Outil de butinage du contenu des documents de collections numériques

Lyne DA SYLVA

École de bibliothéconomie et des sciences de l'information, Université de Montréal

Mots-clés : indexation, collections numériques, index de livre, indexation automatique, accès à l'information, accès au contenu, aide à la lecture

Keywords: indexing, digital collections, back-of-the-book index, automatic indexing, access to information, access to contents, reading aid

Résumé : Cette recherche se veut une contribution à la recherche d'information dans les documents numériques, non pas pour le repérage de documents mais pour l'aide à la lecture et donc l'évaluation de la pertinence de documents repérés. L'introduction d'un outil de butinage est proposée pour accéder au contenu de documents des bibliothèques numériques, soit l'index de livre traditionnel. Celui-ci présente plusieurs avantages en tant qu'outil de navigation, bien que sa création automatique pose quelques difficultés. L'implémentation d'un outil de ce type est esquissée dans ses grandes lignes.

Abstract : Our research is a contribution to information search within digital documents, after the initial steps of document retrieval from a given digital library. We suggest introducing a type of browsing tool to aid in document perusal and thus to help in evaluating its relevance for the user's information need. The tool in question is the traditional back-of-the-book style index. We present its advantages as a browsing tool, some challenges posed by the automatic creation of this type of tool, and a sketch of our current implementation.

1 Introduction

Cette étude porte sur les collections numériques (ou bibliothèques numériques) de textes non structurés et sur les outils pour accéder à leur contenu. Nous proposons l'adjonction d'un certain outil de navigation dans les documents. Des outils puissants de description sont nécessaires pour permettre aux utilisateurs de repérer les documents pertinents à leurs besoins. Les outils de ce type s'arrêtent cependant à la tâche d'extraire un certain nombre de documents de la collection, n'aidant pas ou peu

l'utilisateur à prendre connaissance du contenu de ceux-ci afin d'évaluer leur pertinence réelle. Nous proposons d'ajouter un outil d'aide à la lecture du document afin de faciliter cette tâche de prise de connaissance du contenu.
La section 1 identifie les outils de recherche actuels et leurs lacunes pour accéder au contenu des documents. À la section 2, nous présentons un nouveau type d'outil, l'index de fin de livre, qui est bien connu pour faire des recherches dans des documents imprimés, mais pratiquement inutilisé pour les documents numériques. La section 3 esquisse une implémentation d'un tel outil et la section 4 discute des difficultés rencontrées lors de l'implémentation de cette approche, alors que la conclusion aborde des pistes de recherche futures.

2 Travaux précédents

Les outils de recherche, qui permettent de repérer des documents dans une collection, se déclinent en plusieurs variétés : moteurs de recherche généraux (Goole, Ask.com ou Yahoo!, etc.), moteurs spécialisés (Yahoo! Kids, Google Scholar, etc.) selon les utilisateurs, le domaine, le type de documents ou la région géolinguistique visés, méta-moteurs (Excite, Hotbot, Metacrawler, etc.) et autres. Leur fonction première est d'aider les utilisateurs à trouver un document qui peut répondre à leurs besoins d'information ; il reste à cet utilisateur à consulter le document en question (soit le lire, totalement ou partiellement), pour déterminer si son besoin d'information est satisfait.
Certains outils peuvent servir également à prendre connaissance (bien que sommairement) du contenu des documents. À cet effet, dans plusieurs cas, la liste des documents retournés en réponse à la requête contient un (très) court extrait de chaque document. De plus, ces outils reposent sur l'indexation préalable des documents, qui peut être faite en vocabulaire libre ou avec des vocabulaires contrôlés ; par exemple (Baca, 2003) : *Library of Congress Subject Headings*, *Art and Architecture Thesaurus*, *Thesaurus of Geographic Names*, *Library of Congress Thesaurus for Graphic Materials*. L'indexation fournit ainsi les métadonnées qui servent à décrire chaque document et à les apparier aux mots de la requête. Ces métadonnées peuvent également aider davantage l'utilisateur : elles peuvent aussi nourrir un système de visualisation de l'information, pour regrouper des documents semblables par exemple (comme dans les outils Metacrawler, Clusty, Grokker, etc.). Ceci peut aider l'utilisateur à se faire une idée de leur contenu. Certains systèmes de repérage (Davis, 2006) permettent à la fois la navigation dans une structure préétablie et une recherche par mots-clés ; la combinaison peut aider à mieux cerner la pertinence des documents repérés. Mais

