

INTRODUCTION

L'évaluation en recherche d'information a une longue histoire. La tradition expérimentale dans ce domaine est fortement présente depuis les travaux de Cleverdon¹ dans les années soixante. De ces travaux sont apparus le besoin de collections qui permettent de travailler sur des données partagées par la communauté et des mesures d'évaluation définies selon la tâche à accomplir.

De très nombreuses campagnes d'évaluation ont suivi les travaux de Cleverdon en proposant des collections de volume toujours plus important et pour des tâches de plus en plus diversifiées. D'abord en langue anglaise, la diversification s'est orientée vers d'autres langues, puis vers des données au-delà du texte : hyperliens, structuration des documents, données sociales, ou encore par l'accès à des ressources externes aux documents telles que des ontologies. Outre les documents, les besoins d'informations doivent être créés en accord avec la tâche à accomplir. La première étape est donc la définition aussi précise que possible de la tâche, puis la construction et/ou la collecte de ces besoins. Associé à la collecte de documents, il y a la construction des jugements. Pour la tâche *ad hoc* de base, ces jugements indiquent quels sont les documents pertinents pour tel ou tel besoin d'information, avec toute la difficulté associée à cette notion de pertinence. Pour d'autres tâches ce sont d'autres qualifications des réponses possibles qui sont utiles ou d'autres niveaux de granularité pour la pertinence.

Du côté des mesures, pour la tâche *ad hoc* par exemple, les mesures prédominantes et largement acceptées aujourd'hui utilisent les notions de précision et de rappel. Il existe cependant de très nombreuses mesures sur cette base et des variations de ces mesures pour tenter de coller à telle ou telle particularité de telle ou telle tâche. Enfin, la mise en œuvre de ces mesures pour l'évaluation de tel ou tel système et plus particulièrement concernant l'amélioration des performances nécessite un certain soin. En effet, la différence purement quantitative n'est pas nécessairement la preuve absolue d'un gain réel. Pour cela des tests statistiques sont aussi mis en œuvre lors des comparaisons de systèmes.

Ce numéro spécial s'intéresse à la notion d'évaluation centrée sur la recherche d'information. L'objectif est de présenter des travaux qui fournissent un éclairage relatif à la notion d'évaluation que ce soit du point de vue système ou utilisateur. Cinq articles ont été retenus dont trois correspondent à une version étendue d'articles sélectionnés dans les conférences INFORSID 2014, CORIA 2013 et 2014.

Le premier article d'Orélie Desfriches Doria propose une étude sur l'articulation entre l'évaluation de l'information en lien avec les systèmes d'information, l'organisation du travail, l'évaluation au travail et l'activité.

1. Cyril Cleverdon (1960). Report on the Testing and Analysis of an Investigation into the Comparative Efficiency of Indexing Systems. The College of Aeronautics, Cranfield.

Le deuxième article de Bissan Audeh, Philippe Beaune et Michel Beigbeder présente une mesure visant à comparer le rappel des systèmes de recherche d'information.

Le troisième article de Patrice Bellot, Véronique Moriceau, Josiane Mothe, Éric SanJuan et Xavier Tannier, définit un cadre d'évaluation pour la contextualisation de tweets généralisable à d'autres textes courts. Une collection de référence ainsi que des mesures d'évaluation *ad hoc* sont proposées. Ce cadre d'évaluation a été expérimenté dans le contexte de la campagne INEX Tweet Contextualization. Il s'agit d'une version étendue de l'article présenté dans le cadre de CORIA 2013.

Le quatrième article écrit par Vincent Bouvier et Patrice Bellot, propose une approche faiblement supervisée à base de combinaison de modèles de langue dynamiques et temporels. Des métacritères permettent la désambiguïsation d'entités. Il s'agit d'une version étendue de l'article présenté dans le cadre d'INFORSID 2014.

Le dernier article de Léa Laporte, Sébastien Dejean et Josiane Mothe, propose de nouvelles méthodes de sélection de variables via des SVM parcimonieux pour l'apprentissage d'ordonnancement. Les méthodes ont été évaluées sur des collections de référence reconnues internationalement et une collection issue d'un moteur de recherche commercial. Les algorithmes suppriment jusqu'à 90 % de caractéristiques sans dégrader la qualité d'ordonnancement. Il s'agit d'une version étendue de l'article présenté dans le cadre de CORIA 2014.

Les articles soumis à ce numéro ont fait l'objet d'une évaluation réalisée par au moins trois relecteurs qui, par leurs commentaires, ont participé à la qualité des articles retenus. Nous souhaitons donc remercier le comité de lecture de ce numéro.

MICHEL BEIGBEDER
École des mines de Saint-Etienne
SYLVIE CALABRETTO
INSA Lyon
MAX CHEVALIER
IRIT, Université de Toulouse

COMITÉ DE LECTURE

Aurélien Bénel – TechCICO, Université de Technologie de Troyes
Guillaume Cabanac – IRIT, Université de Toulouse
Jean-Pierre Chevallet – LIG, Université de Grenoble
Vincent Claveau – IRISA, CNRS
Mihaela Juganaru-Mathieu – Laboratoire Hubert Curien, ENS Mines, Saint-Etienne
Gilles Hubert – IRIT, Toulouse
Léa Laporte – LIRIS, INSA de Lyon
Pierre-Edouard Portier – LIRIS, INSA de Lyon
Karen Sauvagnat – IRIT, Université de Toulouse