Noname manuscript No.
(will be inserted by the editor)

Using wavelet analysis for text categorization in digital libraries: a first experiment with Strathprints

Sándor Darányi · Peter Wittek · Milena Dobreva

Received: date / Accepted: date

Abstract Digital libraries increasingly benefit from research on automated text categorization for improved access. Such research is typically carried out by using standard test collections. In this paper we present a pilot experiment of replacing such test collections by a set of 6000 objects from a real-world digital repository, indexed by Library of Congress Subject Headings, and test support vector machines in a supervised learning setting for their ability to reproduce the existing classification. To augment the standard approach, we introduce a combination of two novel elements: using functions for document content representation in Hilbert space, and adding extra semantics from lexical resources to the representation. Results suggest that wavelet-based kernels slightly outperformed traditional kernels on classification reconstruction from abstracts and vice versa from full-text documents, the latter outcome due to word sense ambiguity. The practical implementation of our methodological framework enhances the analysis and representation of specific knowledge relevant to large-scale digital collections, in this case the thematic coverage of the collections. Representation of specific *knowledge* about digital collections is one of the basic elements of the persistent archives and the less studied one (compared to representations of digital objects and collections). Our research is an initial step in this direction developing further the methodological approach and demonstrating that text categorisation can be applied to analyse the thematic coverage in digital repositories.

Keywords digital libraries · text categorization · machine learning · support vector machines · analogical information representation · wavelet analysis

Sándor Darányi
Swedish School of Library
and Information Science
University of Borås
Tel.: +46 (0)33 435 4679
E-mail: sandor.daranyi@hb.se

Peter Wittek
School of Computing
National University of Singapore
Tel.: +65 91239785
E-mail: wittek@comp.nus.edu.sg

Milena Dobreva
Centre for Digital Library Research
University of Strathclyde
Tel.:+44 (0)141 548 4753
E-mail: milena.dobreva@strath.ac.uk

1 Introduction

Evolving digital libraries (DL) [55], with new forms of digital media result in an increasing variety and number of digital objects with indexing and classification needs as part of management for access [21,43] and reuse [7], in the broader context of digital preservation and even broader context of information life cycle management [7]. To apply machine learning (ML) to one of the standard DL circulation activities, namely text categorization [48], is part of the cognitive toolbox deployed [18]. In this context, ML is extensively being experimented with in different development areas and scenarios; to name but a few, for extracting image content from figures in scientific documents for categorization [33,34], automatically assessing and characterizing resource quality for educational DL [54,5], assessing the quality of scientific conferences [37], web-based collection development [42], automated document metadata extraction by support vector machines (SVM, [24]), automatic extraction of titles from general documents [27],

information architecture [17], to remove duplicate documents [9], for collaborative filtering [59], for the automatic expansion of domain-specific lexicons by term categorization [3], for generating visual thesauri [45], or the semantic markup of documents [13]. As part of this direction of research, ML is being tested for its ability to reproduce parts of collections indexed by widespread classification schemes in a supervised learning setting, such as automatic text categorization using the Dewey Decimal Classification (DDC, [52]), or the Library of Congress Classification (LCC) from Library of Congress Subject Headings (LCSH, [20,43]).

All these examples of ongoing work may seem to address completely isolated tasks, but if we look at them from the point of view of the functional entities in a digital preservation (DP) system, as defined by the OAIS[1] reference model [28], we will be able to clearly associate them with the OAIS functional entities (see Fig. 1).

It is not surprising that most of the currently experimented categorisation tools can be applied to enhance the possibilities offered to the consumers within the Access functional entity, which, in the expanded information life cycle terminology, would correspond to use and re-use of digital objects. Here we will try to analyse in more depth how the text categorisation could be used in a real archival environment and what this use could contribute in particular to DP.

With various information retrieval (IR) and text categorization (TC, also known as automatic classification) models becoming more and more available for DLs and generating local demand for new, automated solutions [23], in this paper we test a new TC model in a real world setting for the above purpose. As TC research typically uses standard test collections of documents, we replace them by a small database, the institutional repository of the University of Strathclyde, Glasgow, called Strathprints[2], indexed by the Library of Congress Subject Headings (LCSH). As due to newly registered documents, the database keeps on expanding, a need to index digital objects based on previous practice exists, calling for the application of supervised ML. In such a scenario, part of the existing, indexed collection is used to teach the algorithm the indexing rules and another part to test whether the algorithm had learnt its homework well.

[1] The most popular standard in the area of DP is ISO 14721:2003 (Space data and information transfer systems Open archival information system Reference model), widely known as OAIS. It is a functional framework which presents the main components and the basic data flows within a digital preservation system. In this article we will use the terminology in DP as defined in OAIS.

