RICSH : Recherche d'information contextuelle par segmentation thématique de documents

Rachid Aknouche, Omar Boussaid, Fadila Bentayeb

Laboratoire ERIC, Université Lumière Lyon2
5 avenue Pierre Mendès-France, 69676 Bron Cedex, France

{Rachid.Aknouche, Omar.Boussaid, Fadila.Bentayeb }@univ-lyon2.fr

Résumé. Le but principal des systèmes de recherche d'informations (SRI) classiques est de retrouver dans un corpus de documents l'information considérée comme pertinente pour une requête utilisateur. Cette pertinence est souvent liée à la fréquence d'apparition des termes dans le texte par rapport au corpus sans tenir compte du contexte de la recherche. Partant de ce constat, nous proposons dans cet article une approche pour la recherche d'information contextuelle par segmentation thématique de documents (RICSH). Cette approche s'appuie sur la méthode de pondération *tf-idf* que nous avons adaptée dans notre cas pour indexer le corpus. Cette adaptation se situe au niveau de l'importance du terme et de son pouvoir de discrimination par rapport aux fragments de textes et non au corpus. Ces fragments sont obtenus grâce à un processus d'identification des unités thématiques les plus pertinentes pour chaque document.

1 Introduction

Les systèmes de recherche d'informations classiques utilisent des méthodes de pondération et des mesures de similarités pour retrouver des textes pertinents par rapport à une requête utilisateur. La pondération *tf-idf*[1] est l'une des techniques les plus utilisées dans ces systèmes. Elle permet d'évaluer l'importance d'un terme dans un document par rapport à une collection ou à un corpus. Cependant cette formule, telle qu'elle est souvent présentée dans la littérature, ne tient pas compte du contexte de la recherche d'information (RI). On entend par contexte l'endroit d'apparition des termes recherchés dans un document. Il s'agit notamment des fragments de texte qui sont représentés sous forme de sections et de paragraphes dans un document. Par contre, une recherche pertinente devrait considérer ces éléments lors de la phase de formulation de la requête, dans le processus d'appariement et/ou dans la phase de classement des résultats. Les approches classiques de la RI utilisent les premières techniques, dites traditionnelles, qui sont basées sur la représentation linéaire des documents. D'après (Zargayouna, 2004), ces techniques procèdent à des requêtes plates (recherche par mots clés) et ignorent, par conséquent, la

1. désigne un ensemble de schémas de pondération de termes. *tf* signifie *(Term Frequency)* qui désigne le nombre d'occurrence du terme dans le document et *idf (Inverted Document Frequency)* qui est la valeur inverse du nombre de documents dans lesquels le terme est présent ou le pouvoir de discrimination de ce terme.