Application des méthodes à noyaux à la fouille de données textuelles

Jean-Michel Renders

▶ **To cite this version:**

Jean-Michel Renders. Application des méthodes à noyaux à la fouille de données textuelles. Jun 2004. sic_00001251

HAL Id: sic_00001251

https://archivesic.ccsd.cnrs.fr/sic_00001251

Submitted on 8 Dec 2004

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Application des méthodes à noyaux à la fouille de données textuelles

Jean-Michel Renders
Xerox Research Center Europe
6, Chemin de Maupertuis
38240 Meylan
France
Jean-michel.renders@xrce.xerox.com

RÉSUMÉ. *Les méthodes à noyaux connaissent depuis peu une utilisation intensive dans le domaine du traitement automatique du langage et en particulier pour la résolution de tâches en fouille de données textuelles. Les Noyaux définissent une mesure de similarité généralisée entre objects de structure arbitraire, avec trois propriétés intéressantes : la faculté d'incorporer de la conaissance a priori sur le domaine, la projection implicite des données dans un espace où la représentation est plus riche et la résolution du problème plus aisée, la (relative) indépendance des algorithmes d'apprentissage par rapport à la dimension de ce nouvel espace. Ces propriétés, couplées avec des algorithmes d'apprentissage robuste – pour des tâches telles que le classement, la classfication, la réduction de dimension, le filtrage, ... – sont à la base de résultats remarquables pour des tâches de Fouille de données textuelles, telles que la catégorisation et la classification de documents et de concepts, la désambiguisation sémantique et structurelle, l'extraction d'informations et de relations entre entités, l'extraction automatique de lexiques multilingues, ...*

ABSTRACT. *Kernel methods have recently been introduced to solve Natural Language Processing and Text Mining problems. Kernels define a generalised similarity measure between objects of arbitrary structure, with three interesting properties, namely the ability to incorporate prior knowledge about the problem, the implicit mapping of the data into a new feature space, which allows for very richer representation and where problem solving is easier, and finally the independence of learning algorithms from the dimension of this new feature space ("the Kernel trick").*

These properties, coupled with robust learning algorithms (for classification, clustering, dimension reduction, filtering, ...) provide some remarkable results in Text Mining tasks, such as document categorization, concept clustering, word sense disambiguation, information extraction, relationship extraction and automatic multilingual lexicon extraction.

MOTS-CLÉS : *méthodes à noyaux , noyaux sur données structurées, noyaux pour données textuelles, noyaux interlangues, catégorisation de documents, classification de concepts, désambiguisation sémantique, désambiguisation structurelle, extraction d'information, extraction de relations, extraction automatique de lexiques multilingues*

KEYWORDS: *kernel methods, kernels on structured data, kernels for NLP, kernels for textual data, kernel engineering, cross-lingual kernels, documents categorization, concept clustering, word sense disambiguation, information extraction, relationship extraction, automatic multilingual lexicon extraction*

Actes de CIFT 2004, pages 117 à 119

 Actes de CIFT 2004, pages 117 à 119

1. Résumé étendu

Les méthodes à noyaux connaissent depuis peu une utilisation intensive dans le domaine du traitement automatique du langage et en particulier pour la résolution de tâches en fouille de données textuelles. Les Noyaux définissent une mesure de similarité généralisée entre objects de structure arbitraire, avec trois propriétés intéressantes : la faculté d'incorporer de la conaissance a priori sur le domaine, la projection implicite des données dans un espace où la représentation est plus riche et la résoltuin du problème plus aisée, la (relative) indépendance des algorithmes d'apprentissage par rapport à la dimension de ce nouvel espace.

La première est de pouvoir introduire, lors de la conception d'un noyau particulier, de la connaissance a apriori sur le problème, principalement de la connaissance linguistique obtenue par analyse syntaxique et sémantique. La seconde réside dans la transformation implicite des objets dans un nouvel espace, où la résolution du problème s'avère plus aisée : la structure du problème dans l'espace transformée est plus simple et des méthodes linéaires robustes peuvent être utilisées. Enfin, le nouvel espace des objets transformés peut posséder une dimension très élevée (et même infinie), ce qui permet une représentation très riche, tout en ne pénalisant pas la complexité des algorithmes parce que les algorithmes mis en œuvre ne requièrent jamais de travailler explicitement dans ce nouvel espace (« the kernel trick »).

