hypothèses des approches précédentes

De notre examen émane un certain nombre d'hypothèses tacites posées par les approches d'indexation automatique microscopique et qui représentent souvent la source des difficultés.

- Les concepts traités dans un document sont identifiés par les mots du document. En d'autres termes, il suffit d'identifier les mots ou termes du document et éventuellement de leur trouver des formulations synonymiques normalisées pour cerner les concepts traités. Or, comme nous l'avons souligné ci-dessus, les exigences linguistiques et textuelles faussent la donne lorsque l'on procède par comptage d'expressions explicites : un concept important peut n'être qu'évoqué (et non explicité) alors qu'un concept accessoire peut être répété par figure de style[11].
- Une analyse des termes (mots ou expressions) les plus fréquents, globalement, dans le document est utile pour faire ressortir les concepts traités dans celui-ci. En d'autres termes, l'occurrence et la fréquence des mots est cruciale. Par contre, leur localisation dans le document, et dans la structure de celui-ci, est secondaire[12].
- La structuration des entrées d'index est une étape finale, faite en dernier lieu (ou pas du tout puisqu'elle est « trop difficile » pour les systèmes). Cela respecte d'ailleurs la méthodologie enseignée aux apprentis indexeurs.

La taille de l'index dépend entièrement du nombre d'expressions utiles que le système a réussi à identifier. Ce n'est pas une contrainte de départ (d'ailleurs cette contrainte serait inutile). Et incidemment, chaque terme repéré donne lieu à une seule entrée d'index, contrairement à ce que fait un indexeur humain.

11. On pourrait supposer que les deux phénomènes finissent par s'annuler, mais cela dépend sans doute du style de l'auteur.
12. Notons par contre que Baxendale (1957) se sert de la position pour limiter les termes à compter, mais qu'elle ne s'en sert plus par la suite.

Hypothèses à envisager

Nous jugeons que les hypothèses suivantes, qui ne sont habituellement pas posées, sont pertinentes et suggèrent de nouvelles voies.

La structure d'un document peut guider, d'une part, l'identification des concepts-clés du document et, d'autre part, la localisation et la longueur des passages qu'il est utile d'indexer.

La taille désirée pour l'index peut aider à guider le système dans son choix des expressions à retenir ou à éliminer comme entrée d'index.

Toutes les expressions identifiées ne sont pas également utiles pour suggérer des entrées d'index, ou plus précisément, pas utiles de la même façon. Cela ne veut pas dire que l'on a les termes utiles d'un côté, les termes inutiles de l'autre. Les mots très fréquents peuvent certes identifier le sujet principal du document alors qu'un sous-ensemble des mots assez fréquents peut contribuer à structurer les entrées plutôt qu'à les constituer.

Ces nouvelles hypothèses permettent d'aborder autrement l'indexation automatique microscopique. Spécifiquement, il faut explorer la détection de passages et la façon dont l'information est structurée et subordonnée à l'intérieur de chaque passage. Cela permettra de déterminer quelle(s) entrée(s) proposer pour chaque passage, combien d'entrées pour chacun, etc. Les travaux suivants sont pertinents : en *information retrieval,* Hearst (1997) ainsi que Pevzner et Hearst (2002) présentent un algorithme de détection automatique du changement de thème dans un texte, ce qui permet de découper le texte en unités thématiques ; Aït el Mekki *et al.* (2002) proposent, comme nous, de commencer par la structure d'une monographie afin de l'indexer ; et les travaux de Marcu (1999) en condensation automatique portant sur l'identification automatique de la structure rhétorique du document suggèrent des moyens linguistiques (et non plus seulement statistiques) pour dériver la segmentation du texte. Il est intéressant de noter qu'un projet concret comme VIXIT avait pour sa part reconnu la nécessité de segmenter le texte avant de procéder à son indexation automatique (voir Gingras 1992, 15).

Par ailleurs, pour contrer en partie les difficultés liées à l'identification des termes ou des expressions intéressantes, certains travaux récents en traitement automatique de la langue (TAL) peuvent être mis à contribution. Nous en traçons ici les grandes lignes.

Travaux connexes en TAL

L'objectif d'indexation microscopique peut bénéficier de travaux divers en traitement automatique de la langue. Ceux-ci semblent porter sur des problématiques tout à fait différentes, mais utilisent des techniques qui peuvent aider le repérage des termes.