l'utilisateur a peu de moyens, outre la lecture complète ou partielle du document, pour prendre connaissance de son contenu et évaluer la pertinence pour ses besoins. Un résumé peut alléger cette tâche, mais les résumés se font plutôt rares.
Certains chercheurs proposent des interfaces de navigation basées sur une analyse du contenu des documents (par exemple, Dakka et al., 2005). Elles servent alors à naviguer dans une collection de document, et non à l'intérieur d'un document donné. Quelques chercheurs (dont Hernandez et Grau, 2003) proposent un outil qui peut générer, pour un document, une structure semblable à une table des matières. Yaari et Gan (2000) font ceci à partir d'une analyse hiérarchique des sujets abordés et leur système construit aussi un index thématique, qui consiste essentiellement d'une liste de termes extraits d'une section donnée.
Nous cherchons à contribuer aux efforts de déploiement d'outils d'accès au contenu de documents numériques, qui permettraient aux utilisateurs de naviguer effectivement dans le document par le biais de son réseau conceptuel, et non de son organisation textuelle.

3 Un outil à considérer : l'index de livre

Nous voulons explorer un moyen, autre que le résumé ou la table des matières, qui aiderait l'utilisateur à prendre rapidement connaissance du contenu d'un document, en quelque sorte une aide à la lecture. Un outil à considérer pour l'accès au contenu des documents serait l'index de livre.

3.1 Structure

Un index de livre se présente comme une liste alphabétique d'entrées, chacune structurée en vedette principale et éventuellement de sous-vedettes, menant à une référence de page, par exemple :

> Température, 186-189 (Fenwick, 1997)
> du bain, 138, 141, 227
> de la chambre, 118, 121, 178
> fièvre, 180, 184, 186-188, 187
> pendant la grossesse, 38
> prise de la, 187
> urgence, 38, 174
> *voir aussi* Thermomètre

Chaque entrée représente un thème abordé dans le document ; les sous-vedettes le subdivisent en aspects secondaires, termes spécifiques, etc. Certaines entrées sont simples, constituées uniquement d'une vedette principale. La taille de l'index détermine sa couverture thématique par rapport au contenu global du document. Des renvois de type *voir aussi*

entre les entrées permettent d'établir des liens qui auraient pu échapper à l'utilisateur alors que les renvois de type *voir* (non illustré ici) mènent à des vedettes synonymes. Ce type d'outil est très familier aux utilisateurs de documents papier et il possède des caractéristiques différentes de celles offertes par les autres outils d'accès.

3.2 Comparaison avec autres outils d'accès au contenu

La structure de l'index est différente de celle de la table des matières : cette dernière reflète l'organisation textuelle alors que l'index est plutôt un inventaire de thèmes abordés. La table des matières aborde le contenu d'un document de manière séquentielle ; l'index permet d'y accéder de manière tabulaire (une longue réflexion sur des sujets reliés est présentée dans Vandendorpe, 1999).

L'index diffère d'un résumé en ce qu'il couvre davantage de thèmes, dans plus de détails, regroupant les mentions de ceux-ci qui seraient dispersées dans le document. Et bien sûr, le résumé est généralement un texte suivi (ou une grille textuelle), qui suit dans les grandes lignes l'organisation textuelle, alors que l'index est composé de courts termes juxtaposés et ordonnés (par ordre alphabétique ou un autre ordre systématique).

Par rapport à une fonction de recherche en texte intégral, l'index offre l'avantage de présenter ouvertement à l'utilisateur les thèmes du document ainsi que certaines des relations qui les unissent ; cela peut aider à mieux formuler une requête ultérieure. Une étude de Abdullah et Gibb (2009) a comparé trois types d'outils de navigation pour les livres électroniques (*e-books*) : index de livres, table des matières et fonction de recherche. L'index s'est révélé plus efficient que les autres en termes de rapidité pour trouver l'information, plus performant pour repérer correctement du contenu pertinent et plus convivial pour les utilisateurs.

