

## Extension sémantique du modèle de similarité basé sur la proximité floue des termes

Zoulikha Heddadji<sup>\*,\*\*</sup>, Nicole Vincent<sup>\*</sup>  
Séverine Kirchner<sup>\*\*</sup>, Georges Stamon<sup>\*</sup>

<sup>\*</sup>Université René Descartes  
45, rue des Saints Pères 75270 Paris CEDEX06  
<sup>\*\*</sup>CSTB-84, avenue Jean Jaurès Champs-sur-Marne  
77421 Marne-la-Vallée CEDEX2  
{zoulikha.heddadji, kirchner}@cstb.fr  
{nicole.vincent, Georges.Stamon}@math-info.univ-pris5.fr

**Résumé.** Le modèle flou de proximité repose sur l'hypothèse que plus les occurrences des termes d'une requête se trouvent proches dans un document, plus ce dernier est pertinent. Cette mesure floue est très avantageuse dans le traitement des documents à textes courts, toutefois elle ne tient pas compte de la sémantique des termes. Nous présentons dans cet article l'intégration d'une métrique conceptuelle au modèle de proximité floue des termes pour la formalisation de notre propre modèle.

### 1 Introduction

Dans le cadre de la modélisation des étapes du raisonnement à partir de cas pour la réalisation d'un outil logiciel qui fera office d'un tuteur d'aide pour l'évitement des circonstances de pollution domestique exprimées dans des plaintes (Z. Bellia, 2004), nous souhaitons améliorer la méthode de tri basée sur la contiguïté des termes de la requête dans le texte d'un document source. À l'évidence, il est dans l'intérêt de l'utilisateur du système de retrouver les cas les plus pertinents parmi les plaintes déjà traitées. Généralement, lorsqu'un utilisateur formule une requête au système, il compte retrouver les documents dont la signification du contenu se rapproche le plus de sa demande. Par exemple, pour la résolution d'une nouvelle plainte comportant le terme «couverture», il sera judicieux de retrouver les anciens cas de la mémoire archive relatifs non seulement au terme «couverture» lui-même, mais aussi aux «couettes», aux «duvets», aux «édredons», etc. Les documents contenant ces termes sont sans doute pertinents pour la plainte courante, néanmoins, ils ne seront pas sélectionnés par un modèle de recherche basé uniquement sur les occurrences directes des termes. Une solution incontournable est l'utilisation d'un réseau sémantique pour gérer le vocabulaire très variés qui peut être employé dans les plaintes. Dans l'étape de l'«élaboration» des cas en RàPC nous avons opté pour un modèle semi-structuré pour la constitution de la base. L'interface usager de notre système propose une série d'indexés sous forme de questions, dont les réponses apportent de l'information pour la description du problème. Nous avons proposé de traduire ces indexés sous forme de modèles de balise dans