Stage de M2 LID 2007-2008

Citation et autocitation dans les articles scientifiques

Encadrants : Nadine Lucas et Emmanuel Giguet

Contexte

Les articles académiques sont répertoriés et souvent classés, notamment en médecine, en mathématiques et en informatique. Pour évaluer l'intérêt des articles, on utilise les index de citations, la notoriété mesurée par les citations ultérieures d'un article par d'autres auteurs (un article largement cité est plus important qu'un article peu ou jamais cité). Les systèmes informatiques de référencement utilisent cette mesure : Google Scholar ou CiteSeer pour l'informatique par exemple. Pour éviter que les auteurs qui se citent eux-mêmes ne biaisent l'interprétation des résultats, il importe d'exclure les autocitations. CiteSeer utilise ce principe (2, 6, 7).

À l'inverse, on peut vouloir retrouver les citations et les autocitations pour suivre la démarche d'un auteur ou d'un groupe dans le temps. Cette recherche sert à évaluer l'originalité des travaux. Dans le même contexte et *a contrario*, la recherche de plagiat (reprise de parties d'article sans citation) informe la recherche d'antériorité.

Objectifs

L'objectif du stage est de faire le point sur les travaux concernant la citation, dans différentes communautés de chercheurs (8, 9, 10, 11). Pratiquement, il consiste à intégrer les connaissances disponibles pour repérer automatiquement les citations dans les articles scientifiques, en vérifiant trois points : la fiabilité (détection avec un maximum de certitude) ; la couverture en termes de domaine de savoir (informatique, médecine, physique, géographie…) ; et la robustesse en termes de langue/graphie traités. Pour distinguer les autocitations, il faut reconnaître les citations et les noms d'auteurs et les mettre en correspondance avec les auteurs de l'article examiné.

Les méthodes dans la lignée robuste de CiteSeer (2, 13) seront confrontées avec des méthodes statistiques. Peut-on améliorer la fiabilité des méthodes de détection, ou leur robustesse ? Le premier point à améliorer et évaluer est la gestion des formats et des types de citations (4, 14) ; le second est l'aspect multilingue/multiscript (11) ; le troisième concerne l'évaluation des résultats, quantitativement et qualitativement (en relation avec le temps de traitement).

Travail

Le travail comprend la familiarisation avec les types d'articles et les modes de citation (4, 10). Il suppose le maniement d'outils d'analyse de corpus et de maquettage (plate-forme wims, Java), de la technologie XML et des langages à expressions régulières (Perl, Python). Le stage comprend une réflexion sur les différentes approches d'investigation des articles scientifiques pour situer les complémentarités (1, 14), une part de développement de projet et une part d'évaluation comparative des méthodes existantes.

Du point de vue pragmatique, on attend une procédure efficace de traitement de l'information, capable d'assurer le repérage des citations et des autocitations, indépendamment de la langue, sur des documents en ligne. La démonstration sera faite sur un corpus de documents pris au hasard.

L'évaluation des performances sera assortie d'une description des limites de la procédure. Quels documents peuvent être traités ? (langue, format, discipline, longueur, style…).

Un plus serait la gestion de l'interface utilisateurs. Comment informer les utilisateurs des limites des méthodes utilisées ?On peut aussi évaluer la méthode employée du point de vue du génie linguistique, et les conséquences des choix en génie logiciel.

Références

1 BESAGNI Dominique and BELAÏD Abdel (2004) Citation recognition for scientific publications in digital libraries. *First International Workshop on Document Image Analysis for Libraries* (DIAL'04), Palo Alto, Ca USA. pp. 244-252.

2 CiteSeer Documentation en ligne http://citeseer, http://citeseer.ist.psu.edu

3 DÉJEAN Hervé (1998) *Concepts et algorithmes pour la découverte des structures formelles des langues*. Thèse Université de Caen.

4 FLOTTUM K. & RASTIER F. (eds.) (2003) *Academic discourse, multidisciplinary approaches* Oslo: Novus forlag.

5 GIGUET E. (1998) *Méthode pour l'analyse automatique de structures formelles sur documents multilingues*. Thèse Université de Caen.

6 GILES L., BOLLACKER K., LAWRENCE S. (1998) CiteSeer: an automatic citation indexing system. In: Witten, Ackscyn, Shipman (eds) *Digital libraries*. ACM Press.

7 LAWRENCE S., GILES L., BOLLACKER K. (1999) Digital libraries and autonomous citation indexing. *IEEE Computer* 32:67-71.

8 LUCAS N. (2004) La citation et l'appel à référence bibliographique dans les articles académiques. In: López-Muñoz JM, Marnette S, Rosier L (eds) *Le discours rapporté dans tous ses états: Question de frontières*. L'Harmattan, Paris, p 419-427.

9 MORRIS S. A. (2004) Manifestations of emerging specialties in journal literature: A growth model of papers, references, exemplars, bibliographic coupling, cocitation, and clustering coefficient distribution. *Journal of the American Society for Information Science and Technology* 56:1250-1273.

10 POUDAT C., LOISEAU S. (2004) Evaluation of authorial presence in academic genres. in *Strategies in Academic Discourse*, Tognini-Bonelli, E. and G. Del Lungo Camiciotti (eds.), 51–68.

11 VASSILEVA I. (2000) *Who is the author? A contrastive analysis of authorial presence in English, German, French, Russian and Bulgarian academic discourse*, St Augustin, Asgard

12 VERGNE Jacques (2001). Analyse syntaxique automatique de langue : du combinatoire au calculatoire, Tours, TALN 2001.

13 VOISIN L. (2002) Le correcteur automatique d'articles scientifiques en anglais. Stage de maîtrise sous la dir. de N. Lucas et J. Vergne, GREYC, Université de Caen.

14 ZITT M., BASSECOULARD E. (1998) Méthodes de structuration pour l'analyse stratégique des univers scientifiques: les techniques de citation *VSST'98. Actes de colloque associé au FAUST*, Toulouse, p. 31.