Notes de lecture

Rubrique préparée par Denis Maurel

Université François Rabelais Tours, LI (Laboratoire d'informatique)

Mohand BOUGHANEM, Jacques SAVOY, Recherche d'information : état des lieux et perspectives, *Hermès-Lavoisier*, 2008, 343 pages, ISBN 978-2-7462-2005-8.

Lu par **Aurélie NÉVÉOL**

National Library of Medicine, Bethesda, USA

Cet ouvrage collectif présente la problématique de la recherche d'information depuis l'émergence de la discipline dans les années soixante jusqu'à l'avènement récent du Web qui a introduit de nouveaux défis pour la communauté, caractérisés par le volume de l'information à traiter, ainsi que sa diversité tant au niveau des formats que des supports ou des langues dans lesquels elle se présente. Ainsi, après un premier chapitre introductif, l'ouvrage propose un traitement inégal de thématiques liées à l'émergence du Web ainsi que des problématiques majeures que sont la prise en compte du multilinguisme, le traitement de documents multimédias, la prise en compte du contexte ou les systèmes de questions-réponses.

Recherche d'information fait appel à dix-sept auteurs pour présenter les travaux depuis l'émergence de la discipline jusqu'à aujourd'hui. Le premier chapitre introduit la problématique de la recherche d'information (RI), à savoir l'appariement entre un besoin d'information exprimé, par exemple, sous forme de requête, et l'information cherchée qui se trouve dans les documents d'une collection. Les modèles vectoriel et probabiliste sont présentés, ainsi que les mesures de performances en RI, consacrées par les campagnes d'évaluation TREC.

Le deuxième chapitre évoque plus particulièrement la RI dans le contexte du Web comme « collection » de documents et l'introduction des liens entre documents dans les algorithmes de RI. Les diverses méthodes de classement des documents fondées sur l'usage des hyperliens sont clairement exposées et illustrées. La fin du chapitre, un peu moins bien structurée, commente les comportements de recherche des internautes de différents pays qui utilisent des moteurs de recherche mettant en œuvre les algorithmes décrits.

Le troisième chapitre est une traduction en français d'un article paru dans les actes de SIGIR en 2002. Il décrit une approche statistique de la détection de *pages d'entrée*, ou *pages d'accueil*. Ce travail part du constat que contrairement aux algorithmes de RI classique qui doivent optimiser le rappel (exhaustivité), la recherche de pages d'accueil doit optimiser la précision puisqu'il n'y a que peu de

pages d'accueil à retourner, voire une seule dans beaucoup de cas. Les auteurs explorent trois critères (longueur du document, nombre de liens entrant, profondeur de l'URL) permettant d'estimer la probabilité pour des pages retournées par un système de RI classique, d'être des pages d'accueil. Une évaluation utilisant les corpus des campagnes TREC consacrées à la détection des pages d'accueil est présentée.

Le chapitre 4 commence par un bref historique des langages de balisage qui ont permis de structurer le contenu des documents au sein d'une collection comme le Web, et de donner accès à des entités plus petites que le document. La notion de granularité de l'information est introduite. L'auteur décrit ensuite les contributions respectives des communautés RI et BD (base de données) dans l'exploitation de documents structurés ou semi-structurés. Le milieu du chapitre, et en particulier la partie sur les langages de requêtes, aurait pu être illustré par des exemples concrets à l'intention du lecteur non spécialiste.

Le chapitre 5 est consacré au traitement de documents dans d'autres langues que l'anglais dont la prépondérance sur le Web est en diminution notable. La RI dans les multiples langues naturelles rencontre les obstacles, désormais classiques en TAL, liés aux alphabets comportant des caractères non ASCII, à l'inégalité des ressources linguistiques disponibles, ainsi qu'aux difficultés de segmentation des textes en unités linguistiques significatives particulièrement accrues pour les langues orientales. Assez dense, ce chapitre aborde de nombreux points parfois sans les approfondir ni les rendre accessibles pour le lecteur non spécialiste.

