Exploration de textes et recherche d'information

Dominic Forest, Lyne Da Sylva

Canadian Journal of Information and Library Science, Volume 35, Number 3, September/septembre 2011, pp. 217-222 (Article)

Published by University of Toronto Press
DOI: https://doi.org/10.1353/ils.2011.0019

Exploration de textes et recherche d'information

Introduction

Depuis plusieurs années, de nombreux projets ont été entrepris afin de numériser et de rendre disponible en format numérique le patrimoine informationnel des organismes et des différentes branches du savoir. En favorisant l'accès à des ressources numériques de plus en plus nombreuses et dont la qualité, en termes d'encodage et de métadonnées, est des plus appréciables, ces projets ont en outre motivé le développement de techniques plus efficaces de recherche et d'analyse de l'information textuelle. Ainsi, étant donné le nombre croissant de ressources numériques, on a proposé des techniques et des stratégies visant à assister plus efficacement la recherche et l'analyse des documents textuels, que ce soit sur le web ou dans l'ensemble de la documentation générée par les organisations.

L'intérêt pour les techniques d'exploration de textes à des fins d'analyse et de gestion de l'information numérique s'inscrit dans cette perspective. En outre, ce domaine est actuellement un important lieu de recherche, dont les sources théoriques et pratiques proviennent principalement de travaux dans les domaines de l'exploration de données, de l'intelligence artificielle et de l'apprentissage machine (automatique). Toutefois, l'exploration classique de textes, notamment l'agrégation automatique, la catégorisation automatique et la reconnaissance des entités nommées, intègrent aussi des concepts et des techniques d'analyse qui ont émergé des domaines de la linguistique informatique. Dans cette optique, le domaine de l'exploration de textes se trouve caractérisé par le jumelage de différentes techniques d'analyse provenant tant des approches numériques issues de l'intelligence artificielle et de l'apprentissage automatique que des approches fondées davantage sur l'analyse et le traitement linguistique des données textuelles.

Récemment, plusieurs auteurs (Weiss et al. 2004; Feldman et Sanger 2006; Ibekwe-SanJuan 2007; Srivastava et Sahami 2009) ont indiqué qu'il fallait déterminer la pertinence de l'application de certaines techniques d'exploration de textes à des fins de recherche documentaire en sciences de l'information. Dans ce travail, des prototypes de moteurs de

recherche en ligne intégrant des fonctionnalités de regroupement automatique des résultats ont d'ailleurs été proposés (dont Carrot search, Clusty, etc.). Les résultats de ces prototypes se sont révélés des plus pertinents. Toutefois, peu de travaux ont véritablement exploré et évalué rigoureusement la pertinence de l'emploi de techniques d'exploration de textes dans un contexte de recherche d'informations.

* * *

Ce numéro thématique de la *Revue canadienne des sciences de l'information et de bibliothéconomie* présente les résultats des plus récentes recherches, dont l'objectif est d'exploiter des techniques d'exploration de textes dans une perspective de recherche d'information numérique. Les quatre articles de ce volume ont été sélectionnés parmi seize articles reçus par les principaux spécialistes de ces domaines dans le monde.

La contribution d'Andréani et de ses collègues est intitulée « Normalisation des entités nommées : allier règles déclaratives, ressources endogènes et processus centré sur l'utilisateur ». Cet article présente des techniques de normalisation des entités nommées dans les documents de la société TecKnowMetrix. Rappelons que les entités nommées sont les expressions qui dénotent des individus ou « entités uniques », notamment des lieux géographiques, des noms de personnes, d'organisations ou de produits, et des dates. Les entités nommées visées par ces auteurs sont tout d'abord des noms d'organisations, omniprésents dans leurs corpus de brevets, de publications scientifiques et d'articles de presse techno-économique. Ils visent à ramener à une forme canonique unique les différentes variantes d'une même entité nommée, par exemple Mitsubishi, Mitsubishi KK et Mitsubishi Corp. Cette normalisation est primordiale pour assurer un repérage fiable de toute l'information se rapportant à une entité donnée. Et elle ne peut bénéficier de l'appui de ressources comme un dictionnaire lexical ou terminologique, puisqu'elle s'applique à des noms propres qui n'obéissent à aucune réglementation.

Les techniques de normalisation utilisées comprennent : i) le recours à des lexiques particuliers qui permettent d'identifier la présence, le type et parfois la nationalité d'une organisation, ainsi qu'un dictionnaire de noms de pays du monde; ii) le découpage de noms complexes en leurs constituants de base; iii) la réécriture graduelle des noms d'organisation pour en arriver à une version normalisée; iv) l'appariement de la forme ainsi obtenue à d'autres formes déjà normalisées; et v) le repérage de sous-séquences fréquentes d'une expression donnée à l'intérieur du corpus.