Divers travaux en extraction de terminologie

Les travaux en informatique sur le repérage d'information (*information retrieval*) élaborent des représentations et des algorithmes pour assurer le repérage rapide et efficace de documents (en d'autre termes, des informations à l'intérieur d'un texte ou d'une collection de textes). La finalité est donc la même que celle de l'indexation. Si les index produits pour le repérage d'information diffèrent des index traditionnels des analystes ou des bibliothécaires (essentiellement constitués d'unitermes, ils collent davantage à la terminologie présente dans les textes), on voit une tentative de sophistiquer les techniques pour se rapprocher davantage de ce qu'un index traditionnel permet. Anderson et Pérez-Carballo (2001, 262) soulignent l'état embryonnaire de la recherche portant sur l'utilisation de multitermes en repérage d'information dont certains résultats préliminaires sont décrits dans les actes des conférences TREC (Callan *et al.* 1995, Strzalkowski *et al.* 1997).

Nous relevons ici l'apport notable de l'indexation par sémantique latente (*Latent Semantic Indexing,* Deerwester *et al.* 1990), particulièrement efficace en tant que technique d'indexation macroscopique. Elle capte en quelque sorte les informations sémantiques non explicites dans les textes mais qui émanent de l'association répétée de termes reliés dans les documents. La technique repose sur des calculs mathématiques opérés sur des vecteurs de mots des documents. On dégage, pour la collection de documents, un ensemble de concepts abstraits, dérivés automatiquement et jamais définis explicitement, qui représentent d'une certaine façon la combinaison des relations sémantiques entre les termes des documents.

Cette approche réussit à contourner les problèmes suivants : la synonymie, la polysémie (jusqu'à un certain point), l'aléatoire dans l'expression explicite des concepts clés d'un document, les biais de la fréquence. Il reste à voir comment elle pourrait être intégrée à une approche d'indexation microscopique qui exige non seulement que ces concepts abstraits soient explicités mais également que les concepts soient structurés, et qui considère un document non pas comme un tout mais comme un ensemble de passages.

De plus, les systèmes de traduction automatique les plus performants reposent largement sur de grandes banques terminologiques bilingues (ou multilingues). De nombreux travaux ont porté sur des façons de dériver automatiquement ces banques de termes qui sont très coûteuses à produire de façon manuelle. Dans ce sens, les efforts d'extraction de terminologie peuvent être mis à contribution dans un système d'indexation automatique qui, lui aussi, repose sur une identification efficace des expressions du texte qui représentent des termes d'un domaine de spécialité mais aussi des concepts complexes exposés dans le document.

Condensation automatique

La condensation automatique de documents représente un sujet de recherche très actif depuis la fin des années 1990. Le nombre croissant de documents disponibles en format numérique, et par le fait même des collections de ces documents, fait ressortir le besoin criant, pour un utilisateur, d'avoir des moyens efficaces de prendre rapidement connaissance de cette information.

Ce domaine avait été peu développé jusqu'au début des années 1990. On note néanmoins les travaux initiaux de Baxendale (1958), de Luhn (1957), et les travaux de Rush (1971). On assiste, ces dernières années, à une véritable explosion de projets de recherche et les systèmes commerciaux suivent : la fonction de synthèse automatique du logiciel Word, de Microsoft ; le logiciel grand public Summarizer, de Copernic, etc.

Plusieurs de ces approches procèdent d'abord par l'identification de

termes « intéressants » du document pour ensuite repérer les phrases ou les passages contenant l'essentiel du discours présenté dans le texte. En ce sens, ces systèmes font donc une indexation (automatique) préalable des documents. Leur objectif de condensation oriente leur façon d'identifier les termes intéressants et pose notamment, comme l'on a constaté, qu'il n'est pas nécessairement utile de tenir compte de toutes les sous-parties d'un texte pour le résumer (on peut se concentrer sur l'introduction, sur la conclusion et sur les premières phrases des paragraphes), alors que souvent la recherche de termes intéressants se limite à ces seules parties. Cette approche ouvre une voie intéressante pour l'indexation macroscopique du document puisqu'elle suggère d'aborder l'indexation microscopique en tenant compte de la structure du document.

