Introduction

Trop d'information tue l'information ! Ce vieux dicton est encore d'actualité à l'ère d'internet et des grandes bases de données documentaires. En effet, la production et la diffusion de documents électroniques conduisent à des stockages en masse d'informations souvent mal ou sous-exploitées, ce qui entraîne une perte potentielle de connaissances pourtant disponibles. Afin de pallier ce problème, des recherches se sont développées dans le domaine de la « fouille de texte » avec des méthodes issues de domaines aussi variés que la fouille de données, le traitement automatique du langage naturel, l'acquisition des connaissances, la recherche d'information, etc. Les textes ainsi fouillés comprennent deux parties : le texte lui-même et l'organisation sous-jacente (structures logiques et physiques (souvent hiérarchiques), graphiques ou stylistiques, balisage, liens hypertextuels, annotations, etc.).

Les huit articles présentés ici abordent la problématique de la fouille de texte selon des approches très diverses comme la classification ou catégorisation de textes, le traitement linguistique et psycholinguistique, l'extraction d'information à partir de la structure et de l'organisation des documents ou encore par la mise en œuvre de techniques d'apprentissage dans des systèmes spécifiques capables d'extraire de l'information au sein de grands corpus. Grâce à la pertinence des méthodes proposées et aux validations expérimentales effectuées, chacun de ces articles représente une avancée notable dans ce vaste domaine qu'est la recherche d'information.

En tout premier lieu, dans un article de synthèse, Yannick Toussaint retrace un historique de la fouille de textes et propose une définition mettant l'accent sur les spécificités de ce domaine. Il insiste sur l'importance à accorder au processus d'extraction de connaissances à partir de textes et sur la notion de structure de documents. Enfin, il présente plusieurs méthodes de fouille de textes en insistant sur la nécessité d'exploiter un modèle de connaissances pour envisager le processus de fouille de textes.

Dans l'article suivant, Ludovic Denoyer et Patrick Gallinari proposent un modèle général permettant la classification supervisée de documents structurés multimédias. C'est plus précisément un modèle génératif fondé sur les réseaux bayésiens qui est défini et évalué à partir de deux instances du modèle : l'une pour

les documents uniquement textuels et l'autre appliquée aux pages web (texte avec images). Le formalisme du noyau de Fisher est utilisé pour la définition d'un modèle discriminant. Une expérimentation effectuée sur trois grandes bases de données permet de valider ce modèle.

Jérémy Clech et Djamel Zighed s'intéressent plus particulièrement, dans le troisième article, à la problématique du réétiquetage automatique dans un contexte de catégorisation de textes. L'opération d'étiquetage, qui consiste à affecter les catégories d'appartenance à un document, est souvent réalisée de manière subjective par des experts et peut être considérée, dans certains cas, comme inconsistante. Dans cet article, les auteurs proposent une méthode de relaxation selon deux critères : la similarité et la cohérence locale moyenne en vue d'optimiser la cohérence de l'étiquetage. La collection Reuters-21578 ApteMod a été utilisée comme base expérimentale. Les résultats confirment qu'une méthode de relaxation permet de rendre plus consistant l'étiquetage initial de l'ensemble d'apprentissage d'un corpus de texte en modifiant automatiquement cet étiquetage.

Le quatrième article de Fernando Aguiar et Michel Beigbeder décrit une méthode de classification automatique fondée sur la mesure de similarité entre les nœuds d'un hypertexte. Cette méthode qui prend en compte les aspects structurels (les liens entre les nœuds) et le contenu textuel des nœuds, est implémentée pour la recherche d'information. Le système proposé indexe à la fois les nœuds et leurs contextes. La collection WT10g utilisée dans la Web Track de TREC en 2000 et les topiques 451-500 constituent la base de test utilisée pour la validation du modèle de recherche d'information et du modèle de requête. Ce dernier comporte deux niveaux : sujet et contexte.

L'article de Fabricio Enembreck et de Jean Paul Barthès aborde la recherche d'information sur le web à partir d'un système multi-agents, nommé MAIS. Il met l'accent sur l'importance de la spécialisation des agents pour la recherche et le filtrage et sur l'intérêt de la personnalisation de l'information. Le système MAIS est capable de récupérer et de classer automatiquement l'information à partir de recherches effectuées par des moteurs de recherche (All-The-Web, Altavista et Google). Il a été expérimenté directement sur la toile. Les résultats montrent que la prise en compte par MAIS des pages favorites et des centres d'intérêt de l'utilisateur améliore sensiblement les résultats de la recherche.

Nadine Lucas et Bruno Crémilleux traitent la fouille de données textuelles à partir d'une approche hybride exploitant des segments de textes fondés sur la hiérarchie de mise en forme. Cette approche met en œuvre d'une part des outils de fouille de données, de structuration et d'analyse de document et d'autre part la linguistique du discours. A partir d'une application concernant la détection de l'absence et de la présence de fautes de style dans des articles scientifiques en anglais, ce papier met en évidence l'intérêt d'une méthode fondée sur l'exploitation du contexte étendu d'un marqueur textuel. Deux méthodes de fouilles de textes ont été mises en œuvre pour caractériser la correction des textes : les règles de

caractérisation et les motifs émergents. Les résultats font ressortir plusieurs pistes d'exploration et illustrent la grande complémentarité entre la fouille de textes et la linguistique textuelle.

L'index du style situé à la fin d'un livre peut, selon Lyne Da Sylva, s'avérer une précieuse inspiration pour des outils de fouille de documents numériques. Le travail de recherche présenté dans son article vise à identifier les relations sémantiques présentes dans les index des livres et à déterminer lesquelles peuvent être dérivées automatiquement. La méthode exposée repose sur l'analyse de l'organisation textuelle et conceptuelle du document et sur l'extraction de la terminologie spécifique à celui-ci. Les premiers résultats, très prometteurs, laissent apparaître que l'index constitué de paires de termes (qualifiés de « vedettes/sous-vedettes ») est plus suggestif à l'utilisateur qu'un index contenant de simples listes alphabétiques de termes.

Enfin, la contribution de Bruno Gaume et de Karine Duvignau porte sur l'application de la notion de graphe de terrain de type *small world* à une représentation de dictionnaire intégrant des notions d'ergonomie cognitive fondées sur le plan psycholinguistique. Les premiers résultats permettent de penser que l'élaboration de dictionnaires électroniques en s'appuyant sur une théorie linguistique de l'organisation sémantique du lexique, s'avère en adéquation avec des processus d'acquisition précoce du lexique. Ainsi, la prise en compte d'aspects cognitifs pour la modélisation de dictionnaires électroniques ouvre la voie à des applications novatrices pour les documents électroniques en général.

La fouille de texte est encore « balbutiante ». Les recherches dans ce domaine sont récentes mais héritent de techniques issues de domaines voisins plus anciens : fouille de données, recherche d'information, modélisation de documents et d'hyperdocuments, etc. Les domaines d'application sont immenses et prometteurs.

Béatrice Rumpler

Jean-Marie Pinon