Annotation collective dans le contexte RI : définition d'une plate-forme pour expérimenter la validation sociale

Guillaume Cabanac

Université de Toulouse
Institut de Recherche en Informatique de Toulouse — IRIT UMR 5505 CNRS
cabanac@irit.fr

RÉSUMÉ. *Avec l'avènement du Web participatif, les lecteurs de documents électroniques sont de plus en plus actifs. En particulier, des systèmes d'annotation leur permettent de commenter, de reformuler, de critiquer, etc. des passages de documents. Les approches de RI qui ne considéraient jusqu'alors que le contenu des documents tendent actuellement à exploiter cette dimension participative du Web. L'activité des lecteurs (annotations et débats suscités) peut par exemple améliorer rappel et précision des résultats de RI. Dans des travaux précédents, nous suggérions de discriminer les annotations sur leur « validité sociale » (degré de confirmation exprimé par le groupe dans le débat associé) avant de les prendre en compte dans les processus de RI. Cet article décrit une plate-forme d'expérimentation visant à comparer les algorithmes de validation sociale proposés avec la perception humaine du consensus. Cette expérimentation toujours en cours a mobilisé 173 participants, les données recueillies sont en cours d'analyse.*

ABSTRACT. *With the advent of the participative Web, digital document readers gain in expressiveness. Annotations systems enables them to comment, to rephrase, to criticize, etc. passages from any document. IR approaches that considered document contents only nowadays tend to exploit this participative dimension of the Web. Readers' activity (annotations and related debates) may notably improve IR recall and precision. In previous works, we intended to integrate annotations into the IR process regarding their "social validation," i.e. the confirmation degree expressed by the group who debated. This paper depicts an experimental platform aiming to compare the proposed social validation algorithms with the human perception of consensus. This ongoing experiment has been raising 173 participants; acquired data is under study.*

MOTS-CLÉS : *RI, annotation, débat, validation sociale, expérimentation, perception humaine.*

KEYWORDS: *IR, annotation, debate, social validation, experiment, human perception.*

1. Introduction et motivations

Les systèmes de recherche d'information visent principalement à répondre au besoin d'un individu en extrayant les documents jugés pertinents au regard de sa requête, à partir d'un corpus préalablement indexé. Pour ce faire, les approches développées dans le domaine de la Recherche d'Information (RI) évaluent classiquement l'appariement entre la requête et le contenu des documents du corpus. Dans le même temps, les documents électroniques peuvent être annotés à l'image des documents papier. Cette activité d'annotation électronique répond aux divers besoins des lecteurs : lecture active et critique, prise de notes, appropriation de document en reformulant des passages, etc. De telles annotations sont dites « en contexte » car elles portent sur une partie du document électronique, localisée par le « point d'ancrage » de l'annotation.

Dans ce contexte de document électronique annoté, les travaux concernant le projet Responsa Retrieval (Fraenkel *et al.*, 1999) visent à enrichir la RI, en termes de rappel et de précision, par la prise en compte du contenu des annotations en plus du seul contenu des documents du corpus. Ainsi, pour améliorer le rappel, la requête est comparée au contenu des annotations afin d'enrichir les résultats avec les documents annotés — trouvés de façon indirecte. Concernant le gain en précision, les termes des annotations aident à désambiguïser les passages annotés. Dans le même temps, avec la démocratisation du Web, de nombreux systèmes d'annotation de document électronique ont vu le jour, tant dans le milieu universitaire que dans le secteur privé. Une contribution significative est à l'initiative du W3C avec le projet de système d'annotation Annotea (Kahan *et al.*, 2002). De tels systèmes exploitent la mise en réseau des postes informatiques en stockant les annotations dans un serveur dédié, accessible par tous les utilisateurs. De ce fait, les annotations partagées sont visibles par chaque utilisateur lorsqu'il consulte un document annoté. Ce dernier peut réagir à une annotation en initiant un « fil de discussion » en contexte : un débat portant sur l'annotation en question, auquel d'autres lecteurs peuvent également participer. Les travaux (Agosti *et al.*, 2005; Frommholz *et al.*, 2006) prennent en compte ces débats dans le processus de RI, en tant que faisceaux de preuves complémentaires aux contenus des documents.