Catégorisation ou classification automatique

Les travaux de catégorisation et de classification automatiques visent à regrouper automatiquement des documents en classes sur la base de propriétés partagées.

On peut distinguer deux types d'approches selon la nature de la classification résultante : ou on utilise un plan de classification pré-établi dans lequel on tente d'insérer un document ; ou aucun plan de classification n'est présupposé au départ et l'algorithme de classification regroupe les documents entre eux pour générer une nouvelle classification basée sur les similitudes observées entre les documents. Dans les deux cas, les approches se basent sur les mots ou expressions contenus dans le document (éventuellement en faisant intervenir un thésaurus ou une base de connaissances comparable). Ainsi, le repérage de termes est une étape préalable à la classification.

De plus, la façon de procéder peut inclure une étape où l'on assigne à chaque document des étiquettes, semblables à des termes d'indexation, sur la base de termes significatifs du document. Ces approches sont particulièrement pertinentes pour l'indexation automatique puisqu'elles intègrent diverses stratégies pour identifier les termes clés.

Il faut toutefois préciser que ces travaux se rapprochent davantage de l'indexation macroscopique, c'est-à-dire qu'ils visent la synthèse et non l'analyse des sujets du document.

Méthodes statistiques

Les approches innovatrices de recherche en traitement automatique de la langue depuis les 10 ou 15 dernières années sont les méthodes dites statistiques (ou stochastiques, ou probabilistes). Ces dernières font intervenir des calculs de statistiques sur de très grands corpus (plusieurs millions de mots), portant sur une « transformation » donnée appliquée sur des textes. L'exemple pionnier a été celui de la reconnaissance de la parole, suivi d'autres applications en traitement automatique de la langue écrite, notamment la traduction automatique. Dans ce cas, les statistiques de correspondance entre les phrases de textes sources et textes cibles (traduits), calculées sur un grand nombre de paires de textes, servent par la suite à estimer la traduction la plus probable d'un nouveau texte. Cette technique a l'avantage d'être facilement transposable d'un corpus à un autre, d'une langue à une autre, d'un domaine à un autre, puisque les opérations de base du système sont statistiques et mathématiques plutôt que linguistiques. Cependant, le système requiert un corpus d'entraînement de taille très importante afin de maximiser l'estimation des probabilités. De plus, le corpus doit être « de haute qualité », c'est-à-dire habituellement validé par des humains ; dans la tâche de traduction automatique, le corpus doit être constitué de textes traduits par des traducteurs humains.

Ces méthodes ont été appliquées pour accomplir diverses tâches, dont le résumé automatique (Kupiec *et al.* 1995). Dans le domaine de l'indexation, Leung et Kan (1997) rapportent des travaux similaires. Notons qu'il s'agit là d'une tâche d'indexation macroscopique et que cette approche n'a pas été appliquée à l'indexation microscopique. Pour en juger l'applicabilité, il faudrait détenir un corpus important de monographies, chacune accompagnée d'un index de qualité validé.

Pistes à explorer

Alors que les efforts précédents en indexation automatique microscopique se sont penchés en grande partie sur de meilleures analyses du texte afin de mieux identifier les termes candidats à proposer, les travaux futurs se doivent d'incorporer d'autres aspects de la tâche d'indexation négligés jusqu'à présent. Ainsi les travaux devraient opérer sur plusieurs fronts.

Les techniques d'extraction d'information ou d'extraction de terminologie connues doivent être privilégiées pour procéder au repérage des termes candidats. De nombreux travaux sur cette problématique sont effectués dans un grand éventail de domaines, comme nous l'avons évoqué brièvement ci-dessus : traduction automatique, repérage de l'information, extraction d'information, terminotique, catégorisation ou classification automatique de documents, condensation automatique, etc. On relève ici diverses techniques de repérage de termes par patrons linguistiques, syntaxiques, sémantiques, par statistiques, etc. Le recours à un thésaurus (ou à une autre base de connaissances lexicales, sémantiques et encyclopédiques), est souvent le seul moyen de repérer des expressions synonymes qui devraient être indexées par un seul terme commun.