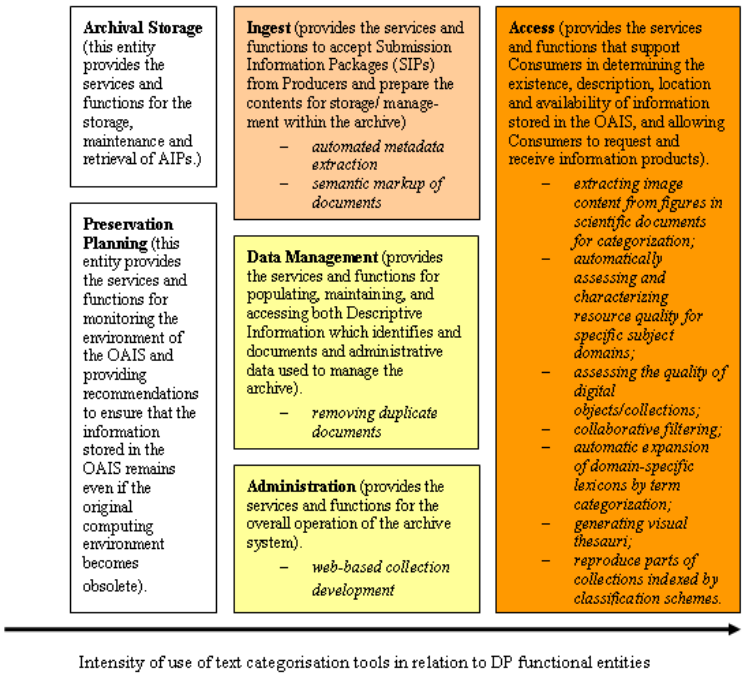
[2] `http://strathprints.strath.ac.uk/`

As part of the experimental setup, we also add two novel elements to the existing methodology. The first is a somewhat unusual model of information representation which is related to content-based image indexing because it uses wavelets for semantic content representation [53,61,56]. The second one is that the setup includes lexical resources as the interpretational context of content elements in text documents [8]. Although such a contextual representation scheme for medical literature in a DL setting had been tested [44], wavelet analysis is mostly experimented with in the image, video, and audio processing context, based on analogical information representation [33,14,31,35,39].

Our reason to test wavelets for text representation is as follows. Whereas we regard document categorization by SVM [30,50,49,4,38,6] a particular implementation of machine learning, an increasingly successful solution to the classical problem of automatic classification, we also envisage information representation by vectors, a standard point of departure for TC by SVM, a limitation of the above attempt, and combine the former with semantic content representation in Hilbert space instead of Euclidean space. In this new approach, instead of term and document vectors, term and document functions are used to represent the semantic content of digital objects, with the advantage that functions, having more parameters than vectors, can host more semantic content in a comprehensive description than vector space based methods. Given these considerations, the research problem of this article is this: how efficient is our algorithm in reproducing an existing classification, i.e. how reliable would it be to develop a scaled up TC application based on this model? As the results will demonstrate, the proposed approach has several theoretical and practical points to make, including bridging the gap between computational linguistics, signal processing, text categorization and digital libraries, plus introducing a robust, innovative way of content processing with the ability to integrate different approaches of text and image indexing.

This paper is organized as follows: in Section 2, we briefly introduce text categorization by machine learning. In Section 3, information representation for text categorization is discussed in more detail, with stress on modelling term dependence by semantic kernels in vector and function space and the role of semantic ordering for term expansion. In Section 4, the material and the method are outlined, with consecutive results in Section 5. In Section 6 we consider the relevance of the proposed approach for OAIS and persistent archives (PA), and we also hint at directions of further research.

Fig. 1 Some possible uses of text categorisation tools in DP



2 Text categorization by machine learning

Text categorization is the task of assigning unlabeled documents into predefined categories. Given a collection of $\{d_1, d_2, \ldots, d_N\}$ documents, and a $C = \{c_1, c_2, \ldots, c_{|C|}\}$ set of predefined categories, the task is, for each document d_j $(j \in \{1, 2, \ldots, N\})$, to assign a decision to file d_j under c_i or a decision not to file d_j under c_i $(c_i \in C)$ by virtue of a function Φ, where the function Φ is also referred to as the classifier, or model, or hypothesis, or rule. Supervised text classification is a machine learning technique for creating the function Φ from training data. The training data consist of pairs of input documents, and desired outputs (i.e., classes).

Support vector machines have been found the most effective by several authors [30]. The proposed semantic text classification method is grounded in the kernel methods underlying support vector machines.

A support vector machine is a kind of supervised learning algorithm. In its simplest, linear form, a support vector machine is a hyperplane that separates a set of positive examples from a set of negative examples with maximum margin [49]. The strength of kernel methods is that they allow a mapping $\phi(.)$ of $\mathbf{x}$ to a higher dimensional space. In the dual formulation of the mathematical programming problem, only the kernel matrix $K(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i)'\phi(\mathbf{x}_i)$ is needed in the calculations.

The simplest linear kernel $(K(\mathbf{x}, \mathbf{y}) = (\mathbf{x}, \mathbf{y}))$ essentially calculates the co-occurances of index terms in two documents. A second or third order polynomial kernel $(K(\mathbf{x}, \mathbf{y}) = ((\mathbf{x}, \mathbf{y}) + 1)^d)$ also regards multiple word co-occurences.

Moreover, any continuous symmetric function $K(\mathbf{x}, \mathbf{y})$ in $L_2 \times L_2$ may be used as an admissible kernel, as long as satisfies a weak form of Mercer's condition [51]: for all $g \in L_2(\mathbf{R}^M)$, $\int\int K(\mathbf{x}, \mathbf{y})g(\mathbf{x})g(\mathbf{y}) \geq 0$, .

This latter approach gave rise to radial basis function kernels ($\exp(-\gamma|\mathbf{x}-\mathbf{y}|^2)$)), where smaller γ values will result in smoother, roughly oval shaped decision boundaries, while larger γ values will give a more irregular decision boundary.

A recent work explored combining wavelets with SVMs proving that these kernels satisfy the above admissibility condition [61]. Two admissible wavelet kernels were studied, $K(\mathbf{x}, \mathbf{y}) = \prod_{i=1}^{M} \psi(\frac{x_i - k}{j})\psi(\frac{y_i - k}{j})$ and the translation-invariant kernel defined as $K(\mathbf{x}, \mathbf{y}) = \prod_{i=1}^{M} \psi(\frac{x_i - y_i}{j})$. The kernel performed well in support vector regression and classification.

3 Information representation for TC

Intuitively, if two documents address similar topics, it is highly possible that they share lots of substantive terms. After having removed the stopwords, such as articles and prepositions, and stemmed the rest, the stemmed terms construct a vector representation for each text document. Let $\mathbf{a}_j$ be a document vector in the vector space model, that is, $\mathbf{a}_j = \sum_{k=1}^{M} a_{kj}\mathbf{e}_k$, where M is the number of index terms, a_{kj} is some weighting (e.g., term frequency), and $\mathbf{e}_k$ is a basis vector of the M-dimensional Euclidean space. This representation is also referred to as the bag-of-words (BOW) model.