Une limite des travaux de la littérature réside dans le fait qu'ils exploitent les annotations et débats sans toutefois évaluer leur pertinence. Ce sont pourtant des contributions subjectives provenant de lecteurs qui sont le plus souvent anonymes ou inconnus de l'utilisateur qui pose une requête. C'est pourquoi nous avons proposé d'évaluer la pertinence d'une annotation en fonction des opinions exprimées dans le débat associé (Cabanac *et al.*, 2007), en s'appuyant sur le concept de « validation sociale » introduit dans (Cabanac *et al.*, 2005). La validité sociale d'un débat reflète le degré de consensus ou de controverse atteint dans ce dernier, en fonction des opinions échangées par le groupe social qui débat. Les approches de RI précédemment citées pourraient alors être améliorées, tout comme des domaines connexes : RI contextuelle, opinion mining, analyses des tendances sur le Web… Cependant, avant d'explorer plus en détail ces pistes de recherche ébauchées dans (Cabanac *et al.*, 2007), nous souhaitons évaluer expérimentalement les algorithmes de validation sociale afin de vérifier la légitimité de ces pistes. Cet article a donc pour but de définir une plate-forme d'expé-

rimentation pour ces algorithmes. Ainsi, la section 2 rappelle brièvement leur objectif (agréger les opinions contenues dans un débat argumentatif) ainsi que leur fonctionnement. La contribution de cet article est exposée dans la section 3 : nous définissons un protocole pour évaluer expérimentalement, avec des individus, l'adéquation entre les résultats de l'algorithme de validation sociale et la perception humaine du consensus. Nous formulons l'hypothèse expérimentale correspondante ainsi que la méthode envisagée pour évaluer sa pertinence. Enfin, la section 4 décrit la plate-forme qui a d'ores et déjà permis à 173 personnes de prendre part à cette expérimentation en cours.

2. Validation sociale : agrégation des opinions d'un débat argumentatif

La validité sociale d'un débat argumentatif représente l'opinion obtenue en agrégeant les divers arguments exprimés au cours de ce débat. Le but recherché est d'obtenir un résultat aussi proche de la perception humaine du consensus que possible. Nous modélisons un débat à évaluer par un graphe orienté acyclique. Les nœuds du graphe représentent les arguments exprimés. Deux arguments peuvent être liés par un arc étiqueté avec une seule opinion : réfutation ($\mathcal{R}$), neutre ($\mathcal{N}$) ou confirmation ($\mathcal{C}$). Un débat peut être visualisé sous la forme d'un fil de discussion (une hiérarchie d'arguments) comme dans la Figure 1. Sur cet exemple, l'argument racine a_1 d'Alice initie le débat ; parmi les réponses obtenues, a_2 réfute a_1 (flèche barrée) et a_3 confirme a_2 (flèche normale). Nous avons proposé deux algorithmes différents pour calculer la validation sociale, ils traduisent deux approches différentes. La première approche (Cabanac *et al.*, 2005) est basée sur des mesures statistiques alors que la seconde (Cabanac *et al.*, 2006) est fondée sur la théorie bipolaire de l'argumentation en intelligence artificielle (Cayrol *et al.*, 2005). Concrètement, les algorithmes proposés calculent une valeur $v(a) \in [-1; 1]$ qui représente le degré de consensus ou de controverse du débat initié par l'argument a. Ils prennent notamment en compte le fait qu'une confirmation réfutée a moins de poids qu'une confirmation confirmée, et ce de façon récursive. Plus spécifiquement, la valeur $v(a) \to -1$ traduit la controverse : le groupe réfute globalement l'argument a. À l'opposé, $v(a) \to 1$ traduit le consensus lorsque le groupe confirme globalement a. Enfin, $v(a) \to 0$ lorsque la force globale des réfutations est contrebalancée par celle des confirmations : le groupe n'a pas trouvé un accord.

