Présentation de Sujet de Recherche

Améliorer les descripteurs de documents semi-structurés en utilisant les informations contextuelles [1]

Antoine Doucet*,**

** Université de Caen, Département informatique*
Campus Côte de Nacre F-14032 Caen Cedex
doucet@info.unicaen.fr

*** University of Helsinki, Department of Computer Science*
P.O. Box 26 (Teollisuuskatu 23), FIN-00014 University of Helsinki, Finland

1. Introduction

Le nombre et la taille des collections de documents semi-structurés ne cessent de croître. Pourtant, la plupart des méthodes de recherche d'information se focalisent sur le texte seul ou uniquement sur certains éléments de sa structure. Dans le cadre d'une cotutelle de thèse avec l'université d'Helsinki, j'étudie la notion de séquence fréquente maximale (Ahonen, 1999) en tant que descripteur de document, et les différents moyens d'y attacher des informations contextuelles.

2. Présentation et Problématique

Des informations telles que la structure d'un document, l'analyse syntaxique des mots qu'il contient, ainsi que d'éventuelles données géographiques sont autant d'éléments qui doivent permettre d'améliorer la qualité des descripteurs. Dans la mesure où ce type d'information est désormais présent dans la plupart des documents textuels, adapter les algorithmes existants, afin qu'ils tiennent compte de ces informations contextuelles, présente un intérêt certain. Par exemple, les documents géographiques contiennent des cartes, des tableaux et des images.

Inclure ce type de données dans les descripteurs est un nouveau challenge pour améliorer leur qualité. Les applications sont nombreuses, et comprennent notamment

[1] Projet numéro 73643 (Doremi) financé par l'Académie de Finlande (« *Suomen Akatemia* »).

la recherche globale d'informations, la fouille de données appliquée à de grandes collections de document (Ahonen *et al*., 1999), et la navigation dynamique.

Ainsi, j'envisage d'explorer différentes manières d'inclure ces informations contextuelles, qui peuvent prendre plusieurs formes:

- **Information structurelle** : Le chemin dans lequel un mot apparaît (à l'intérieur d'un document SGML), est d'évidence une information supplémentaire au texte brut. En tenir compte dans les méthodes d'extraction de descripteurs est nécessairement une amélioration.

- **Information géographique** : Les cartes, tableaux, images... sont des informations géographiques typiques. Mais ce type de données est radicalement différent de la donnée usuelle: le mot. Les données géographiques ne sont pas séquentielles; l'ordre dans lequel elles apparaissent a généralement peu ou pas d'importance. Inclure des données non-séquentielles à des algorithmes basés sur des données séquentielles est une sérieuse difficulté.

- **Information morphologique** : A l'aide d'un analyseur syntaxique, les mots peuvent être reliés à des vecteurs caractéristiques, contenant des données morphologiques. Ce type d'information permet, par exemple, de désambiguïser le sens des mots, au-delà de leur simple graphie. Ceci agrandit clairement le spectre de la découverte de connaissances.

Toutefois, en prenant ces données en compte, de nombreuses difficultés apparaissent. Elles sont principalement liées à la complexité algorithmique. Par exemple, la quantité d'information par mot, obtenue à l'aide d'un analyseur syntaxique est potentiellement énorme. Une première étape consiste à trouver un moyen de présélectionner ces informations. Il faut aussi trouver un moyen de représenter ces informations. Une première approche pourrait être de simplement remplacer chaque mot par un vecteur caractéristique. Mais il reste alors à intégrer ces diverses informations, correspondant à un seul mot, au sein des algorithmes de traitement. C'est un vrai problème, dans la mesure où les méthodes ont été conçues pour des informations uni-valuées (le mot lui-même et rien d'autre).

C'est à cela que peut schématiquement se résumer la difficulté de mon objectif: adapter des idées conçues pour des informations séquentielles et uni-valuées au cas d'informations semi-séquentielles et muti-valuées.

3. Bibliographie

Ahonen-Myka H., "*Finding All Frequent Maximal Sequences in Text*", in ICML 99 Workshop on Machine Learning in Text Data Analysis, Ljubljana, Slovenia, 1999, p. 11-17.

Ahonen-Myka H., Heinonen O., Klemettinen M., Verkamo A. I., "*Finding Co-occurring Text Phrases by Combining Sequence and Frequent Set Discovery*", in IJCAI 99 Workshop on Text Mining: Foundations, Techniques and Applications, Bled, Slovenia, 1999, p. 1-9.