Leur approche interactive est basée sur des règles et repose sur la validation par un utilisateur.

Des résultats d'évaluation par un expert humain sont présentés, en distinguant les cas où la normalisation est exacte, partielle ou incorrecte (avec bruit ou avec silence). Les normalisations exactes (variant selon le type de publication) représentent en moyenne 84 % pour les noms des organisations, 83,8 % pour le type (« académique » ou « entreprise ») et 62,4 % pour le pays d'origine de l'organisation. Les résultats sont inférieurs lorsque l'on combine deux ou trois des critères. L'évaluation met surtout en lumière le fait que 86,4 % des entités nommées ont fait l'objet d'une la procédure de normalisation avant d'atteindre la forme canonique attendue, ce qui indique que la procédure est primordiale pour obtenir un repérage efficace.

L'article illustre ainsi l'utilité des techniques d'exploration de texte (ici, fortement basées sur le traitement automatique de la langue) pour la recherche d'information.

L'objectif de l'article « Bilingual document clustering: evaluating cognates as features » de Denicia-Carral et de ses collègues est de regrouper (par *clustering*) des documents écrits en deux langues différentes, de manière à ce que les documents parlant des mêmes sujets se retrouvent dans les mêmes catégories, peu importe leur langue d'expression. La technique du *clustering* se fait généralement en retenant un certain nombre de caractéristiques pour chaque document et en rassemblant les documents qui partagent assez de caractéristiques communs. Le défi est toujours de choisir judicieusement ces caractéristiques. Dans cet article, les caractéristiques principales utilisées pour réaliser le *clustering* sont les liens apparentés (cognats) : des mots qui s'écrivent de manière identique ou presque, dans les deux langues, comme « construction » ou « céréale/cereal » en français et en anglais. L'idée est que les cognates sont très faciles à repérer (car ce repérage n'exige que très peu de traitements linguistiques) et assez fiables comme caractéristiques communes. (Bien sûr, certains cognats sont de faux amis, comme « librairie » et « library », mais on compte sur la relative rareté de ce phénomène pour minimiser les problèmes.)

Ainsi, l'objectif est atteint par des méthodes indépendantes des ressources linguistiques externes, et il est en réalité en grande partie indépendant de la langue—mais les auteurs insistent sur le fait qu'il faut que les langues

soient relativement proches (même alphabet, bon nombre de racines communes, etc.) pour que cette technique soit efficace.

Deux méthodes d'extraction de liens apparentés ont été utilisées : l'extraction de paires similaires (selon un calcul de similarité graphique), puis l'extraction de paires similaires dont le contexte est aussi similaire, c'est-à-dire qu'il contient des paires d'entités nommées identiques. Deux méthodes différentes de *clustering* ont aussi été mises à contribution (Direct et Star).

L'évaluation s'est effectuée sur un corpus bilingue anglais-espagnol. Les deux valeurs de base qui ont été définies correspondent à d'autres approches au même problème : d'abord, un *clustering* défini après la traduction des documents d'une langue vers l'autre, et ensuite un *clustering* basé uniquement sur les entités nommées. Les résultats dépassent l'efficacité de ces références, dont le premier exige notamment davantage de ressources, comme des corpus parallèles ou des dictionnaires bilingues. Les auteurs désirent pousser la recherche pour identifier d'autres méthodes indépendantes de la langue qui pourraient fournir des caractéristiques pour améliorer la classification.

Ce travail apporte des améliorations à un certain type de recherche d'information, celle des collections textuelles multilingues. En effet, le regroupement des documents au-delà de la langue facilite le repérage de tous les documents parlant d'un même sujet.

L'article « Modélisation automatique de connecteurs logiques par analyse statistique du contexte » de Charton et Torres-Moreno est de nature plus technique et traite de l'identification d'un certain type de synonymes, soit des connecteurs logiques (« donc », « en conséquence », « par conséquent », etc.) qui sont en quelque sorte interchangeables.

La méthode présentée par les auteurs est basée sur la recherche de contextes partagés par différents connecteurs logiques. Quand assez de contextes se répètent pour deux connecteurs logiques, on conclut qu'ils sont synonymes. Cela rappelle d'autres travaux sur la recherche de synonymes concernant les mots pleins et non les mots vides comme les connecteurs logiques, qui sont des adverbes, ou des locutions adverbiales ou des syntagmes prépositionnels.