D'autres pistes méritent d'être explorées dans la dérivation des relations sémantiques, notamment l'identification des liens entre deux concepts traités dans le même passage. La relation identifiée pourra servir à structurer une entrée correspondante. Cette relation pourra aussi suggérer différentes entrées d'index utiles pour un seul concept repéré.

Enfin, ce n'est qu'en tenant compte de la structure textuelle et organisationnelle du texte que l'on pourra atteindre l'objectif d'indexer chaque passage pertinent. Il s'agit ici de partir de la structure du texte, et non des mots, pour guider le choix des entrées.

Dès que l'on considère la décomposition en passages comme primordiale, une analogie s'impose : le passage est au document ce qu'un document est à une collection. Il est donc envisageable d'appliquer, pour le repérage d'un passage, les techniques utilisées pour repérer un document dans une collection. Et ainsi, la distinction microscopique/macroscopique,

présente tout au long de ce texte, doit être revue autrement.

Une problématique additionnelle doit en outre s'ajouter aux précédentes. L'élaboration de nouveaux systèmes d'indexation automatique microscopique doit être accompagnée de méthodes d'évaluation objective des résultats obtenus. En effet, il est facile d'extraire des suites de mots d'un texte et même de proposer des façons de les regrouper, mais il est difficile d'avoir des métriques précises de la qualité atteinte ou, chose plus importante, de l'utilité réelle des résultats. Le premier scénario d'évaluation qui vienne à l'esprit est de demander à des experts (des indexeurs, en l'occurrence) de s'exprimer sur la qualité des index produits automatiquement. Cependant, cette approche se bute invariablement à la difficulté que deux experts humains ne s'entendront pas nécessairement sur la façon d'évaluer ni sur l'appréciation qu'ils feront. De surcroît, ils auront de la difficulté à évaluer un produit qui différera sans doute largement de ce qu'ils auraient eux-mêmes produit. Une approche moins subjective (mais pas nécessairement plus simple) est d'évaluer l'efficacité de l'outil (l'index produit) dans l'accomplissement d'une tâche, celle de repérer de l'information dans un document. C'est notamment l'approche de Bennion (1980). Cette technique a l'avantage de faire abstraction des considérations esthétiques de la présentation de l'index pour cibler l'essentiel de sa fonction, soit celle d'aider à repérer un passage intéressant.

Les efforts d'indexation automatique microscopique méritent véritablement d'être poursuivis. Ils peuvent mener, au minimum, à des systèmes d'aide à l'indexation pour l'humain et permettre, chemin faisant, de mieux décrire le processus de production d'un index, ce qui est intéressant d'un point de vue pédagogique.

En outre, de nouveaux contextes d'application se présentent, qui tolèrent une qualité moindre que celle que l'on attend des index préparés par les humains. Nous pensons ici à deux types de contextes : ceux où le volume à traiter est trop important pour jamais être entrepris par des humains et ceux où l'existence éphémère des documents ne justifie pas un travail humain. Dans le premier cas, un outil même approximatif peut aider à naviguer dans des documents volumineux de façon complémentaire à ce que pourrait donner la recherche plein texte. Le second cas est illustré par les réponses des moteurs de recherche à une requête : pour chaque page Web jugée pertinente (mais parfois dépourvue d'outil d'aide au repérage), le système pourrait fournir un index « localisé ». Dans ces cas, la qualité moindre ne remet pas nécessairement en cause l'utilité de l'outil si l'utilisateur comprend ses limites et son fonctionnement.

Sources consultées

Abeillé, Anne. 1993. *Les nouvelles syntaxes.* Paris : Armand Collin.

Aït El Mekki, Touria et Adeline Nazarenko. 2002. L'index, une représentation synthétique de document. In *Atelier « Le résumé de texte automatique : solutions et perspectives »*, Paris, 14 décembre 2002. Disponible en ligne. Page consultée le 25 février 2003. Adresse URL <http://www.atala.org/je/021214/AitElMekki. pdf>.

Anderson, J.D. and J. Pérez-Carballo. 2001. The nature of indexing : How humans and machines analyze messages and texts for retrieval. Part II : Machine Indexing, and the allocation of human versus machine effort. *Information Processing & Management* 3 (2) : 255-277.

Artandi, Susan. 1963. *Book indexing by computer.* New Brunswick, N.J. : S.S. Artandi.