Given this representation, semantic relatedness of a pair of text fragments is computed as the cosine similarity of their corresponding term vectors which is defined as:

$$S(\mathbf{a}_i, \mathbf{a}_j) = \frac{\mathbf{a}_i \mathbf{a}_j}{|\mathbf{a}_i||\mathbf{a}_j|}. \quad (1)$$

3.1 Linear Semantic Kernels

One enrichment strategy is to use a semantic smoothing kernel while calculating the similarity between two documents. Any linear kernel for texts is characterized by $K(\mathbf{a}_i, \mathbf{a}_j) = \mathbf{a}_i' S' S \mathbf{a}_j$, where S is an appropriately shaped matrix commonly referred to as the semantic smoothing matrix [50,49,4,38,6]. The presence of S changes the orthogonality of the vector space model, as this mapping should introduce term dependence. A recent attempt tried to manually construct S with the help of a lexical resource [50]. The entries in the symmetric matrix S express the semantic similarity between the terms i and j. Entries in this matrix are inversely proportional to the length of the WordNet [19] hierarchy path linking the two terms. The performance, measured over the 20NewsGroups corpus, showed an improvement of 2 % over the the basic vector space method. Moreover, the semantic matrix S is almost fully dense, hence computational issues arise. In a similar construction, [6] defined the matrix entries as weights of superconcepts of the two terms in the WordNet hierarchy. Focusing on special subcategories of Reuters-21578 and on the TREC Question Answering Dataset, they showed consistent improvement over the baseline. As [38] pointed out, polysemy will remain a problem in semantic smoothing kernels. A more complex way of calculating the semantic similarity as the matrix entries was also proposed [4]. For a more general discussion on semantic similarity see Section 3.3.1.

An early attempt to overcome the untenable orthogonality assumption of the vector space model was proposed under the name of generalized vector space model [58]. The article which proposed the model did not provide empirical results, and since then the model has been regarded of large theoretical importance with less impact on actual applications. The model takes a distributional approach, focusing on term co-occurrences. The underlying assumption is that term correlations are captured by the co-occurrence information. That is, two terms are semantically related if they co-occur often in the same documents. By eliminating orthogonality, documents can be seen as similar even if they do not share any terms. The term co-occurrence matrix is AA', hence the model takes A' as the semantic similarity matrix S. A major drawback of the generalized vector space model is that it replaces the orthogonality assumption with another questionable assumption. The computational needs are tremendous too, if the dimensions of A are considered. Moreover, the co-occurrence matrix is not sparse anymore.

Latent semantic indexing (or latent semantic analysis) was another attempt to bring more linguistic and psychological aspects to language processing via a kernel. Conceptually, latent semantic indexing is similar to the generalized vector space model, it measures semantic information through co-occurrence analysis in the corpus. From the algorithmic perspective it is an enormous problem that textual data have a large number of relevant features. This results in huge computational needs and the classification models may overfit the data. The number of features can be reduced by multivariate feature extraction methods. In latent semantic indexing, the dimension of the vector space is reduced by singular value decomposition [16].

Using rank reduction, terms that occur together very often in the same documents are merged into a single dimension of the feature space. The dimensions of the reduced space correspond to the axes of greatest variance. For latent semantic indexing, by dual representation the kernel matrix is $K = V\Sigma_k^2 V'$, where Σ_k is a diagonal matrix containing the k largest singular values of the singular value decomposition of the vector space, and V holds the right singular vectors of the decomposition. The new kernel matrix can be obtained directly from K by applying an eigenvalue decomposition of K [12]. The computational complexity of performing an eigenvalue decomposition on the kernel matrix is a major drawback of latent semantic indexing.

3.2 Text Representation Enrichment Strategies by Term Expansion

In order to eliminate the bottleneck of the traditional BOW representation, previous approaches in term ex-

pansion enriched this convention by external lexical resources such as WordNet.

As a first step, these methods generate new features for each document in the dataset. These new features can be synonyms or homonyms of document terms as in [26,47], or expanded features for terms, sentences and documents as in [22], or term context information for word sense disambiguation such as topic signatures [2, 1].

Then, the generated new features replace the old ones or are appended to the document representation, and construct a new vector representation $\hat{\mathbf{a}}_i$ for each text document. The similarity measure of document pairs is defined as:

$$S(\hat{\mathbf{a}}_i, \hat{\mathbf{a}_j}) = \frac{\hat{\mathbf{a}}_i \hat{\mathbf{a}_j}}{|\hat{\mathbf{a}}_i||\hat{\mathbf{a}_j}|}. \quad (2)$$

3.3 Proposed Framework

The basic assumption of our framework is that terms can be arranged in an order such that consecutive terms are semantically related. Hence each term acquires a unique position, and this position ties the term to its semantically related neighbors. However, given a BOW representation with a cosine similarity measure, this position would not improve classification performance. Therefore we suggest to associate a mathematical function with each term, thus mapping terms and documents to the L_2 space, and using the inner product of this space to express similarity. The choice of function will determine to which extent neighboring terms, i.e., the enriching terms, are considered in calculating the similarity between two documents. This section first introduces an algorithm that produces the aforementioned semantic order, then the semantic kernels in the L_2 space are discussed.

3.3.1 An Algorithm for a Semantic Ordering of Terms

The proposed kernels assume that there is a semantic order between terms. Let V denote a set of terms $\{t_1, t_2, \ldots, t_n\}$ and let $d(t_i, t_j)$ denote the semantic distance between the terms t_i and t_j. The initial order of the terms is not relevant, though it is assumed to be alphabetic. Let $G = (V, E)$ denote a weighted undirected graph, where the weights in the set E are defined by the distances between the terms.