Comme outil de navigation, il offre l'avantage de délimiter la couverture conceptuelle du document, proposant des termes pour une requête éventuelle. Il inclut des variantes des termes, pour repérer des thèmes peu importe la terminologie choisie. Et il peut représenter un outil additionnel, la recherche en texte intégral étant aussi souvent disponible. Enfin, Wathen and Burkell (2002) rapportent que les utilisateurs recherchent la familiarité de l'imprimé dans les environnements Web. L'inclusion d'un index de ce type, très connu des utilisateurs, présenterait donc de nombreux avantages.

La fonctionnalité de butinage qu'il incorpore est intéressante. Brown (1988) relève la distinction entre le butinage (ou navigation) et la recherche en termes de l'opposition entre ce que l'on recherche et l'endroit où il se trouve : pour le butinage, l'utilisateur procède de l'endroit vers l'information recherchée (*from where to what*) alors que pour la recherche, le mouvement est inverse, de l'information recherchée

à l'endroit où elle se trouve (*from what to where*). Pour la recherche, il faut donc savoir au départ ce que l'on veut trouver. Alors que la navigation permet une appropriation graduelle d'un contenu même si l'utilisateur n'a aucune connaissance préalable de celui-ci. Ertzscheid (2003) étudie les comportements différents qui caractérisent les deux activités liées au repérage d'information.

4 Construction automatique d'un index de livres

La construction d'un tel outil de manière automatique présente plusieurs obstacles. Les premiers prototypes (Artandi, 1963 ; Earl, 1970) n'ont guère eu de succès, étant limités par leur méthodologie qui consistait essentiellement à lister alphabétiquement les mots les plus fréquents du document. Or, une implémentation réussie doit contourner deux pièges : d'abord, éviter de confondre fréquence avec importance, ce qui implique de limiter les entrées d'index aux occurrences significatives d'un sujet ; ensuite, aborder le problème de la structuration des entrées, généralement faite (par les humains) sur la base de relations sémantiques, très difficiles à capter automatiquement.
Nous avons toutefois proposé une méthode inspirée de la méthodologie des indexeurs humains et qui permet d'atteindre de meilleurs résultats. Elle repose sur trois principes : (i) les indexeurs indexent des passages, et non des mots, et cette séparation en passages dicte le repérage des thèmes importants à l'intérieur de celui-ci ; (ii) à fréquence égale, les mots n'ont pas tous la même utilité dans l'index ; (iii) certains relations sémantiques entre les thématiques sont explicitées par les mots du passage. Les détails de l'implémentation sont présentés dans Da Sylva et Doll (2005) et sont esquissés ci-dessous. Mais il est utile d'approfondir d'abord la dernière idée, portant sur les relations sémantiques évoquées dans les entrées structurées de l'index.

4.1 Entrées complexes et relations utiles

Un index vise à aider l'utilisateur à repérer des passages utiles. Pour ce faire, l'index présente (en vedette principale) les concepts-clés du document. Lorsque plusieurs passages font référence à un même concept-clé, l'entrée énumère tous ces endroits en autant de numéros de page. Lorsque cette liste devient trop longue, elle est inutile à l'utilisateur, qui est confronté à un trop grand nombre de passages potentiellement intéressant pour sa recherche d'information. Il est préférable, alors, de distinguer chacune des références en introduisant des sous-vedettes qui explicitent l'aspect selon lequel le concept est envisagé dans chacun des passages. Cette présentation est souhaitable aussi bien dans un index créé manuellement que dans un autre construit automatiquement.