Le chapitre 6 présente les systèmes de questions-réponses (QR), qui permettent d'extraire en trois étapes une information précise des documents d'une collection : l'analyse de la question, la recherche de passages pertinents et l'extraction de la réponse. L'auteur expose ensuite les particularités du Web qui peut être utilisé par les systèmes de QR comme une collection fortement redondante ou encore comme ressource pour la création de bases de connaissances dédiées.

Le chapitre 7 aborde un domaine actuellement très actif en RI, qui consiste à prendre en compte divers aspects du contexte des recherches, tels que des caractéristiques propres à l'utilisateur ou à ses interactions avec le système de RI. Ce contexte peut être exprimé explicitement, ou inféré automatiquement. Les auteurs décrivent les extensions des modèles vectoriel et probabiliste qui découlent de la prise en compte du contexte. Ils évoquent la question de l'évaluation de tels systèmes de RI contextuelle, qui reste à approfondir malgré la création récente de tâches dédiées dans le cadre de TREC.

Le chapitre 8 rappelle le but des travaux autour du « Web sémantique » qui, depuis la fin des années 90, cherchent à construire autour du Web une plate-forme « universellement accessible » grâce à une description sémantique des documents. Ainsi, des outils ont été développés en vue de procéder à de telles descriptions et d'exploiter l'interopérabilité entre les documents annotés. Les auteurs exposent le rôle théorique et pratique des ressources terminologiques et des ontologies qui sont

au cœur de la RI dans le Web sémantique en tant qu'outils privilégiés permettant d'associer concepts et contenu informationnel des documents et requêtes.

Le chapitre 9 propose un panorama des travaux en indexation et RI sur les images fixes, les vidéos et les documents audio. Pour une grande part, il s'agit de transposer les méthodes développées avec succès pour le texte, par exemple en utilisant le texte accompagnant les éléments images, audio ou vidéo, ou en exploitant les liens entre ces éléments. Les résultats sont mitigés, en particulier pour les documents audio et vidéo. Il en ressort qu'un réel traitement des documents multimédias (particulièrement au niveau de l'indexation) reste un défi majeur à relever pour la communauté RI.

Le chapitre 10 décrit les interfaces permettant de visualiser les résultats d'un système de RI afin de faciliter leur traitement par l'utilisateur. Trois caractéristiques de ces interfaces sont présentées (aspects cognitifs, dimensionnalité, usage des couleurs) ainsi que leur mise en œuvre dans une large sélection de systèmes. La question de l'évaluation des interfaces de visualisation est d'autant plus complexe qu'il est difficile de dissocier l'interface du système de RI dont elle présente les résultats.

L'écueil auquel se heurte *Recherche d'information* est d'ordre éditorial. L'ouvrage manque parfois de consistance et de précision dans les notations utilisées. La première partie de l'ouvrage (chapitres 1, 2 et 5) gagnerait en clarté si les notions abordées étaient systématiquement définies, et la terminologie fixée. Par exemple, la notion d'indexation contrôlée n'est jamais réellement évoquée avant le chapitre 8 consacré au Web sémantique. Même si l'essentiel de la discussion porte sur l'indexation libre et automatique au sens de Salton, il aurait été souhaitable, d'une part de mentionner l'existence d'un autre type d'indexation et d'autre part de définir les termes « mot-clé », « mot », « terme » qui semblent parfois employés de manière interchangeable sans qu'on sache réellement de quoi parlent les auteurs. Le chapitre 2 introduit le terme de « dépistage » de l'information sans en donner aucune définition. On pourrait au début croire qu'il s'agit d'un certain type de recherche d'information, mais il semble finalement qu'il ne s'agit que d'un synonyme de la « recherche » d'information qui est le thème central de l'ouvrage. Des renvois entre les différents chapitres auraient favorablement joué sur la cohésion globale, par exemple dans le chapitre 5 qui reprend longuement des notions exposées dans les chapitres 1 et 2. En revanche, les chapitres 6 et suivants qui présentent les aspects les plus novateurs de la RI, paraissent beaucoup plus clairs, mieux structurés et constituent finalement l'attrait principal de l'ouvrage.

Thierry FONTENELLE, Practical Lexicography. A Reader, *Oxford University Press*, 2008, 405 pages, ISBN 978-0-19-929234-9.