Cette présentation a pour but d'expliquer les fondements et les idées maîtresses qui se cachent derrière les noyaux, en prenant des exemples dans le domaine de la fouille de données textuelles. Nous ébaucherons ainsi une sorte d'ingénierie des noyaux, en nous concentrant sur leur construction. Nous partirons d'abord de règles de construction et de combinaisons générales (comment combiner deux types élémentaires de noyau pour en concevoir un nouveau, plus adapté à la tâche ?). Nous examinerons aussi comment spécialiser ou adapter des noyaux par rapport aux données (comment concevoir des noyaux optimaux compte tenu d'un ensemble d'exemples disponibles, pas forcément annotés ?).

Ensuite, nous nous focaliserons sur des noyaux spécialement conçus pour des entités textuelles. On peut voir un texte comme un « sac de mots », un ensemble de concepts, une chaîne de caractères, une séquence de mots, une séquence de concepts, un arbre (ou un graphe) de dépendance. Suivant notre vision et notre représentation de ces entités textuelles, nous obtiendrons différents noyaux. Dans ce domaine, nous décrirons ainsi les noyaux les plus importants qui sont : les noyaux « sacs de mots », les noyaux à sémantique latente, les noyaux pour chaîne ou séquence de mots, les noyaux pour arbres et pour graphes acycliques orientés. Nous nous focaliserons sur la conception de noyaux spécifiques aux entités textuelles, débutant avec des noyaux encodant de la connaissance liguistique (noyaux à lissage sémantique), poursuivant avec les noyaux de convolution (noyaux définis sur une structure de données qui peut se définir récursivement) et terminant avec les noyaux

à base de modèles génératifs (noyaux pour lesquels la « topologie » du problème est traduite en une fonction noyau).

Nous verrons ensuite comment ces différents noyaux peuvent être mis en œuvre pour la résolution de tâches de fouilles de données textuelles telles que la catégorisation de textes, l'extraction et la classification de concepts, la désambiguisation sémantique et structurelle, l'extraction d'information, l'extraction de relations entre entités et l'extraction automatique de lexiques spécialisés multilingues.

2. Bibliographie

N. Cancedda, E. Gaussier, C. Goutte, and J.-M. Renders. Word-sequence kernels. Journal of *Machine Learning Research* 3:1059-1082, 2003.

M. Collins and N. Duffy. Convolution kernels for natural languages. *NIPS'2001*.

C. Cumby and D. Roth. On kernel methods for relational learning. *ICML'2003*.

T. Gartner. A survey of kernels for structured data. *SIGKDD explorations 2003*.

D. Haussler. Convolutional kernels on discrete structures. Technical report, Department of Computer

J. Kandola, J. Shawe-Taylor, and N. Cristianini. Learning semantic similarity. *NIPS'2002*.

H. Kashima, K. Tsuda, and A. Inokuchi. Marginalized kernels between labeled graphs. *ICML'2003*.

R.I. Kondor and J. Lafferty. Diffusion kernels on graphs and other discrete input spaces. *ICML'2002*.

H. Lodhi, J. Shawe-Taylor, N. Cristianini, C.Watkins. Text classification using string kernels. Journal f Machine Learnin Research 2:419-444, 2002. *SANN'2002*.

C. Saunders, J. Shawe-Taylor, and A. Vinokourov. String kernels, Fisher kernels and finite state automata.*NIPS'2002*.

J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press. 2004

G. Siolas and F. d'Alche-Buc. Mixtures of probabilistic PCAs and Fisher kernels for word and document odeling. *ICANN'2002*.

J. Suzuki, Y. Sasaki, and E. Maeda. Kernels for structured natural language data. *NIPS'2003*.

K. Tsuda, M. Kawanabe, G. Ratsch, S. Sonnenburg, and K.-R. Muller. A new discriminative kernel rom probabilistic models. *Neural Computation* 14:2397-2414, 2002.

C. Watkins. Dynamic alignment kernels. Technical report, Department of Computer Science, Royal olloway, University of London, 1999.

D. Zelenko, C. Aone, and A. Richardella. Kernel methods for relational extraction. *Journal of Machine Larning Research* 3