Matching Evolving Hilbert Spaces and Language for Semantic Access to Digital Libraries

Peter Wittek[1], Sándor Darányi[2], and Milena Dobreva[3]

[1] Department of Computer Science. National University of Singapore, Computing 1, Law Link, Singapore 117590,
[2] Swedish School of Library and Information Science, University of Borås and Göteborg University, Allégatan 1, 50190 Borås, Sweden
[3] Centre for Digital Library Research, Information Services Directorate, University of Strathclyde, Livingstone Tower, 26 Richmond Street, Glasgow, G1 1XH United Kingdom

1 Background

Extended by function (Hilbert) spaces, the 5S model of digital libraries (DL) [1] enables a physical interpretation of vectors and functions to keep track of the evolving semantics and usage context of the digital objects by support vector machines (SVM) for text categorization (TC). For this conceptual transition, three steps are necessary: (1) the application of the formal theory of DL to Lebesgue (function, L2) spaces; (2) considering semantic content as vectors in the physical sense (i.e. position and direction vectors) rather than as in linear algebra, thereby modelling word semantics as an evolving field underlying classifications of digital objects; (3) the replacement of vectors by functions in a new compact support basis function (CSBF) semantic kernel utilizing wavelets for TC by SVMs.

2 Experimental results

We processed 5946 abstracts with LCSH metadata for machine learning (ML) from the Strathprints digital repository, University of Strathclyde [2]. Keywords were obtained by a WordNet-based stemmer using the controlled vocabulary of the lexical database resulting in 11586 keywords in the abstracts, and ranked according to the Jiang-Conrath distance based with the algorithm described in [3]. With altogether 176 classes, the research question was how efficiently SVM kernels can reproduce fine-grained text categories based on abstracts only. The corpus was split to 80% training data and 20% test data, without validation. Multilabel, multiclass classification problems were split into one-against-all binary problems and their micro-, macro-averaged precision and recall values plus F_1

score calculated. Only C-SVMs were benchmarked, with the C penalty parameter left at the default value of 1. The implementation used the `libsvm` library [4] with linear, polynomial and RBF kernels on vectors to study classification performance. Polynomial kernels were benchmarked at second and third degree, RBF kernels by a small value ($\gamma = 1/$size of feature set) parameter as well as relatively high ones ($\gamma = 1$ and 2). A B-spline kernel with multiple parameters was benchmarked with the length of support ranging between 2 and 10. In terms of the micro- and macroaverage F_1 measures, in three out of four cases the wavelet kernel outperformed the traditional kernels while reconstructing existing classification tags based on abstracts (1). In all, the wavelet kernel performed best in the task of reconstructing the existing classification on a deeper level from abstracts.

Table 1. Results on the StrathPrints collection

Kernel	Linear	Poly	RBF	CSBF				
Support length	-	-	-	2	4	6	8	10
Microaverage P	0.514	0.457	**0.680**	**0.516**	0.503	0.485	0.472	0.466
Macroaverage P	0.603	0.595	**0.951**	**0.611**	0.595	0.572	0.558	0.547
Microaverage R	**0.433**	0.348	0.012	**0.444**	0.441	0.439	0.435	0.433
Macroaverage R	**0.364**	0.295	0.174	**0.362**	0.360	0.361	0.357	0.358
Microaverage F_1	**0.470**	0.395	0.023	**0.478**	0.470	0.461	0.453	0.449
Macroaverage F_1	**0.454**	0.395	0.294	**0.455**	0.449	0.442	0.435	0.432

3 Acknowledgement

Research by the second and third authors was funded by the SHAMAN EU project (grant no.: 216736).

References

1. Gonçalves, M., Fox, E., Watson, L., Kipp, N.: Streams, structures, spaces, scenarios, societies (5s): A formal model for digital libraries. ACM Transactions on Information Systems **22**(2) (2004) 312
2. Dawson, A., Slevin, A.: Repository case history: University of Strathclyde Strathprints (2008)
3. Wittek, P., Darányi, S., Tan, C.: Improving text classification by a sense spectrum approach to term expansion. In: Proceedings of CoNLL-09, 13th Conference on Computational Natural Language Learning, Boulder, CO, USA (2009) 183–191
4. Chang, C.C., Lin, C.J.: LIBSVM: A library for support vector machines. (2001) http://www.csie.ntu.edu.tw/˜cjlin/libsvm.