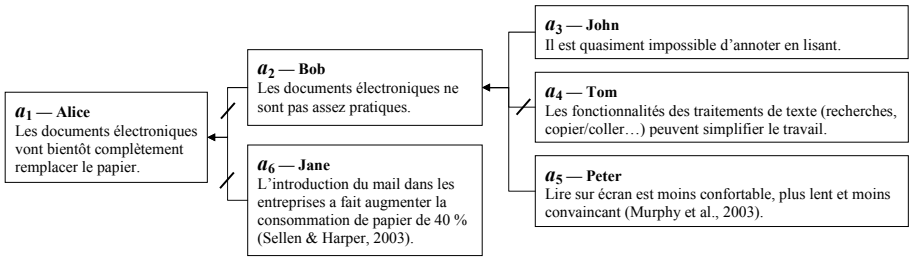

Figure 1. *Représentation graphique d'un débat comprenant six arguments.*

3. Protocole d'expérimentation écologique pour la validation sociale

Cette section définit un protocole d'expérimentation pour comparer la validation sociale avec la perception humaine du consensus. Ainsi, l'hypothèse expérimentale à (in)valider est : nos algorithmes fournissent des résultats proches de la perception humaine du consensus. L'adjectif « écologique » signifie que les participants (appelés cobayes par la suite) ne réaliseront pas l'expérimentation dans un laboratoire, sous la surveillance des expérimentateurs, mais dans leur environnement quotidien.

Pour valider les algorithmes, nous comparons leurs résultats avec la perception humaine du consensus exprimée par rapport à un corpus de débats argumentatifs. L'absence d'un tel corpus a été soulignée dans (Agosti *et al.*, 2005), c'est pourquoi nous avons pris l'initiative d'en constituer un. Afin de proposer des discussions réalistes et diversifiées, nous avons exploité des cartes d'argumentation (Twardy, 2004) publiées par des universitaires sur le Web [1] ainsi que les dialogues argumentatifs dans (Cayrol *et al.*, 2004). Après traduction des ressources françaises, nous avons obtenu 13 débats argumentatifs en anglais (soit 222 arguments) dont le contenu est accessible sur la page `http://www.irit.fr/~Guillaume.Cabanac/expe/corpus`. Leurs caractéristiques [2] en termes de nombre d'arguments $\langle 5\,;\ 17{,}1\,;\ 34\,;\ 8{,}1\rangle$, de profondeur $\langle 3\,;\ 4{,}2\,;\ 7\,;\ 1{,}3\rangle$ et de largeur $\langle 3\,;\ 7{,}9\,;\ 15\,;\ 3{,}2\rangle$ sont variées, il en est de même pour leurs thématiques : consommation de tabac, réchauffement climatique, alternatives à une intervention chirurgicale, orientation à l'université, interprétations de la Bible…

3.1. *Tâches d'un participant : étiquetage des arguments et synthèse des opinions*

Pour un cobaye, participer à l'expérimentation consiste à réaliser deux tâches (notées ❶ et ❷) pour chacun des 13 débats. Afin de limiter l'abandon en début d'expérimentation, les débats ont été ordonnés par difficulté croissante, estimée subjectivement à partir de leurs caractéristiques et des thématiques abordées. Le résultat du travail des cobayes est stocké dans une base de données dédiée (Figure 2). L'analyse de ces données permettra d'(in)valider l'hypothèse étudiée.

❶ **L'étiquetage des arguments** consiste à évaluer l'*Opinion* exprimée dans un Argument (3 choix exclusifs : Réfutation, Neutre, Confirmation) ainsi que le type du contenu grâce aux éléments de la classe *Commentaire* (3 choix non exclusifs : Modification, Question, Exemple). La combinaison des types *Opinion* et *Commentaire* permet de moduler la force d'un Argument : une Confirmation+Exemple est plus forte qu'une Confirmation seule, par exemple. Cette tâche ❶ vise à identifier objectivement l'opinion exprimée par un Argument au sujet de son père. Par exemple, la lecture des arguments a_2 et a_1 de la Figure 1 permet d'identifier une réfutation entre ces derniers.

1. cf. les sites Web `http://goreason.com` et `http://austhink.com/reason/tutorials` de l'université de Melbourne en Australie, `http://jostwald.com/ArgumentMapping` de l'université d'Ohio aux États-Unis.