Various lexical resource-based [8] and distributional measures [40] have been proposed to measure semantic relatedness and distance between terms. Terms can be corpus- or genre-specific. Manually constructed general-purpose lexical resources include many usages that are infrequent in a particular corpus or genre of documents. For example, one of the 8 senses of *company* in WordNet is a *visitor/visitant*, which is a hyponym of *person* [32]. This sense of the term is practically never used in newspaper articles, hence distributional attributes should be taken into consideration. Composite measures that combine the advantages of both approaches have also been developed [46,29]. This paper relies on the Jiang-Conrath composite measure [29], which has been shown to be superior to other measures [8], and we also found that this measure works the best for the purpose. The Jiang-Conrath metric measures the distance between two senses by using the hierarchy of WordNet. By denoting the lowest super-ordinate of two senses s_1 and s_2 in the hierarchy with LSuper(s_1,s_2), the metric is calculated as follows:

$$d(s_1, s_2) = \mathrm{IC}(s_1) + \mathrm{IC}(s_2) - 2\mathrm{IC}(\mathrm{LSuper}(s_1, s_2)),$$

where IC(s) is the information content of a sense s based on a corpus. Distance between given two terms is calculated according to the following equation: $\mathrm{d}(t_1, t_2) = \max_{s_1 \in \mathrm{sen}(t_1), s_2 \in \mathrm{sen}(t_2)} d(s_1, s_2)$, where t_1 and t_2 are two terms, and sen(t_i) is the set of senses of t_i. The distance between two terms is usually defined as the minimum of the sense distances. We chose maximum because it ensures that only closely related terms will be placed to adjacent positions by the algorithm below.

Finding a semantic ordering of terms can be translated to a graph problem: a minimum-weight Hamiltonian path G' of G gives the ordering by reading the nodes from one of the paths to the other. G is a complete graph, therefore such a path always exists, but finding it is an NP-complete problem. The following greedy algorithm is similar to the nearest neighbor heuristic for the solution of the traveling salesman problem. It creates a graph $G' = (V', E')$, where $V' = V$ and $E' \subset E$. This G' graph is a spanning tree of G in which the maximum degree of a node is two, that is, the minimum spanning tree is a path between two nodes.

Step 1 Find the term at the highest stage of the hierarchy in a lexical resource.

$t_s = \mathrm{argmin}_{t_i \in V} \mathrm{depth}(t_i)$.

This seed term is the first element of V', $V' = \{t_s\}$. Remove it from the set V:

$V := V \backslash \{t_s\}$.

Using WordNet, this seed term is *entity*, if the vocabulary of the text collection contains it.

Step 2 Let t_l denote the leftmost term of the ordering and t_r the rightmost one. Find the next two elements of the ordering:

$t'_l = \text{argmin}_{t_i \in V} d(t_i, t_l)$,

$t'_r = \text{argmin}_{t_i \in V \setminus \{t'_l\}} d(t_i, t_r)$.

Step 3 If $d(t_l, t'_l) < d(t_r, t'_r)$ then add t'_l to V', $E' := E' \cup \{e(t_l, t'_l)\}$, and $V := V \setminus \{t'_l\}$. Else add t'_r to V', $E' := E' \cup \{e(t_r, t'_r)\}$ and $V := V \setminus \{t'_r\}$.

Step 4 Repeat from *Step 2* until $V = \emptyset$.

The above algorithm can be thought of as a modified Prim's algorithm [11], but it does not find the optimal minimum-weight spanning tree.

3.3.2 Semantic Kernels in the L_2 Space

The L_2 space shares resemblance with a real vector space. Square-integrable functions replace real-valued vectors, and the dot product is replaced by the following inner product: $(f_i, f_j) = \int f_i f_j dx$, for some f_i, f_j in the given L_2 space.

Lately, Hoenkamp has also pointed out that the L_2 space can be used for information retrieval when he introduced a Haar basis for the document space [25]. He utilized a signal processing framework within the context of latent semantic indexing. In order to apply an L_2 representation for text classification, the problem is approached from a different angle than by Hoenkamp, taking discounting expansion terms as our point of departure.

Assigning a basis function $w(x-k)$ to the term in the kth position in a semantic order, a document j can be expressed as follows:

$$f_j(x) = \sum_{k=1}^{M} a_{kj} w(x-k), \tag{3}$$

where x is in $[1, M]$, and it is the variable of integration in calculating the inner product of the L_2; x can be regarded as a "dummy" variable carrying no meaning in itself. The above formula will be referred to as a document function. If the support of the basis function is compact, the computational cost of the kernel is similar to that of a linear kernel. Furthermore, with compactly supported B-splines as the basis, the integral can be calculated expicitely due to the convolution property of B-splines, making the eventual computations easier.

The inner product of the $L_2[1, M]$ space is applied to express similarity between two documents in similar vein as the dot product does in a real-valued vector space:

$$(f_i, f_j) = \int_{[1,M]} f_i(x) f_j(x) dx, \tag{4}$$

where f_i and f_j are the representations of the documents in the L_2 space ($f_i, f_j \in L_2([1, M])$).

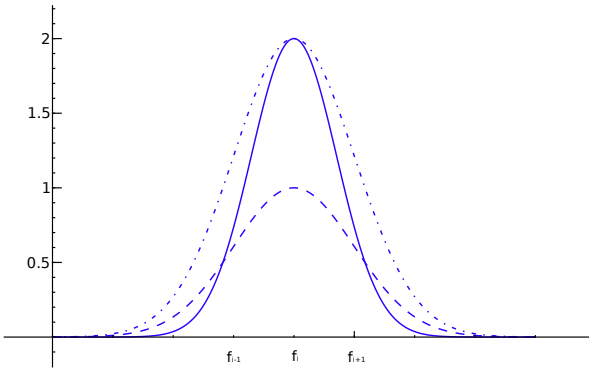


Fig. 2 Two documents with matching term *brand name*. Dotted line: Document-1. Dashed line: Document-2. Solid line: Their product.

With the above formula, a matching term in two documents will be counted to its full term frequency or tfidf score, while semantically related terms will be counted less and less according their semantic similarity to the matching term. Assuming that the terms *brand*, *brand name*, and *trade name* follow each other in the semantic order, consider the following example. The first document has the term *brand name*, and so does the second document. In Figure 2, it can be seen brand name is counted the same way as it would be in a BOW model with its full term frequency score, brand and trade name are counted to a lesser extent, while other related terms are considered even less.