Il est alors important de déterminer quelle sous-vedette devrait être utilisée pour distinguer chaque référence. D'une part, pour simplifier la tâche du système, on peut supposer que la sous-vedette est présente explicitement dans le texte source (la dériver automatiquement exige des ressources sémantiques considérables) ; on voudra donc, dans l'implémentation, extraire au besoin des paires de termes, dont l'un sera finalement la vedette principale et l'autre sera la sous-vedette. D'autre part, on veut limiter le type de sous-vedettes utilisées (pour limiter le nombre de paires extraites), ce qui peut être fait en définissant les types de relations utiles dans un index.
Dans un travail précédent, nous avons analysé le type de relations entre la vedette principale et les sous-vedettes dans un certain nombre d'index créés par des indexeurs humains. Elles sont présentées dans Da Sylva (2004). Elles incluent entre autres la relation hyperonymique (voir la figure 1).

Type	Relation	Exemple
Syntagmatique	Mot – Terme avec ce mot	Grammaire - grammaire de dépendance
	Coordination	Café - et grossesse
Paradigmatique	Hyperonyme – Hyponyme	Mammifères – félins
	Tout – partie	Voiture – moteur
	Thème – Facette ou aspect	Robotique – développement

Figure 1. Relations principales observées dans les index de livres

4.2 Implémentation

Notre prototype d'indexation (Da Sylva et Doll, 2005) fonctionne de la manière suivante :

1. Segmentation du texte en segments thématiques. La méthodologie utilisée s'inspire de l'approche de Hearst (1997) et repose sur l'analyse de la cohésion lexicale : une coupure thématique est postulée entre deux segments quand le score calculé à partir d'indicateurs lexicaux (mots répétés, absence d'anaphores, etc.) chute. Nous avons modifié l'algorithme pour assurer la relative uniformité des segments. Cette segmentation sert à définir les passages auxquels les entrées d'index font référence.

2. Extraction des mots (et des suites de plusieurs mots, appelés multitermes) après lemmatisation et comptage des fréquences. Sur la base de la fréquence des mots (à l'intérieur des segments comme dans le document dans son ensemble), on déterminera la saillance d'un sujet dans un segment donné.

3. Identification, dans le texte, de paires de mots ou de multitermes qui pourront former des couples vedette principale/sous-vedette. Ces paires doivent relever de types précis, identifiés dans l'étude préalable (Da Sylva, 2004). Cette méthode permet de produire des entrées structurées comme celle donnée en exemple au début de cet article.
4. Pour chacun des éléments de la liste de candidats-termes (que sont les mots, termes et paires identifiés dans les étapes 2. et 3.), calcul d'un poids; pour chaque segment, on ne retiendra dans l'index que les candidats-termes dont les poids sont les plus élevés (au-delà d'un certain seuil).
5. Sélection des candidats-termes les plus saillants, regroupement sur la base des vedettes principales partagées et mise en ordre alphabétique.

Charlet et al. (2004) et Nazarenko et Aït El-Mekki (2005) présentent un outil très similaire à celui que nous avons développé de manière indépendante. L'introduction, dans le processus de construction de l'index, d'un auteur humain leur permet de contourner plusieurs problèmes liés à la limite de l'analyse automatique de la langue.
Notre originalité tient au traitement que nous accordons aux différents types de liens sémantiques qui peuvent tenir entre une vedette principale et une vedette secondaire. La figure 2 présente des exemples d'entrées d'index produites par notre prototype. Certaines entrées sont des mots simples, d'autres des multitermes, d'autres encore des paires de termes. Les numéros font référence aux segments obtenus par la segmentation automatique. Même dans ce court extrait, on voit que les concepts sont reliés dans l'index même quand ils sont disséminés dans le document.

béton béton armé, 5 limite, 5 dalles de béton ordinaire coulées, 4 renforcement du béton avec des fibres d'acier, 5 béton armé, 5 utilisation du béton précontraint, 4 béton ordinaire, 4	cheveu, 10 fibre, 9 fibres de noix de coco, 9 fines fibres, 3 renforcement du béton avec des fibres d'acier, 5 béton armé, 5

Notre prototype n'a pas encore fait l'objet d'une évaluation objective, sauf l'aspect segmentation de texte (Da Sylva, 2006), qui se compare favorablement à l'approche de Hearst (1997).