2. Les quadruplets suivants sont de la forme ⟨minimum ; moyenne ; maximum ; écart-type⟩.

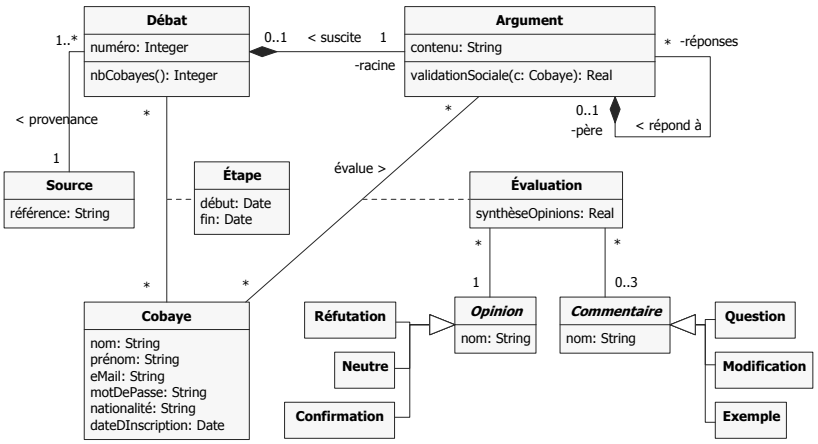

Figure 2. *Schéma conceptuel UML de la base de données de l'expérimentation.*

❷ **La synthèse des opinions** consiste à agréger mentalement, pour un argument père donné, les opinions exprimées dans ses fils. Cette valeur agrégée est stockée dans l'attribut synthèseOpinions de la classe Évaluation, c'est un nombre réel exprimé sur une échelle graduée entre « réfuté » et « confirmé ». Par exemple, on pourrait estimer que a_2 est confirmé à 66 % en synthétisant les opinions de a_3, a_4 et a_5 de la Figure 1. Cette estimation est rationnelle car elle reflète le taux $\frac{|\text{confirmations}|}{|\text{réfutations}|}$. Notons que ce raisonnement n'est pas suggéré aux cobayes pour ne pas influencer leurs décisions.

3.2. *Analyse des données fournies par les cobayes lors de l'expérimentation*

Les données recueillies durant l'expérimentation permettent de comparer le résultat des algorithmes de validation sociale avec la perception humaine des cobayes. Nous pouvons constituer une liste contenant des quadruplets $q = \langle$cobaye, argument, synthèseOpinions, argument.validationSociale(cobaye)$\rangle$ à partir de la base de données. Le troisième élément du quadruplet noté q_3 représente la synthèse des arguments fils qui répondent à l'argument père q_2. Le quatrième élément q_4 est le résultat de l'algorithme de validation sociale, calculé à partir des *Opinions* et *Commentaires* identifiés par le cobaye q_1. Notons que la validation sociale est calculée à partir des mêmes éléments que ceux utilisés par le cobaye pour fournir l'agrégation mentale q_3. Dans un premier temps, nous pourrons évaluer le degré d'accord inter-cobaye pour chaque débat, grâce au coefficient κ (Fleiss, 1971), afin de répondre à la question : identifient-ils les mêmes opinions dans l'étape ❶ ? Dans un second temps, pour évaluer la différence entre les deux séries (évaluation mentale q_3 *versus* évaluation algorithmique q_4) nous envisageons deux approches complémentaires. La première consiste à évaluer la corrélation entre les deux séries, en calculant le coefficient r de Pearson que l'on peut interpréter grâce aux tables de (Cohen *et al.*, 2003) par exemple. Puis, nous désirons confirmer

ces résultats par des tests statistiques de significativité. Pour ce faire, il existe dans la littérature statistique une batterie de tests qui permettent de s'assurer que deux méthodes fournissent des échantillons différents, avec une probabilité d'erreur de α %. Dans le domaine de la RI (Hull, 1993) détaille les différents types de tests disponibles. Pour de tels contextes une marge d'erreur de $\alpha = 5$ % est couramment admise, elle peut être toutefois assouplie dans le domaine des sciences humaines où la variabilité humaine est importante. Concrètement, il existe deux familles de tests : les tests paramétriques adaptés pour deux échantillons suivant une loi normale (en forme de cloche, centrée sur zéro) et les tests non-paramétriques qui ne font pas d'hypothèse sur la distribution des échantillons. Dans le premier cas qui est le plus courant, deux tests sont envisageables : le t-test de Student pour comparer deux méthodes ou l'ANOVA pour comparer plus de deux méthodes. Pour comparer les deux séries q_3 et q_4, nous retenons le t-test qui prend en compte la moyenne des erreurs entre deux séries, leur écart type et leur taille. En appliquant le t-test bilatéral sur données pairées, nous obtiendrons une valeur p à comparer avec α ; la différence entre les résultats des deux méthodes n'est pas statistiquement significative lorsque $p > \alpha$. Nous pourrons ensuite examiner les résultats présentant une différence importante entre q_3 et q_4. Cela nous permettra dans un premier temps de « contrôler » le jugement des cobayes, en invalidant notamment les évaluations irrationnelles (p. ex. le cobaye juge comme réfuté un argument qui ne possède pourtant que des réponses qui le confirment à l'unanimité). Ce type de bruit peut être dû à l'incompréhension de la tâche à réaliser, par exemple.