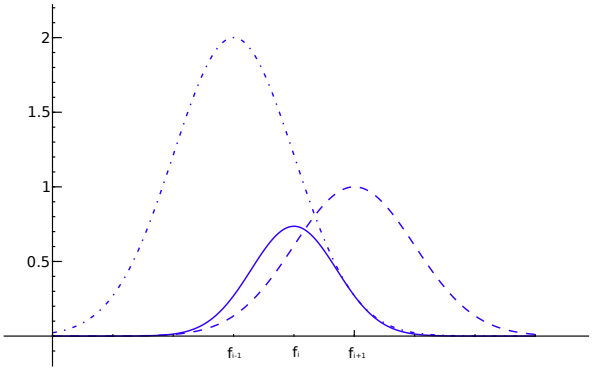


Fig. 3 Two documents with no matching term but with related terms *brand* and *trade name*. Dotted line: Document-1. Dashed line: Document-2. Solid line: Their product.

Now if the two documents do not share the exact term, only related terms occur, for instance, *trade name*

Table 1 Size of training and test sets

	Abstracts	Full text
Training	4757	5495
Test	1189	1374

and *brand*, respectively, then the term *brand name*, placed between *trade name* and *brand* in the s semantic order, will be considered only to some extent for the calculation of similarity (see Figure 3).

4 Case study

The kernel outlined in this paper has been benchmarked on standard test collection and has been shown to be superior to certain semantic kernels [57]. In this paper we are interested how it performs on a real-world digital library. Strathprints is "an institutional eprint repository for making research papers and other scholarly publications widely available on the Internet" at the University of Strathclyde, UK [15], its hosting and technical support provided by the Department of Computer and Information Sciences (CIS). Eprints and usage statistics software have been installed, configured and managed by the Centre for Digital Library Research (CDLR) at the same university. Its digital objects are indexed by the LCSH classification scheme. From an ML point of view, this is a multilabel scenario where an instance of the collection may belong to several categories. Out of 6869 records, the size of the database on 13th June 2009, we downloaded and processed 5946 abstracts, the rest being records without abstracts or duplicates. With 14 ppt files removed, only abstracts in doc, html and pdf format were indexed by their LCSH tags. Keywords were obtained by a WordNet-based stemmer using the controlled vocabulary of the lexical database resulting in 21718 keywords in the full-text documents and 11586 in the abstracts. Keywords were ranked according to the JCN distance based with the algorithm described in Section 3.3.1.

With 20 top classes and altogether 176 classes, the immediate research question was how efficiently SVM kernels can reproduce different levels of increasingly fine-grained text categories based on fulltext vs. abstracts only. The corpus was split to 80% training data and 20% test data; validation was not applied (Table 1).

5 Results

We split the multilabel, multiclass classification problems into one-against-all binary problems and calculated the micro-, and macro-averaged precision and recall values, and then their average, the F_1 score [60]. For both sets of measurements, this was the most important observation parameter.Only C-SVMs were benchmarked, with the C penalty parameter left at the default value of 1. The implementation used the `libsvm` library [10]. The results consisted of two parts:

(a) SVM and kernels on Strathprints. We used the widespread linear, polynomial and RBF kernels on vectors to study classification performance. Polynomial kernels were benchmarked at second and third degree, RBF kernels were benchmarked with a small value (γ = 1/size of feature set) parameter as well as relatively high one (γ = 1 and 2). Because of the size of the fulltext-based vocabulary, the implementation of a linear semantic matrix (Section 3.1) was prohibitive.

(b) Wavelets on Strathprints. 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 (Tables 2, 3, 6 and 7). In a repeated experiment based on fulltext documents, the

Table 2 Results on abstracts with traditional kernels, top-level categories

	Linear	Polynomial	RBF
Microaverage P	0.744	0.722	**0.880**
Macroaverage P	0.685	0.617	**0.976**
Microaverage R	**0.686**	0.623	0.017
Macroaverage R	**0.464**	0.398	0.064
Microaverage F_1	**0.714**	0.669	0.033
Macroaverage F_1	**0.553**	0.484	0.121

Table 3 Results on abstracts with traditional kernels, refined categories

	Linear	Polynomial	RBF
Microaverage P	0.514	0.457	**0.680**
Macroaverage P	0.603	0.595	**0.951**
Microaverage R	**0.433**	0.348	0.012
Macroaverage R	**0.364**	0.295	0.174
Microaverage F_1	**0.470**	0.395	0.023
Macroaverage F_1	**0.454**	0.395	0.294

Table 4 Results on abstracts with L_2 kernels, top-level categories

Support length	2	4	6	8	10
Microaverage P	**0.732**	0.713	0.704	0.696	0.695
Macroaverage P	**0.685**	0.660	0.657	0.647	0.642
Microaverage R	**0.680**	0.672	0.671	0.668	0.666
Macroaverage R	**0.469**	0.462	0.467	0.462	0.459
Microaverage F_1	**0.705**	0.692	0.687	0.682	0.680
Macroaverage F_1	**0.557**	0.544	0.546	0.539	0.536

Table 5 Results on abstracts with L_2 kernels, refined categories

Support length	2	4	6	8	10
Microaverage P	**0.516**	0.503	0.485	0.472	0.466
Macroaverage P	**0.611**	0.595	0.572	0.558	0.547
Microaverage R	**0.444**	0.441	0.439	0.435	0.433
Macroaverage R	**0.362**	0.360	0.361	0.357	0.358
Microaverage F_1	**0.478**	0.470	0.461	0.453	0.449
Macroaverage F_1	**0.455**	0.449	0.442	0.435	0.432