On présente et on évalue la méthode pour divers textes (débats au Sénat, œuvres littéraires). Les résultats évalués portaient sur des contextes où un

équivalent d'un connecteur logique avait automatiquement remplacé autre; un spécialiste humain a évalué la justesse du résultat. Les substitutions correctes représentent de 79 % à 88 % des cas, selon le type de texte.

L'application à la recherche d'information consisterait notamment à décortiquer la structure des phrases et à aider ainsi au repérage d'expressions cibles. Le point de vue des auteurs quant à l'application de leur méthode à la substitution de synonymes laissent entrevoir encore d'autres applications pour la recherche d'information.

L'article « A sentiment-based digital library of movie review documents using Fedora » de Na et de ses collègues présente une bibliothèque numérique de critiques de films, qui permet des recherches sur la base des opinions exprimées dans les critiques (positives, négatives ou neutres) en ce qui a trait à divers aspects des films. Ainsi, non seulement peut-on rechercher un titre de film donné, mais il est aussi possible de repérer des films dont le travail du metteur en scène a été jugé favorablement ou des films jugés pauvres ou moins bons. L'accès est donc basé autant sur des champs de métadonnées que sur des aspects « émotifs » ou subjectifs des documents, obtenus par une analyse automatique de contenu.

D'abord, pour permettre de classifier l'analyse de sentiments à l'égard des différentes perspectives des films (metteur en scène, distribution ou film dans son ensemble), les phrases des critiques de film ont été étiquetées selon ces aspects. L'étiquetage a été effectué à l'aide de techniques d'extraction d'information, dont l'identification d'entités nommées, le repérage de coréférence entre deux expressions (p. ex. le fait que Carrey et Jim Carrey dénotent la même personne) et la résolution de pronoms (ou la recherche de leur antécédent). Ensuite, la classification des critiques selon ces aspects est effectuée par apprentissage automatique supervisé, à l'aide d'une machine à vecteurs de support (*Support Vector Machine*). Les auteurs évaluent la justesse de cette classification automatique oscillant entre 69,1 % et 90,48 % selon l'aspect considéré.

La suite de l'article décrit la bibliothèque numérique de critiques de films élaborée. Cette dernière permet la navigation et la recherche axées sur les sentiments exprimés envers le metteur en scène, la distribution et le film dans son ensemble. La bibliothèque comprend deux modules principaux : une interface web et une base de données (emmagasinées dans l'application Fedora), qui contient le résultat de l'analyse de sentiments et de la classification automatique décrite dans la première partie de l'article.

L'interface de recherche contient des champs qui correspondent aux métadonnées retenues (titre du film, distribution, metteur en scène). Les critiques de chaque film sont stockées séparément dans des fichiers différents, ce qui permet de ne consulter que ces parties. Lors d'une interrogation, l'utilisateur peut choisir de lire uniquement les jugements positifs, ou négatifs, ou neutres, ou encore toutes les critiques. Des éléments d'interface (une main avec le pouce vers le haut ou vers le bas) synthétisent les jugements et accélèrent la navigation parmi les résultats.

Ce travail présente donc un système complet, de la collecte de documents à leur annotation, en passant par la création de l'interface de recherche et de navigation. Il illustre clairement comment les techniques d'extraction de données sont utilisées pour soutenir le repérage d'information, qui peut prendre en compte des aspects affectifs des documents.

Le domaine de l'exploration de textes est un territoire de recherche des plus actifs, dont les lieux d'application sont nombreux, notamment la découverte d'informations, l'analyse d'opinions, et le suivi thématique. Ce numéro spécial ne présente toutefois que les plus récents développements de la recherche qui intègrent des techniques d'exploration de textes à des fins d'assistance à la recherche d'information. En plus de présenter les récents travaux dans ce domaine, nous sommes d'avis que ce volume permettra de bien comprendre la pertinence de jumeler les techniques d'exploration de textes à celles de la recherche d'information.

Dominic Forest et Lyne Da Sylva, École de bibliothéconomie et des sciences de l'information, Université de Montréal

Références

Feldman, R. et J. Sanger. 2006. *The text mining handbook*: *Advanced approaches in analysing unstructured data*. Cambridge: Cambridge University Press.

Ibekwe-SanJuan, F. 2007. *Fouille de textes : méthodes, outils et applications*. Paris : Hermès.

Srivastava, A. et M. Sahami (dir.). 2009. *Text mining: Classification, clustering, and applications*. Boca Raton: CRC Press.

Weiss, S.N., M. Indurkhya, T. Zhang, et F.J. Damerau. 2005. *Text mining: Predictive methods for analyzing unstructured information*. Berlin: Springer-Verlag.