Lu par **Michael ZOCK**

Laboratoire d'Informatique Fondamentale de Marseille

Beaucoup de choses ont changé dans le monde de la lexicographie, moins entre 1747 et 1980 (1747 étant la date de publication du premier article de cette anthologie), qu'entre 1980 et aujourd'hui, l'époque couverte par les articles de ce volume. Comme l'éditeur le note dans son introduction : l'arrivée des ordinateurs et des corpus électroniques a révolutionné notre manière de fabriquer des dictionnaires. Ce bouleversement est accompagné d'une remise en question de certaines de nos croyances concernant le sens d'un mot, sa représentation ou son illustration via *un exemple. Comme le livre se veut résolument pratique, on trouvera donc des réponses à des questions qu'un lexicographe est amené à se poser en exerçant son métier. Les mots ont-ils un sens, et, si oui, comment le représenter ? Quels termes utiliser dans une définition (prise en compte de l'utilisateur et de ses connaissances linguistiques) ? Qu'est-ce qu'un bon exemple ? Faut-il le construire à la main ou vaut-il mieux le chercher dans un corpus ? Dans ce dernier cas, comment éviter de noyer l'utilisateur sous la masse de données ? etc.*

Ce recueil de 416 pages comprend outre l'introduction, l'index et la bibliographie, une vingtaine d'articles, groupés dans douze sections. La première est consacrée aux apports de la linguistique théorique et aux aspects métalexicographiques (micro et macrostructure des dictionnaires). Elle contient trois articles. Celui de Johnson, intitulé *The Plan of a Dictionary of the English Language*, date du XVIII^{e} siècle. L'auteur y explique comment écrire et organiser un dictionnaire. Les deux autres consistent en une réflexion des apports de la théorie linguistique à la création d'un dictionnaire. Les deux sections suivantes sont consacrées aux corpus, leur conception et représentativité, à la toile comme corpus particulier et au rôle des linguistes à l'époque des sources électroniques. La section 4 est consacrée au *sens*. Hanks et Kilgariff débattent respectivement de la question de savoir si les mots ont un sens, alors que Stock aborde le problème des sens multiples, la polysémie. La section suivante est consacrée aux problèmes des collocations et à celui des expressions idiomatiques. Les auteurs des sections 6 à 8 s'intéressent aux définitions, aux exemples, à la grammaire et à l'usage des dictionnaires. La section 9 est consacrée aux dictionnaires bilingues. La 10^{e} section, intitulée « outils pour lexicographes », contient deux grands classiques, un sur les « associations de mots » et l'information mutuelle, l'autre sur le *sketch engine*, outil permettant de résumer en une page le comportement grammatical et collocationnel d'un mot dans un corpus donné. Enfin, Grefenstette médite sur le rôle des linguistes et lexicographes au prochain millénaire. Les deux dernières sections, appelées

respectivement « *réseaux sémantiques* » et « *usage de dictionnaires* » contiennent une introduction à *WordNet* et un article sur l'usage contrôlé des dictionnaires.

Pour apprécier ce livre à sa juste valeur il peut-être utile de savoir que : (a) le présent ouvrage est censé accompagner le nouveau manuel de lexicographie de Atkins et Rundell (2008), intitulé *Oxford Guide to Practical Lexicography* et publié chez le même éditeur (*Oxford University Press*). Si le rôle de cet ouvrage est de former des lexicographes en leur fournissant un support de cours, le rôle de celui-ci est de compléter le manuel d'Atkins et Rundell par des lectures afin d'éclairer le lecteur ou afin de l'emmener à se poser certaines questions; (b) la maison d'édition avait défini une série de contraintes en termes de contenus et de longueur, rendant difficile, pour ne pas dire impossible, l'intégration d'articles hors du champ direct de la lexicographie pratique. Ces contraintes expliquent en partie certains choix de l'éditeur, dont certains peuvent surprendre (voir plus loin). Ceci étant dit, c'est un très bel ouvrage, composé d'une introduction exemplaire et d'une série d'articles qui, dans l'ensemble, sont d'une excellente qualité. Sans aucun doute cet ouvrage sera très précieux pour des lexicographes et lexicologues.

