Classification de documents : calcul d'une distance structurelle

Karim Djemal, Chantal Soulé-Dupuy
Nathalie Vallès-Parlangeau

Université de Toulouse, IRIT, 118, route de Narbonne, F-31062 Toulouse cedex4, FRANCE
Karim.djemal@irit.fr
{Chantal.Soulé-Dupuy, Nathalie.Valles-Parlangeau}@univ-tlse1.fr

Résumé. La classification des documents numériques garantit un accès rapide et ciblé à l'information. Si nous considérons qu'un document est représenté par sa ou ses structures, définir des classes de documents revient à définir des classes de structures. Une classe structurelle représente donc des structures « proches ». Ainsi, associer la structure d'un document à sa classe structurelle revient à calculer une distance dite « structurelle ». Elle tiendra compte à la fois de l'organisation des éléments (position des nœuds, chemin), du coût d'adaptation des représentants des classes ainsi que de la représentativité des sous-graphes. Sur un corpus de documents représentant des notices de livres issus de la bibliothèque de l'université, nous discuterons de la construction de cette distance, de l'intérêt de chacun des trois paramètres utilisés.

1 Introduction

Les bases documentaires classiques retournent souvent de longues listes de documents qui doivent être parcourus afin de trouver l'information pertinente. La classification de documents offre une alternative permettant l'optimisation du processus de restitution sous forme de cluster (Soulé-Dupuy 2001).

Suivant l'objectif recherché, la classification des documents peut se faire suivant différents critères : par nature de documents (texte, video,…), par type de documents (cv, poème, lettre,…) ou encore par le contenu (thème, mots-clés,…). Nous nous intéressons ici à la classification par type de document, un type de document étant caractérisé par une structure particulière. On considère que la structure documentaire entre deux types de documents est suffisamment discriminante pour établir des classes dites « structurelles ». Les questions qui se posent à ce niveau sont : quels sont les facteurs discriminants entre des structures de documents (organisation des nœuds, types de nœuds, types de relations entre nœuds...) ? Qu'est-ce qu'une distance structurelle ? Est-ce qu'une classe doit regrouper des documents respectant une même DTD ou des documents structurellement « proches » ?

Après la présentation d'un état de l'art sur la classification de documents et la présentation du contexte de notre étude, nous abordons ces questions en proposant un calcul de distance que nous appelons « distance structurelle ». Cette distance se base sur trois critères dont nous essaierons de trouver la meilleure combinaison. Enfin, nous détaillons l'impact de ce calcul sur la classification d'un corpus de documents représentant des notices de livres issus de la bibliothèque de l'université.