Baxendale, P.B. 1958. Machine-made index for technical literature – an experiment. *IBM Journal for Research and Development* 2 : 354-361.

Bennion, Bruce C. 1980. Performance testing of a book and its index as an information retrieval system. *Journal of the American Society for Information Science* 31 (4) (juillet) : 264-270.

Bertrand-Gastaldy, Suzanne et Gracia Pagola. 1994. *Le contrôle du vocabulaire et l'indexation assistée par ordinateur ; une approche méthodologique et un procédurier pour l'utilisation de SATO.* Montréal : Université de Montréal, École de bibliothéconomie et des sciences de l'information (janvier).

Bertrand-Gastaldy, Suzanne et Gracia Pagola. 1992. L'analyse du contenu textuel en vue de la construction de thésaurus et de l'indexation assistées par ordinateur ; applications possibles avec SATO. *Documentation et bibliothèques* 38 (2) (avril-juin) : 75-89.

Broccoli, Kevin. 1998. Indexes : An Old Tool for a New Medium. *Contentious* 1 (8) (november) Page consultée le 24 février 2003.

Adresse URL : <http://www.contentious.com/articles/1-8/guest1-8c.html>.

Callan, J.P., Croft, W.B. and J. Broglio. TREC and TIPSTER experiments with inquiry, 1997. *Information Processing & Management* 31 (3) : 327-332, 343. Reproduit dans Sparck Jones, Karen, Willet, P. 1997. *Readings in information retrieval.* San Francisco : Morgan Kaufman : 436-439.

Chen, Kuang-hua and Hsin-his Chen. 1994. Extracting Noun Phrases from Large-Scale Texts : A Hybrid Approach and Its Automatic Evaluation. *Proceedings of the 32nd Annual Meeting of ACL*, New Mexico : 234-241.

Collier, R. 1993. Knowledge acquisition from technical texts using natural language processing techniques. In *Proceedings of the 2nd Workshop on the Cognitive Science of Natural Language Processing*, 1-15, Seán Ó Nualláin and Andy Way (Organisers), Dublin City University : Dublin, Eire.

Deerwester, Scott C, Dumais, Susan and George W. Furnas. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science* 41 (6) (September) : 391-407.

Dillon, Martin. 1982. Thesaurus-based automatic book indexing. *Information Processing & Management* 8 (4) 1982 : 167-178.

Dillon, M. and A.S. Gray. 1983. FASIT : a fully automatic syntactically based indexing system. *Journal of the American Society for Information Science* 34 (2) : 99-108.

Driscoll, James R. Rajala, David A. Shaffer, William H. and Donald W. Thomas. 1991. The operation and performance of an artificially intelligent keywording system. *Information Processing and Management* 27 (1) : 43-54.

Earl, L. L. 1970. Experiments in automatic extraction and indexing. *Information Storage and Retrieval* 6 : 313-334.

Gingras, Maurice. 1992. VIXIT : un système d'analyse et de repérage de l'information textuelle pour la gestion des ressources humaines. *Documentation et bibliothèques* 38 (2) (avril-juin) : 115-116.

Hearst, Marti. 1997. TextTiling : Segmenting Text into Multi-Paragraph Subtopic Passages. *Computational Linguistics* 23 (1) : 33-64.

Hodge, Gail M. and Jessica L. Milstead, 1998. *Computer support to Indexing.* Philadelphia, PA : National Federation of Abstracting and Information Services.

Jacquemin, Christian. 2001. *Spotting and Discovering Terms through Natural Language Processing.* Cambridge, Mass : MIT Press.

Jacquemin, C., Daille, B., Polanco, X. and J. Royauté. 2002. In vitro evaluation of a program for machine-aided indexing. *Information Processing & Management* 38 (1) : 765-792.

Kilgarriff, Adam and Martha Palmer (reds). 2000. *Computers and the Humanities, Special Issue.* SENSEVAL : Evaluating word sense disambiguation program 34 (1-2) (mai).

Klement, Susan. 2002. Open system versus closed system indexing. *The Indexer* 23 (1) (April) : 23-31.

Korycinski, D. and A. F. Newell. 1990. Natural-language processing and automatic indexing. *The Indexer* 17 : 21-29.