5 Quelques défis

On peut objecter qu'un index créé automatiquement doit offrir plus que simplement l'extraction des mots et expressions dans le texte, sinon il est d'une utilité limitée. Cependant, deux propriétés d'un index, même produit par pure extraction de termes du document, en justifie la création. D'abord, il offre un inventaire des concepts présents dans le document, explicitant du fait même la couverture conceptuelle aussi bien que lexicale ; c'est en quelque sorte une « photographie conceptuelle » de celui-ci. Et il indique également des relations entre les concepts (exprimées dans les entrées structurées en vedette principale et sous-vedettes). Ensuite, il restreint l'apparition des expressions à celles qui sont le plus importantes, alors qu'une fonction de recherche repérera chacune des occurrences.

Mais il est clair qu'il est préférable d'inclure dans l'index des expressions que l'on ne peut pas trouver directement dans le texte : par exemple, des synonymes ou des hyperonymes de termes du document. Si le document parle de « vélo », on voudrait trouver à l'index un renvoi « bicyclette, voir vélo » même si ce deuxième terme n'apparaît pas dans le texte. Également, un ouvrage qui parlerait de différents types de rongeurs, mais toujours dénotés par leur race spécifique (« souris », « rat », « écureuil », etc.), gagnerait à avoir une entrée « rongeurs » qui regrouperait chaque type. La difficulté réside alors à trouver des ressources lexicales externes au document qui contiennent ces informations. Nazarenko et Aït El-Mekki (2005) bénéficient d'une bonne solution à ce problème, ayant accès à une large base lexicale qui contient, pour chaque terme, des variantes aussi bien que des hyperonymes ou hyponymes. Un thésaurus général (disponible en format numérique) comme WordNet peut souvent fournir l'information nécessaire. En l'absence de ceci (par exemple, pour des langues pour lesquelles ces ressources n'existent pas), on doit imaginer d'autres stratégies. Comme par exemple l'analyse de grands corpus afin d'en extraire des généralisations pertinentes, parmi lesquelles pourront se trouver les relations qui nous intéresse (comme dans Ruiz-Casado, 2007, ou Hearst, 1992, par exemple). En plus, un thésaurus thématique serait avantageux : il contiendrait des connaissances disciplinaires spécialisées qui échappent aux thésaurus généraux.

On peut préférer une fonction de recherche en ce qu'elle nous amène directement à l'endroit dans le texte où l'objet de notre recherche apparaît. En contraste, l'entrée d'index nous amène normalement à une région textuelle (un passage, un paragraphe, une page) où il est du ressort de l'utilisateur de localiser l'endroit pertinent.

En outre, dans l'extraction des mots et termes du document, l'identification de ceux-ci se fait normalement sur la base de la chaîne

alphabétique, et non sur la base du sens du mot. Les ambiguïtés dues à la polysémie et à l'homographie amenuisent la performance du système.
En somme, cette tâche rencontre beaucoup des problèmes déjà identifiés dans d'autres applications de traitement automatique de la langue.

6 Conclusion

Pour faciliter l'accès au contenu de documents numériques, nous proposons un outil ancien, bien connu des utilisateurs et des indexeurs, l'index de livres. Cet outil serait particulièrement utile dans le cas de monographies assez importantes où la structure est peu apparente. Il offre un accès différent au texte, complémentaire à un résumé, à une fonction de recherche ou à une table des matières.
Nous avons proposé une implémentation qui tient compte, dans certains de ses aspects du moins, de la méthodologie des indexeurs humains et des propriétés attendues d'un index de qualité. Bien qu'une évaluation objective reste à faire, l'approche générale nous semble suffisamment motivée pour constituer un chantier de recherche intéressant.
Plusieurs pistes de recherche restent à explorer. Parmi celles-ci : l'ajout de ressources lexicales externes (comme un thésaurus général de la langue ou un thésaurus particulier au domaine de la monographie) ; l'évaluation par des utilisateurs et par des indexeurs professionnels ; la construction de différentes présentations de la structure conceptuelle définie par l'index, y compris des représentations en ontologies ou Topic Maps ; et le raffinement des types d'expressions qui constituent les entrées proposées pour l'index.

