Sciences

Informatique. Un outil permet de retrouver la langue, voire l'auteur d'un texte.

La linguistique comprimée par le zip

Par Denis DELBECQ

Le mercredi 13 mars 2002

Comment faire simple pour étudier ce qui est compliqué? Trois chercheurs italiens spécialistes de la complexité ont imaginé une méthode d'analyse du contenu de textes. Elle ne requiert aucune connaissance sémantique ou linguistique, et pour cause. Elle s'appuie sur le zip, un outil diffusé sur l'Internet pour compresser les textes et réduire leur temps de transmission. Le trio l'a détourné et affirme pouvoir découvrir la langue d'un texte, classer les langues par proximité et même découvrir l'auteur d'un écrit...

La compression de textes repose sur une idée simple. Rechercher les répétitions de caractères, mots et groupes de mots dans un texte, sans se soucier du sens. Et ensuite les coder par des symboles plus compacts.

La compression de textes repose sur une idée simple. Rechercher les répétitions de caractères, de mots et de groupes de mots dans un texte, sans se soucier du sens. Et, ensuite, les coder par des symboles plus compacts. La phrase «Avant l'heure, c'est pas l'heure. Après l'heure, c'est plus l'heure» (67 caractères) deviendra, par exemple, «Avant @ % pas @. Après @ % plus @» (32 caractères). Le «dictionnaire» des mots ainsi codés vient s'ajouter en fin de texte. Plus le document est long et plus la compression est aisée.

Proximité. Pour comparer les langues, l'équipe de Vittorio Loreto (université la Sapienza, Rome, Italie) a utilisé la déclaration des droits de l'homme, l'un des écrits les plus traduits au monde. Il en résulte que compresser un texte en basque - langue non indo-européenne - sur la base d'un dictionnaire de compression établi pour le même texte en italien n'est pas du tout efficace. Pour le français, en revanche, la proximité avec l'italien donne de meilleurs résultats. La comparaison croisée des différentes langues a permis de les classer par degré de proximité, sous forme d’un arbre que les linguistes connaissent bien.

Puis le trio italien a repris le même principe pour découvrir l’auteur d’un texte. Cette fois, la langue était commune, l’italien, mais les chercheurs disposaient des créations de onze auteurs. Quand le fragment étudié et le texte de référence sont d’une même plume, le style induit des répétitions de tournures de phrase. Quand le fragment ajouté au texte appartient à un autre auteur, la compression est dégradée. Sur la base d’une centaine de textes de onze écrivains, Loreto affirme avoir retrouvé le bon auteur dans 93 cas sur 100...

Publiés dans une revue de physique (1), ces travaux ont été accueillis avec scepticisme chez les spécialistes de l’analyse automatique de textes. *«C’est intéressant, mais je ne suis pas surpris de constater que du serbo-croate ressemble à du serbo-croate»,* résume Christian Leclere, de l’université de Marne-la-Vallée. Un avis partagé par Franz Guenthner, à l’université de Munich, qui a participé à la mise au point de l’outil de détection de langue installé sur Altavista, un moteur de recherches de documents sur l’Internet. *«On sait qu’en anglais, par exemple, 90 % des textes sont rédigés avec 10 % du vocabulaire.»* François Bourdoncle, qui dirige Exalead, une start-up appliquant les méthodes d’analyse statistique qu’il avait développées à l’Ecole des mines, juge également le travail peu novateur.

Piste. Mais Vittorio Loreto balaie les critiques et poursuivra ses recherches, notamment pour le classement thématique de documents. *«Ce n’est pas l’application de démonstration qui est importante, c’est la méthode. Elle peut s’appliquer à de nombreux domaines.»* Pour lui, la compression reste une piste pour améliorer la pertinence des résultats obtenus sur les moteurs de recherche d’informations, et devrait intéresser les as du *data mining,* qui fouillent les immenses bases de données informatiques d’entreprise.

Ce travail de forçat de l’analyse «minière» de données, les spécialistes du génome le pratiquent tous les jours. En quête de sens dans les longues chaînes d’ADN pour retrouver les gènes, comparer individus et espèces ou découvrir les protéines, ils utilisent une armada informatique lourde et onéreuse. Une idée simple comme celle de Loreto pourrait-elle aider les bio-informaticiens? *«Pourquoi pas,* répond William Saurin, directeur scientifique de Genomining, une start-up spécialisée dans l’analyse de l’ADN. *Il faut parfois des idées simples pour résoudre des* *questions compliquées. La compression permet, par exemple, de retrouver le patrimoine d'espèces qui se sont développées à partir d'un ancêtre commun.»* Le bio-informaticien Ming Li, de l'université de Californie à Santa Barbara, a d'ailleurs mis au point un outil similaire à celui de Loreto, qu'il utilise pour comparer des séquences d'ADN.

Les applications de la compression ouvrent des perspectives considérables aux chercheurs. *«Elle permet de traiter n'importe quelle suite de caractères sans se soucier de la façon dont elle a été obtenue»,* résume Vittorio Loreto. Des caractères, mais aussi des nombres. En comparant des séries d'enregistrements sismiques «compressées», on pourrait repérer des motifs répétitifs qui échappent à l'intuition humaine. Une équipe de géologues du gouvernement américain travaille sur la question. Reste l'éternelle quête du Graal: comment découvrir s'il n'existe pas un peu d'ordre dans les fluctuations boursières? La réponse est peut-être dans le zip.

(1) *Physical Review Letters* du 28 janvier 2002.