Lanteigne, Diane. 1994a. Prolégomènes au développement d'un système d'aide à l'indexation de monographies. *ICO Québec* (printemps).

Lanteigne, Diane. 1994b. *Amorce de développement d'un système d'aide à l'indexation de livres de langue française avec SATO : formalisation de la chaîne des traitements et réalisation d'un index de livre.* Université de Montréal, École de bibliothéconomie et des sciences de l'information, Faculté des arts et des sciences.

Leung, C.H. and W. K. Kan. 1997. A statistical learning approach to automatic indexing of controlled index terms. *Journal of the American Society for Information Science* 48 (1) (January) : 55-66.

Luhn, H.P. 1957. A statistical approach to mechanized encoding and searching of literary information. *IBM Journal of Research and Development* 1 : 309-317.

Marcu, Daniel. 1999. Discourse Trees are Good Indicators of Importance in Text. In Mani, Inderjeet and Mark Maybury. *Advances in Automatic Text Summarization.* Cambridge, Mass. : MIT Press : 123-136.

Mihalcea, R. And D.I. Moldovan. 2001. Document indexing using named entities. *Studies in Informatics and Control* 10 (1) (March) : 21-27.

MUC-5. 1993. *Proceedings of the Fifth Message Understanding Conference (MUC-5).* San Francisco, CA : Morgan Kaufmann Publishers.

MUC-7. 2003. *Proceedings of the Seventh Message Understanding Conference (MUC-7).* Disponible en ligne. Page consultée le 27 février 2003. Adresse URL : <http://www.itl.nist.gov/iaui/894.02/related_projects/muc/proceedings/muc_7_toc.html>.

Mulvany, Nancy. 1999. Software tools for indexing : revisited. *The Indexer* 21 (4) (October) : 160-163.

Pevzner, L. and M. Hearst. 2002. A Critique and Improvement of an Evaluation Metric for Text Segmentation. *Computational Linguistics* 28 (1) (March) : 19-36.

Plamondon, L., Lapalme G. and L. Kosseim. 2002. The QUANTUM question answering system. In Voorhees, E.M. and D.K. Harman. *Information Technology : Tenth Text Retrieval Conference, TREC 2001.* Gaithersburg, MD, USA : NIST : 579-585.

Royauté J., Schmitt, L. and E. Olivetant. 1992. Les expériences d'indexation à l'INIST. In *Proceedings of the 15th International Conference on Computational Linguistics (COLING'92)* (23-28 août), Nantes.

Rush, J.E., Salvador, R. and A. Zamora. 1971. Automatic abstracting and indexing. II. Production of indicative abstracts by application of contextual inference and syntactic coherence criteria. *Journal of the American Society for Information Science* 22 (3) : 260-274.

Salton, Gerard. 1988. Syntactic Approaches to Automatic Book Indexing. In *Proceedings of the 26th Annual Meeting of the Association for Computational Linguistics.* (7-10 June) State University of New York at Buffalo, Buffalo, New York. Morristown, N.J. : Association for Computational Linguistics : 204-210.

Smajda, F. Retrieving collocations from text : Xtract. *Computational Linguistics* 19 (1) : 143-177.

Smart, Godfrey. *SIMPR : Structured Information management : Processing and Retrieval.* Esprit Project 2083. The SIMPR project : the results. SIMPR document n° SIMPR-CRI-1992-41.10e.

Strzalkowski, T., Lin, F. and J. Pérez-Carballo. 1997. Natural language information retrieval : TREC-6 report. In Voorhees, E. M. and D. Harman. *Information technology : the sixth Text Retrieval Conference* (TREC-6). Gaithersburg, MD : US Department of Commerce, Technology Administration, National Institute of Standards and Technology : 209-228.

Turney, Peter D. 2000. Learning algorithms for keyphrase extraction. *Information Retrieval* 2 (4) : 303-336.

Waller, Suzanne. 1999. *L'analyse documentaire. Une approche méthodologique.* Paris : ADBS Éditions.

Wellisch, Hans H. 1994. Book and Periodical Indexing. *Journal of the American Society for Information Science* 45 (8) : 620-627.

Wright, Jan C. 2000. Sonar Bookends automatic index generator : a review. *Key Words* 8 (5) (September-October) : 153, 179-85.