Titre du projet : Application de techniques de forage de textes à des fins de gestion et d'analyse thématique de documents textuels non structurés

Résumé :

Depuis les dix dernières années, on observe une hausse considérable du nombre d'initiatives visant à numériser et à rendre disponible le patrimoine informationnel des organisations et des différentes branches du savoir. Les conséquences découlant de ces initiatives sont importantes et très nombreuses. Elles ont entre autres engendré l'émergence d'applications permettant différentes opérations complexes d'analyse et de gestion des documents. Malgré la diversité de ces applications, on constate qu'un axe de recherche partagé par l'ensemble des disciplines reliées à l'analyse et à la gestion des documents textuels réside dans la compréhension et de l'informatisation des processus d'identification des contenus thématiques et de l'analyse thématique.

Le projet que nous présentons aborde précisément la problématique de l'identification des thèmes et de l'assistance à l'analyse thématique des documents textuels. L'objectif général du projet est de développer et de valider une méthodologie informatique fondée sur la classification et la catégorisation automatiques permettant d'assister efficacement l'identification des thèmes et, surtout, l'analyse thématique de documents textuels. Il vise ainsi à effectuer un transfert de concepts et de méthodologies provenant, d'une part, des recherches théoriques et pluridisciplinaires sur l'analyse thématique et, d'autre part, des recherches appliquées en classification et en catégorisation automatiques des données afin de développer une méthodologie et un prototype d'application flexible visant à assister le chercheur dans son travail d'analyse thématique des textes. Le défi principal de ce projet réside donc dans l'arrimage entre ces deux objectifs : une opérationnalisation de l'analyse thématique avec les stratégies de classification et de catégorisation informatiques des textes.

Une première particularité de notre démarche consiste à jumeler les processus de classification et de catégorisation en appliquant d'abord le processus de classification sur les données initiales et, par la suite, le processus de catégorisation sur les résultats de la classification.

Par ailleurs, la classification et la catégorisation sont des opérations traditionnellement appliquées aux documents entiers. Contrairement à cette démarche, nous proposons une manière alternative de réaliser ces processus. Ainsi, au niveau de la classification, notre démarche consiste d'abord à segmenter chacun des documents puis à soumettre au classifieur les différents segments de texte. Cette démarche, lorsqu'elle est jumelée au processus de catégorisation, possède l'avantage d'attribuer plusieurs catégories thématiques à chaque document (ce qui est plus difficilement réalisable lorsque les documents sont traités en entier).

Finalement, dans bon nombre d'applications d'analyse et de gestion des documents textuels, le processus de catégorisation est effectué en utilisant un plan de classification ou une taxinomie de catégories prédéfinies. Le développement de ces taxinomies, bien qu'il puisse être assisté dans certains cas à l'aide d'applications informatiques, s'avère coûteux et très complexes. Dans ce projet, nous démontrons qu'il est possible, en l'absence de taxinomies, d'employer certains termes du lexique initial du corpus comme étiquettes thématiques.