4. Plate-forme en ligne pour l'expérimentation écologique

Le protocole présenté dans la section précédente est implanté dans une plate-forme d'expérimentation en ligne conforme aux standards de la psychologie expérimentale (Reips, 2002). Cette plate-forme fédère trois composants : une page Web, un logiciel pour recueillir les données et une base de données stockant ces données. La page Web `http://www.irit.fr/~Guillaume.Cabanac/expe` présente l'objectif de l'expérimentation et décrit le travail attendu de la part des participants. Elle donne accès au logiciel d'expérimentation à proprement parler ; ce dernier est développé en Java-Swing et déployé sur le poste informatique du participant via Java Web Start. Suite à son inscription, un tutoriel décrit la tâche à réaliser pour chacun des 13 débats. Enfin, le logiciel affiche le premier débat dans l'interface graphique (Figure 3).

Le choix des participants est un point important pour toute expérimentation. Sélectionner uniquement des collègues ou des étudiants entraîne des biais dénoncés notamment dans (Reips, 2002). C'est pourquoi nous avons plutôt souhaité mobiliser des participants indépendants et diversifiés, sur la base du volontariat. Ainsi, c'est à partir d'avril 2007 que nous avons progressivement envoyé un appel à participation sur des listes de diffusions nationales (rtp-doc, bulle-i3, liste-egc…) et internationales (chi-students et chi-web de l'ACM, www-annotation du W3C, semanticweb, webir…). Au 20 mars 2008, 173 participants se sont inscrits à l'expérimentation et 109 l'ont effectivement commencée. Le nombre de participants décroît en fonction du rang du débat car l'expérimentation peut être stoppée (puis reprise) à tout moment. Ainsi, sur

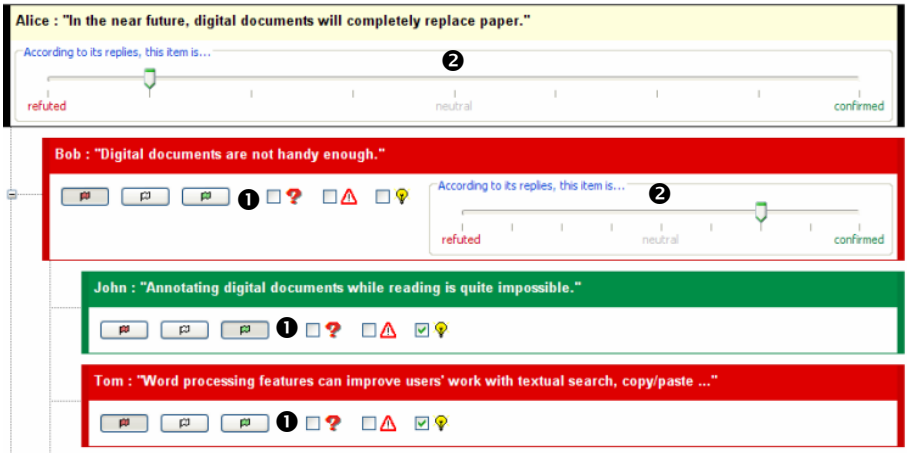

Figure 3. *Détail de l'interface graphique pour l'évaluation d'un débat.*

les 109 participants qui ont évalué au moins un débat, 48 ont effectivement terminé l'expérimentation à l'heure actuelle. Pour cette raison, nous continuons à solliciter les personnes afin d'obtenir un nombre représentatif de participations.