Plutôt que de répéter des louanges faites par d'autres, certes justifiées, mais accessibles sur le Web[1], nous nous astreindrons ici à formuler nos regrets et nos suggestions, espérant qu'ils puissent être pris en compte dans un prochain volume, car, contrairement à la maison d'édition, nous pensons qu'il y a de la place pour un autre ouvrage ou une réédition remaniée. Nous tenons néanmoins à préciser que nous avons le plus grand respect pour le travail des auteurs et celui gigantesque de l'éditeur, dont la qualité d'introduction montre sa très grande compétence et maîtrise du sujet. L'introduction fournit non seulement une excellente initiation à la discipline, mais elle offre également les moyens d'y participer, ne serait-ce qu'intellectuellement.

Maintenant, voyons donc de plus près nos regrets : qualité et adéquation des choix des articles retenus (représentativité du domaine) et thèmes oubliés.

En ce qui concerne les articles retenus, on notera que plus de la moitié (12/22) ont été écrits par cinq auteurs, dont deux sont également auteurs de l'autre volume mentionné et publié par la même maison d'édition. Ceci peut donner l'impression qu'une très grande partie de la discipline a été forgée par ces auteurs, ce qui paraît contestable. Non pas qu'ils n'aient pas joué un rôle important, mais peut-être pas à ce point-là. Aussi peut-on se demander s'il n'y avait pas d'autres personnes ayant marqué la discipline et méritant leur place ici. Nous pensons immédiatement à des noms comme Mel'čuk ou Verlinde et Binon, pour ne citer qu'eux et pour rester dans le champ strict de ce livre, la lexicographie pratique. Sachant que l'éditeur connaît bien le travail de ces personnes on peut se demander jusqu'à quel point il était vraiment libre de ses choix, question qu'on peut également se poser concernant d'autres choix faits dans ce livre.

[1] http://www.cairn.info/revue-francaise-de-linguistique-appliquee-2008-1-page-134.htm

Notre dernier regret porte sur la diffusion d'informations concernant la structuration de dictionnaires (macrostructure), les problèmes d'interface, de navigation ou d'accès lexical. Certes, l'introduction à WordNet et les articles de Fontenelle et de Rundell y répondent partiellement, mais les problèmes d'indexation, d'accès par le sens**,** la fusion ou le croisement d'un dictionnaire avec un thésaurus, la notion de primitives ou de vocabulaire contrôlé pour formuler une requête se trouvent dispersés dans cet ouvrage et pas facilement repérables même à travers l'index donné, or, on aurait aimé voir un chapitre entier consacré aux problèmes d'accès et aux problèmes afférents.

Si ces critiques paraissent fortes, qu'on ne s'y trompe pas, notre but ici n'est pas de dénigrer ou de rabaisser un travail de très haut niveau. Critiquer est toujours facile et tout choix est susceptible d'être remis en question. Nous voulons simplement montrer que si les choses ont changé depuis 1747, notamment ces vingt-cinq dernières années, beaucoup de choses restent à faire. Le lecteur l'aura compris, la vie du lexicographe n'est pas forcément devenue plus simple avec l'arrivée des moyens modernes (ordinateurs et corpus électroniques), parfois bien au contraire. Et c'est pour nous permettre de faire face à ces nouveaux défis que ce livre peut nous servir de guide, ce qui est déjà un service énorme en soi. Au fil de l'étude de cet ouvrage, le lecteur découvrira donc que, contrairement à ce que Samuel Johnson laissait sous-entendre, l'activité du lexicographe, loin d'être une simple affaire de persévérance, est une activité intellectuelle exigeante, demandant de l'imagination, du savoir-faire ET de la réflexion. C'est un des très grands mérites de l'éditeur d'avoir mis cela en évidence, en écrivant une introduction tout à fait remarquable et en réunissant dans ce livre des articles souvent difficiles à trouver et sans aucun doute précieux pour tous ceux qui s'intéressent aux dictionnaires et à leur fabrication.