Table 6 Results on full texts with traditional kernels, top-level categories

	Linear	Polynomial	RBF
Microaverage P	0.728	0.591	**0.949**
Macroaverage P	0.555	0.302	**0.992**
Microaverage R	**0.738**	0.694	0.568
Macroaverage R	**0.612**	0.564	0.289
Microaverage F_1	**0.733**	0.638	0.711
Macroaverage F_1	**0.582**	0.393	0.448

Table 7 Results on full texts with traditional kernels, refined categories

	Linear	Polynomial	RBF
Microaverage P	0.487	0.335	**0.880**
Macroaverage P	0.571	0.366	**0.980**
Microaverage R	**0.523**	0.514	0.230
Macroaverage R	0.576	**0.586**	0.440
Microaverage F_1	**0.505**	0.406	0.370
Macroaverage F_1	0.573	0.451	**0.610**

Table 8 Results on full texts with L_2 kernels, top-level categories

Support length	2	4	6	8	10
Microaverage P	0.714	0.724	**0.728**	0.712	0.695
Macroaverage P	0.490	0.482	**0.538**	0.533	0.532
Microaverage R	**0.738**	0.731	0.728	0.713	0.666
Macroaverage R	**0.612**	0.588	0.584	0.574	0.459
Microaverage F_1	0.726	0.728	**0.728**	0.712	0.680
Macroaverage F_1	0.545	0.530	**0.560**	0.553	0.496

opposite was the case, even if the difference in favour of traditional kernels was a minor one (Tables 4, 5, 8, and 9). We argue that this goes back to word sense disambiguation [36]: abstracts had less ambigous terms than complete documents. In all, the wavelet kernel performed best in the task of reconstructing the existing classification on a deeper level from abstracts.

6 Conclusion and future research

We feel that the greater information representation capacity functions as mathematical objects hold over vectors has a utilization potential for OAIS Access on the one hand. Typically, vectors can represent information by norm and angle only, whereas functions, depending on their types, have e.g. wavelength, amplitude, frequency, driven by many extra parameters which can be assigned specific meanings, plus the opportunity of a physical implementation (such as using spectrography for displaying the conceptual spectrum [56]). Further, their use brings DLs closer to e-science by applying physics as a metaphor to the modelling of word semantics. For example our ongoing experiments to use vector valued functions to characterize evolving semantics in databases relate this Hilbert space-based TC implementation to phase space TC and IR, with all their implications for advanced access in an OAIS setting. On the other hand, in terms of permanent archives (PA, [41]), this will give one room to experiment with self-describing objects and knowledge on collection level in novel ways.

Table 9 Results on full texts with L_2 kernels, refined categories

Support length	2	4	6	8	10
Microaverage P	0.462	0.474	**0.488**	0.480	0.463
Macroaverage P	0.556	0.558	**0.573**	0.569	0.557
Microaverage R	**0.505**	0.501	0.493	0.485	0.478
Macroaverage R	**0.561**	0.499	0.482	0.473	0.463
Microaverage F_1	0.482	0.488	**0.491**	0.483	0.471
Macroaverage F_1	**0.558**	0.523	0.523	0.521	0.510

The overall good quality of the first test encourages further research into the following problems:

- In a digital library circulation process, how to recognize and separate misclassified items from accurately categorized ones while moving from a supervised machine learning environment to an unsupervised one;
- How to scale up text categorization quality assessment in a web harvesting environment;
- How to best exploit the increased information representation capability of functions for advanced access to semantic content.

Acknowledgements For the first and the third author, funding for this research was provided by the SHAMAN EU project. The authors express their gratitude to Dennis Nicholson and Emma McCulloch (Univeristy of Strathclyde) for permission to work with Strathprints and consultations.

References

1. Agirre, E., Alfonseca, E., de Lacalle, O.: Approximating hierarchy-based similarity for WordNet nominal synsets using topic signatures. In: Proceedings of GWC-04, 2nd Global WordNet Conference, pp. 15–22. Brno, Czech Republic (2004)
2. Agirre, E., De Lacalle, O.: Clustering WordNet word senses. In: Proceedings of RANLP-03, 4th International Conference