5. Conclusion et perspectives

Les systèmes de recherche d'information classiques répondent à un besoin informationnel en fournissant des documents dont le contenu est jugé pertinent par rapport à la requête posée. Afin d'améliorer la pertinence des résultats, des travaux dans le domaine de la RI prennent en compte les contenus subjectifs tels que les annotations collectives et débats associés aux documents, qui sont contribués par différents lecteurs. Nous avons souligné dans (Cabanac *et al.*, 2007) l'importance que revêt la discrimination de ces éléments en fonction de leur « validation sociale », concept introduit dans (Cabanac *et al.*, 2005). Par exemple, une annotation réfutée par un groupe social qui a débattu n'a certainement pas la même pertinence qu'une annotation confirmée. Afin de vérifier que les résultats de l'algorithme de validation sociale sont conformes à la perception humaine du consensus, nous avons défini dans cet article un protocole et une plate-forme d'expérimentation en ligne, de type écologique. Cet adjectif fait référence aux 173 volontaires hors laboratoire qui ont été mobilisés par l'intermédiaire d'annonces sur des listes de diffusion nationales et internationales, pour obtenir un échantillon le plus représentatif possible — cette mobilisation est toujours d'actualité. Leur participation consiste à évaluer 13 débats constitués à partir de ressources en ligne en ❶ identifiant l'opinion de chacun des 222 arguments puis en ❷ synthétisant les opinions exprimées par les arguments pères en fonction des opinions de leurs fils. À partir des 109 participations actuelles et des nouvelles inscriptions sur le site `http://www.irit.fr/~Guillaume.Cabanac/expe`, nous pourrons réaliser

les analyses présentées dans cet article. Dans l'hypothèse où les résultats seraient probants, l'intégration de la validation sociale dans les processus de RI exploitant la dimension sociale du Web devrait permettent d'améliorer leur pertinence.

6. Bibliographie

Agosti M., Ferro N., « Annotations as Context for Searching Documents », *CoLIS'05 : Proceedings of the 5th International Conference on Conceptions of Library and Information Sciences*, vol. 3507 of *LNCS*, Springer, p. 155–170, 2005.

Cabanac G., Chevalier M., Chrisment C., Julien C., « A Social Validation of Collaborative Annotations on Digital Documents », *in* J.-F. Boujut (ed.), *International Workshop on Annotation for Collaboration*, CNRS, p. 31–40, November, 2005.

Cabanac G., Chevalier M., Chrisment C., Julien C., « Validation sociale d'annotations collectives : argumentation bipolaire graduelle pour la théorie sociale de l'information », *INFORSID'06 : 24e congrès de l'INFormatique des Organisations et Systèmes d'Information et de Décision*, Éditions Inforsid, p. 467–482, June, 2006.

Cabanac G., Chevalier M., Chrisment C., Julien C., « Collective Annotation : Perspectives for Information Retrieval Improvement », *RIAO'07 : Proceedings of the 8th conference on Information Retrieval and its Applications*, CID, May, 2007.

Cayrol C., Lagasquie-Schiex M.-C., Bipolarité en argumentation, Rapport de recherche n° 2004-07-R, IRIT, Toulouse, February, 2004.

Cayrol C., Lagasquie-Schiex M.-C., « Gradual Valuation for Bipolar Argumentation Frameworks », *in* L. Godo (ed.), *ECSQARU'05 : Proceedings of the European Conference on Symbolic and Qualitative Approaches to Reasoning and Uncertainty*, vol. 3571 of *LNCS*, Springer, p. 366–377, 2005.

Cohen J., Cohen P., West S. G., Aiken L. S., *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*, 3rd edn, Lawrence Erlbaum Associates, Hillsdale, NJ, August, 2003.

Fleiss J. L., « Measuring Nominal Scale Agreement Among Many Raters », *Psychological Bulletin*, vol. 76, n° 5, p. 378–382, November, 1971.

Fraenkel A. S., Klein S. T., « Information Retrieval from Annotated Texts », *J. Am. Soc. Inf. Sci.*, vol. 50, n° 10, p. 845–854, 1999.

Frommholz I., Fuhr N., « Probabilistic, Object-oriented Logics for Annotation-based Retrieval in Digital Libraries », *JCDL'06 : Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries*, ACM Press, New York, NY, USA, p. 55–64, 2006.

Hull D., « Using Statistical Testing in the Evaluation of Retrieval Experiments », *SIGIR'93 : Proceedings of the 16th annual international ACM SIGIR conference on Research and development in information retrieval*, ACM Press, New York, NY, USA, p. 329–338, 1993.

Kahan J., Koivunen M.-R., Prud'Hommeaux E., Swick R. R., « Annotea : an open RDF infrastructure for shared Web annotations », *Comp. Netw.*, vol. 32, n° 5, p. 589–608, August, 2002.

Reips U.-D., « Standards for Internet-Based Experimenting », *Experimental Psychology*, vol. 49, n° 4, p. 243–256, 2002.

Twardy C., « Argument Maps Improve Critical Thinking », *Teaching Philosophy*, vol. 27, n° 2, p. 95–116, June, 2004.