7 Références bibliographiques

[1] N. Abdullah et F. Gibb. Students Attitudes towards e-Books in a Scottish Higher Education Institute: Part 3 -- Search and Browse Tasks. *Library Review*, 58(1), 2009, 17-27.

[2] S. Artandi. *Book indexing by computer*. S.S. Artandi, New Brunswick, N.J. 1963.

[3] M Baca. Practical issues in applying metadata schemas and controlled vocabularies to cultural heritage information. *Cataloging & Classification Quarterly*, 36(3/4), 2003, 47-55.

[4] P. J. Brown. Linking and searching within hypertext. *Electronic Publishing*, 1(1), 1988, 45-53.

[5] J. Charlet, T. Aït el Mekki, D. Bourigault, A. Nazarenko, R. Teulier et B. Toledano. CEDERILIC : constitution d'un livre et d'un index

numériques. In : *Actes du Colloque International sur le Document Electronique (CIDE)*, 2004.

[6] W. Dakka, G. P. G. Ipeirotis et K.R. Wood. Automatic construction of multifaceted browsing interfaces. In *CIKM*, 2005, 768-775.

[7] L. Da Sylva. *Experiments in Proportional and Variable Automatic Text Segmentation* (poster). 19th Conference of the Canadian Society for Computational Studies of Intelligence (AI'06). 2006, Université Laval, Québec.

[8] L. Da Sylva et F. Doll. A Document Browsing Tool: Using Lexical Classes to Convey Information. In G. Lapalme et B. Kégl. *Advances in Artificial Intelligence: 18th Conference of the Canadian Society for Computational Studies of Intelligence, Canadian AI 2005 (Proceedings)*, New York : Springer-Verlag, 2005, 307-318.

[9] L. Da Sylva. Relations sémantiques pour l'indexation automatique. Définition d'objectifs pour la détection automatique. *Document numérique,* 8, 3 (2004), 135-155.

[10] L. Davis. Designing a search user interface for a digital library. *Journal of the American Society for Information Science and Technology*, 57(6), 2006, 788-791.

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

[12] O. Ertzscheid. Comportements de navigation et documents électroniques : propositions d'invariants. In : C. Faure, J. Madelaine (réds), *Document électronique Dynamique. Actes du sixième colloque international sur le document électronique : CIDE.6*, Europia Productions, Paris, 2003.

[13] E. Fenwick. *Mon bébé, je l'attends, je l'élève* (traduction de *The Canadian Medical Association complete book of mother & baby care*). Reader's Digest Association, Montréal. 1992.

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

[15] M. Hearst. Automatic Acquisition of Hyponyms from Large Text Corpora. In *Proceedings of the 14th International Conference on Computational Linguistics*, 1992, 539-545.

[16] N. Hernandez et B. Grau. What is this text about? Combining topic and meta descriptors for text structure presentation. In *Proceedings of the 21st annual international conference on Documentation (ACM SIGDOC)*, San Francisco, 12-15 Oct. 2003, 117-24.

[17] A. Nazarenko et T. Aït El Mekki. Building back-of-the-book indexes. *Terminology*, Special issue on Application-driven Terminology engineering, 11(11), 2005, 199-224.

[18] N. Hernandez et B. Grau. What is this text about? Combining topic and meta descriptors for text structure presentation. In: *Proceedings of the 21st annual international conference on Documentation (ACM SIGDOC*), San Francisco, 12-15 Oct. (2003), 117-124.

[19] M. Ruiz-Casado, E. Alfonseca et P. Castells. Automatising the learning of lexical patterns: An application to the enrichment of WordNet by extracting semantic relationships from Wikipedia. *Data & Knowledge Engineering*, 61(3), 484-99, 2007.

[20] Vandendorpe, C. Du papyrus à l'hypertexte: essai sur les mutations du texte et de la lecture, Boréal, Montréal, 1999.

[21] Y. Yaari et R. Gan. *NLP-assisted exploration of texts*. In In Proceedings RIAO'2000 Content-Based Multimedia Information Access, Paris, 2000, 2000.