on Recent Advances in Natural Language Processing, pp. 121–130. John Benjamins Publishing, Amsterdam, Netherlands, Borovets, Bulgaria (2003)
3. Avancini, H., Lavelli, A., Sebastiani, F., Zanoli, R.: Automatic expansion of domain-specific lexicons by term categorization. ACM Transactions on Speech and Language Processing **3**(1), 1–30 (2006)
4. Basili, R., Cammisa, M., Moschitti, A.: Effective use of WordNet semantics via kernel-based learning. In: Proceedings of CoNLL-05, 9th Conference on Computational Natural Language Learning, pp. 1–8. ACL, Morristown, NJ, USA, Ann Arbor, MI, USA (2005)
5. Bethard, S., Wetzer, P., Butcher, K., Martin, J., Sumner, T.: Automatically characterizing resource quality for educational digital libraries. In: Proceedings of JCDL-09, 9th Joint International Conference on Digital Libraries, pp. 221–230. ACM Press, New York, NY, USA, Austin, TX, USA (2009)
6. Bloehdorn, S., Basili, R., Cammisa, M., Moschitti, A.: Semantic kernels for text classification based on topological measures of feature similarity. In: Proceedings of ICDM-06, 6th IEEE International Conference on Data Mining. Hong Kong (2006)
7. Brocks, H., Kranstedt, A., Jäschke, G., Hemmje, M.: Modeling context for digital preservation. In: N. Nguyen, E. Szczerbicki (eds.) Smart Information and Knowledge Management: Advances, Challenges, and Critical Issues. Springer (2009)
8. Budanitsky, A., Hirst, G.: Evaluating WordNet-based measures of lexical semantic relatedness. Computational Linguistics **32**(1), 13–47 (2006)
9. de Carvalho, M., Gonçalves, M., Laender, A., da Silva, A.: Learning to deduplicate. In: Proceedings of JCDL-06, 6th ACM/IEEE-CS Joint Conference on Digital Libraries, pp. 41–50. ACM Press, New York, NY, USA, Chapel Hill, NC, USA (2006)
10. Chang, C.C., Lin, C.J.: LIBSVM: A library for support vector machines (2001). `http://www.csie.ntu.edu.tw/\~{}cjlin/libsvm`
11. Cormen, T., Leiserson, C., Rivest, R.: Introduction to algorithms. MIT Press, Cambridge, MA, USA (2001)
12. Cristianini, N., Shawe-Taylor, J., Lodhi, H.: Latent semantic kernels. Journal of Intelligent Information Systems **18**(2), 127–152 (2002)
13. Cui, H.: An application for semantic markup of biodiversity documents. In: Proceedings of JCDL-08, 8th ACM/IEEE-CS Joint Conference on Digital Libraries, pp. 421–421. ACM Press, New York, NY, USA, Pittsburgh, PA, USA (2008)
14. Datta, R., Joshi, D., Li, J., Wang, J.: Image retrieval: Ideas, influences, and trends of the new age. ACM Computing Surveys **40**(2), 1–60 (2008)
15. Dawson, A., Slevin, A.: Repository case history: University of Strathclyde Strathprints (2008). URL `http://www.rsp.ac.uk/repos/casestudies/pdfs/strathclyde.pdf`
16. Deerwester, S., Dumais, S., Furnas, G., Landauer, T., Harshman, R.: Indexing by latent semantic analysis. Journal of the American Society for Information Science **41**(6), 391–407 (1990)
17. Efron, M., Elsas, J., Marchionini, G., Zhang, J.: Machine learning for information architecture in a large governmental web site. In: Proceedings of JCDL-04, 4th ACM/IEEE-CS Joint Conference on Digital Libraries, pp. 151–159. ACM Press, New York, NY, USA, Tuscon, AZ, USA (2004)
18. Esposito, F., Malerba, D., Semeraro, G., Fanizzi, N., Ferilli, S.: Adding machine learning and knowledge intensive techniques to a digital library service. International Journal on Digital Libraries **2**(1), 3–19 (1998)
19. Fellbaum, C.: WordNet: An Electronic Lexical Database. MIT Press, Cambridge, MA, USA (1998)
20. Frank, E., Paynter, G.: Predicting library of congress classifications from library of congress subject headings. Journal of the American Society for Information Science and Technology **55**(3), 214–227 (2004)
21. Fuhr, N., Tsakonas, G., Aalberg, T., Agosti, M., Hansen, P., Kapidakis, S., Klas, C., Kovács, L., Landoni, M., Micsik, A.: Evaluation of digital libraries. International Journal on Digital Libraries **8**(1), 21–38 (2007)
22. Gabrilovich, E., Markovitch, S.: Feature generation for text categorization using world knowledge. In: Proceedings of IJCAI-05, 19th International Joint Conference on Artificial Intelligence, vol. 19. Lawrence Erlbaum Associates, Edinburgh, UK (2005)
23. Hagedorn, K., Chapman, S., Newman, D.: Enhancing search and browse using automated clustering of subject metadata. D-Lib Magazine **13**(7/8), 1082–9873 (2007)
24. Han, H., Giles, C., Manavoglu, E., Zha, H., Zhang, Z., Fox, E.: Automatic document metadata extraction using support vector machines. In: Proceedings of JCDL-03, 3rd ACM/IEEE-CS Joint Conference on Digital Libraries, pp. 37–48. IEEE Computer Society Press, Los Alamitos, CA, USA, Houston, TX, USA (2003)
25. Hoenkamp, E.: Unitary operators on the document space. Journal of the American Society for Information Science and Technology **54**(4), 314–320 (2003)
26. Hotho, A., Staab, S., Stumme, G.: WordNet improves text document clustering. In: Proceedings of SIGIR-03, 26th International Conference on Research and Development in Information Retrieval. ACM Press, New York, NY, USA, Toronto, Canada (2003)
27. Hu, Y., Li, H., Cao, Y., Meyerzon, D., Zheng, Q.: Automatic extraction of titles from general documents using machine learning. In: Proceedings of JCDL-05, 5th ACM/IEEE-CS Joint Conference on Digital Libraries, pp. 145–154. ACM Press, New York, NY, USA, Denver, CO, USA (2005)
28. ISO 14721: Reference model for an Open Archival Information System (OAIS) fCCSDS 650.0-B-1 Blue book
29. Jiang, J., Conrath, D.: Semantic similarity based on corpus statistics and lexical taxonomy. In: Proceedings of ROCLING-97, International Conference on Research in Computational Linguistics, pp. 19–33. Taipei, Taiwan (1997)
30. Joachims, T.: Text categorization with support vector machines: Learning with many relevant features. In: Proceedings of ECML-98, 10th European Conference on Machine Learning, pp. 137–142. Springer-Verlag, London, UK, Chemnitz, Germany (1998)
31. Li, T., Ogihara, M., Li, Q.: A comparative study on content-based music genre classification. In: Proceedings of SIGIR-03, 26th International Conference on Research and Development in Information Retrieval, pp. 282–289. ACM Press, New York, NY, USA, Toronto, ON, Canada (2003)
32. Lin, D.: Automatic retrieval and clustering of similar words. In: Proceedings of ACL-98, 36th Annual Meeting of Association for Computational Lingiustics, vol. 36, pp. 768–774. ACL, Morristown, NJ, USA, Montréal, Québec, Canada (1998)
33. Lu, X., Mitra, P., Wang, J., Giles, C.: Automatic categorization of figures in scientific documents. In: Proceedings of JCDL-06, 6th ACM/IEEE-CS Joint Conference on Digital Libraries, pp. 129–138. ACM Press, New York, NY, USA, Chapel Hill, NC, USA (2006)
34. Lu, X., Wang, J., Mitra, P., Giles, C.: Deriving knowledge from figures for digital libraries. In: Proceedings of WWW-07, 16th International Conference on World Wide Web, pp. 1229–1230. ACM Press, New York, NY, USA, Banff, AB, Canada (2007)

