

Contextual-RTM: un cadre général pour la modélisation de thématiques dans les réseaux de documents

Jean Dupuy^{***}, Adrien Guille^{**}, Julien Jacques^{**}

^{*}MeetSYS

^{**}Université de Lyon, Lyon 2, ERIC EA3083

Résumé. Les longs parcours de bases de connaissance ou de grands corpus de documents en réseau peuvent parfois perdre le lecteur et lui faire manquer certains liens entre les documents qu’il consulte. Nous proposons ici un cadre de modélisation thématique pour les réseaux de documents, Contextual-RTM, qui généralise *Relational Topic Model* (RTM). Alors que RTM est agnostique à la position des liens dans les documents, Contextual-RTM les prend en compte, permettant une meilleure contextualisation. Nous définissons le contexte comme une fenêtre de taille ajustable centrée autour du lien et proposons 3 méthodes d’agrégation du contexte : uniforme, positionnelle et sémantique. Contextual-RTM se montre compétitif sur des tâches d’identification de mots à l’origine de liens entre documents. Nous intégrons ces méthodes dans un système d’aide à la lecture capable d’inférer localement des liens latents entre documents. Ainsi le lecteur garde une trace de ses précédentes lectures, et s’en voit recommandé de nouvelles.

1 Introduction

Nous nous intéressons à la prédiction de liens entre documents, avec comme objectif la mise au point d’un système de recommandation et d’aide à la lecture. Nous nous attachons plus particulièrement aux réseaux de documents où les liens entre les pages sont localisés dans le texte. C’est le cas par exemple des liens hypertextes dans les pages Wikipédia. C’est également le cas des citations dans les articles scientifiques.

Nous portons une attention particulière à la localisation des liens dans le texte, et aux mots qui les entourent (que nous appelons leur contexte) car nous pensons que cette information est porteuse de sens et permet de mieux expliquer l’existence d’un lien entre deux documents. Par exemple les liens qui se trouvent dans l’introduction d’un article scientifique sont généralement vers d’autres articles relativement généraux, ceux dans le corps de l’article plus spécialisés sur une tâche similaire à celle qui est traitée dans l’article, et ceux en conclusion vers des tâches spécifiques mais différentes.

Tandis que les systèmes existants recommandent un ensemble de documents à lire, étant donné le dernier document consulté, nous souhaitons mettre au point un système