35. Lyu, M., Yau, E., Sze, S.: A multilingual, multimodal digital video library system. In: Proceedings of JCDL-02, 2nd ACM/IEEE-CS Joint Conference on Digital Libraries, pp. 145–153. ACM Press, New York, NY, USA, Portland, OR, USA (2002)
36. Manning, C., Schütze, H.: Foundations of Statistical Natural Language Processing. MIT Press, Cambridge, MA, USA (1999)
37. Martins, W., Gonçalves, M., Laender, A., Pappa, G.: Learning to assess the quality of scientific conferences: a case study in computer science. In: Proceedings of JCDL-09, 9th Joint International Conference on Digital Libraries, pp. 193–202. ACM Press, New York, NY, USA, Austin, TX, USA (2009)
38. Mavroeidis, D., Tsatsaronis, G., Vazirgiannis, M., Theobald, M., Weikum, G.: Word sense disambiguation for exploiting hierarchical thesauri in text classification. In: Proceedings of PKDD-05, 9th European Conference on the Principles of Data Mining and Knowledge Discovery, pp. 181–192. Springer, Porto, Portugal (2005)
39. Miller, N., Wong, P., Brewster, M., Foote, H.: TOPIC ISLANDS–a wavelet-based text visualization system. In: Proceedings of InfoVis-98, IEEE Symposium on Information Visualization, pp. 189–196. IEEE Computer Society Press, Los Alamitos, CA, USA, Research Triangle Park, NC, USA (1998)
40. Mohammad, S., Hirst, G.: Distributional measures as proxies for semantic relatedness (2005). Submitted for publication
41. Moore, R., Rajasekar, A., Baru, C., Ludaescher, B., Gupta, A., Marciano, R.: Persistent archives (2005)
42. Pant, G., Tsioutsiouliklis, K., Johnson, J., Giles, C.: Panorama: extending digital libraries with topical crawlers. In: Proceedings of JCDL-04, 4th ACM/IEEE-CS Joint Conference on Digital Libraries, pp. 142–150. ACM Press, New York, NY, USA, Tuscon, AZ, USA (2004)
43. Paynter, G.: Developing practical automatic metadata assignment and evaluation tools for internet resources. In: Proceedings of JCDL-05, 5th ACM/IEEE-CS Joint Conference on Digital Libraries, pp. 291–300. ACM Press, New York, NY, USA, Denver, CO, USA (2005)
44. Purcell, G., Rennels, G., Shortliffe, E.: Development and evaluation of a context-based document representation for searching the medical literature. International Journal on Digital Libraries **1**(3), 288–296 (1997)
45. Ramsey, M., Chen, H., Zhu, B., Schatz, B.: A collection of visual thesauri for browsing large collections of geographic images. Journal of the American Society for Information Science **50**(9), 826–834 (1999)
46. Resnik, P.: Using information content to evaluate semantic similarity in a taxonomy. In: Proceedings of IJCAI-95, 14th International Joint Conference on Artificial Intelligence, vol. 1, pp. 448–453. Montréal, Québec, Canada (1995)
47. Rodriguez, M., Hidalgo, J.: Using WordNet to complement training information in text categorization. In: Procedings of RANLP-97, 2nd International Conference on Recent Advances in Natural Language Processing. John Benjamins Publishing, Amsterdam, Netherlands (1997)
48. Sebastiani, F.: Text categorization. In: A. Zanasi (ed.) Text Mining and its Applications, pp. 109–129. WIT Press, Southampton, UK (2005)
49. Shawe-Taylor, J., Cristianini, N.: Kernel Methods for Pattern Analysis. Cambridge University Press, New York, NY, USA (2004)
50. Siolas, G., d'Alché Buc, F.: Support vector machines based on a semantic kernel for text categorization. In: Proceedings of IJCNN-00, IEEE International Joint Conference on Neural Networks. IEEE Computer Society Press, Los Alamitos, CA, USA, Austin, TX, USA (2000)
51. Smola, A., Schölkopf, B., Müller, K.: The connection between regularization operators and support vector kernels. Neural Networks **11**(4), 637–649 (1998)
52. Wang, J.: An extensive study on automated dewey decimal classification. Journal of American Society for Information Science and Technology **60**(11) (2009)
53. Wang, J., Wiederhold, G., Firschein, O., Xin Wei, S.: Content-based image indexing and searching using Daubechies' wavelets. International Journal on Digital Libraries **1**(4), 311–328 (1998)
54. Wetzler, P., Bethard, S., Butcher, K., Martin, J., Sumner, T.: Automatically assessing resource quality for educational digital libraries. In: Proceedings of WICOW-09, 3rd Workshop on Information Credibility on the Web, pp. 3–10. ACM Press, New York, NY, USA, Madrid, Spain (2009)
55. Wilson, B.: A special issue on digital library evolution. D-Lib Magazine **12**(3) (2006)
56. 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, pp. 183–191. Boulder, CO, USA (2009)
57. Wittek, P., Tan, C.: Compactly supported basis functions as support vector kernels for classification. Transactions on Pattern Analysis and Machine Intelligence (2011)
58. Wong, S., Ziarko, W., Wong, P.: Generalized vector space model in information retrieval. In: Proceedings of SIGIR-85, 8th International Conference on Research and Development in Information Retrieval, pp. 18–25. ACM Press, New York, NY, USA, Montréal, Québec, Canada (1985)
59. Xia, Z., Dong, Y., Xing, G.: Support vector machines for collaborative filtering. In: Proceedings of ACMSE-06, 44th Annual Southeast Regional Conference, pp. 169–174. ACM Press, New York, NY, USA, Melbourne, FL, USA (2006)
60. Yang, Y.: An evaluation of statistical approaches to text categorization. Information Retrieval **1**(1), 69–90 (1999)
61. Zhang, L., Zhou, W., Jiao, L.: Wavelet support vector machine. IEEE Transactions on Systems, Man, and Cybernetics **34**(1), 34–39 (2004)