

Structure-Based Document Model with Discrete Wavelet Transforms and Its Application to Document Classification

Supphachai Thaicharoen¹

Tom Altman¹

Krzysztof J. Cios²

¹ Department of Computer Science and Engineering,
University of Colorado Denver, Campus Box 109,
PO Box 173364, Denver, CO 80217-3364, U.S.A.

Email: supphachai.thaicharoen@email.cudenver.edu, tom.altman@ucdenver.edu

² Virginia Commonwealth University,
Richmond, VA 23238, U.S.A.;
IITiS PAN, Poland.
Email: kcios@vcu.edu

Abstract

Term signal is an existing text representation that depicts a term as a vector of frequencies of occurrences in a number of user-defined partitions of a document. Although term signal augments the traditional vector space model with patterns of term occurrences, its document division is not coherent with the actual logical structure of a document. In this paper, we propose a novel document model, termed Structure-Based Document Model with Discrete Wavelet Transforms (SDMDWT), that exploits the structural information of documents and mathematical transforms for document representation. The proposed SDMDWT model enhances the existing term signal concept by additionally taking into consideration document's structural information during document division. We evaluated the proposed model on two different domains of standard data sets, WebKB 4-Universities and TREC Genomics 2005, using Support Vector Machines binary classification. The experimental results show that using our SDMDWT model for document representation demonstrates promising improvements of classification performances over existing document models.

1 Introduction

Document representation is one of the important tasks in text mining particularly for document classification and clustering. Its traditional approach is based on the “bag of words” or vector space model (VSM) where a document is represented by a vector of weights of unique terms selected from a data set. Weights are typically computed from frequency of term occurrences either within a document (term frequency) or across a data set (document frequency), or both.

In addition to the vector space representation, Park et al. proposed a concept of term signal that takes into account both frequency information and patterns of term occurrences in a document (Park et al. 2004, 2002 a,b , Park, Palaniswami & Ramamohanarao 2005, Park & Ramamohanarao 2004, Park, Ramamohanarao & Palaniswami 2005). Term signal is a representation that describes frequency of a term

in physical locations of a document. It augments the traditional vector space model with patterns of term occurrences. With term signal, a document is first divided into a number of partitions based on a sequence and the number of terms in the document. Then, a term is represented as a vector of frequencies of term occurrences in those partitions. Finally, a document, consisting of a number of chosen terms, is represented as a vector of term signals. Park et al. additionally used a number of mathematical transforms such as Cosine Transforms, Fourier Transforms and Discrete Wavelet Transforms on their document representation, and computed document ranking based on query terms. Pryczek and Szczepaniak applied this term signal concept to document classification using Fourier Transforms (Pryczek & Szczepaniak 2006). Using the term signal with mathematical transforms for document representation was shown to be better than the traditional vector space representation for information retrieval in Park et al.'s and for document classification in Pryczek and Szczepaniak's studies. In this paper, we used document representation model based on this term signal concept with Discrete Wavelet Transforms as one of baseline models, and referred to this model as the *Spectral Space Model with Discrete Wavelet Transforms* (SPSMDWT).

Another method for enhancing document model is by additionally including structural information of documents into document representation. With the increase of publicly available full-text databases such as PubMed Central¹ and the widespread uses of semi-structure documents such as XML and HTML pages, the document structural approach became increasingly studied. Hakenberg et al. exploited structural information of full-text biomedical articles by assigning different weights to term occurrences in different sections, which resulted in performance improvements on document classification from the baseline classifier (Hakenberg et al. 2005). Denoyer and Gallinari proposed a Bayesian network model for semi-structure document classification that can be used to handle structural information and different types of document content (Denoyer & Gallinari 2004). In their study, a document is viewed as a tree where each node is corresponding to a document component, and links between two nodes represent dependencies or relations between document components. Model parameters are learned from training documents for each class, and a Bayesian network model is then built for each document. Based on experimental results, their proposed generative models were superior to the models that did not take into account the structural information of documents. Bratko and

Copyright ©2008, Australian Computer Society, Inc. This paper appeared at the Seventh Australasian Data Mining Conference (AusDM 2008), Glenelg, Australia. Conferences in Research and Practice in Information Technology (CRPIT), Vol. 87, John F. Roddick, Jiuyong Li, Peter Christen and Paul Kennedy, Ed. Reproduction for academic, not-for profit purposes permitted provided this text is included.

¹<http://www.pubmedcentral.nih.gov/>

Filipič investigated a number of document models, named tagging, splitting and stacking, that utilize document structural information for document categorization (Bratko & Filipič 2006). For the tagging approach, the same words that occur in different sections of a document are treated as different words. For the splitting approach, texts in different document sections are modeled and evaluated separately, and the results are combined to give the final prediction. Finally, the stacking approach is similar to the splitting approach in that texts in different sections of a document are modeled and evaluated separately. The difference is that the final prediction is generated by a meta classifier that is built from the prediction results of classifications on different sections. Bratko and Filipič found that the stacking approach gave the best result.

In this paper, we propose a novel document representation model, termed Structure-Based Document Model with Discrete Wavelet Transforms (SDMDWT). The proposed SDMDWT model is built upon the concept of term signal by additionally taking into consideration document structure. With the SDMDWT model, a whole data set is first analyzed and the overall structure of its documents is captured. Then, original documents are pre-processed and converted into intermediate semi-structured documents in order to facilitate further processing. Finally, structured-based document model is constructed for each document and the Discrete Wavelet Transforms are applied. Our choice of using Discrete Wavelet Transforms for our model rather than other mathematical transforms is based on the latest work by Park et al. (Park, Ramamohanarao & Palaniswami 2005).

We evaluated our method on two different domains of standard data sets, WebKB 4-Universities and TREC Genomics 2005, using Support Vector Machines binary classification. Support Vector Machines (SVM) have been shown to work well with high-dimensional data and to be suitable to document classification (Joachims 1998). We utilized the VSM and SPSMDWT as baseline document models. The experimental results show that our SDMDWT model outperforms VSM and SPSMDWT on both standard data sets based on F-measure, micro-averaged F-measure and macro-averaged F-measure.

This paper is organized as follows. We present the technical background in Section 2, which covers the term signal concept, term and signal weighting schemes and Discrete Wavelet Transforms. Then, we describe our proposed document model, the data sets used in this paper and the pre-processing framework including a feature selection approach in Section 3. Next, we explain our evaluation method and provide experimental results in Section 4. Finally, we discuss experimental results and conclude this paper in Section 5.

2 Technical background

In this section, we describe background methods that are useful to understand the rationale behind our approach. We begin with the concept of term signal that our proposed method is built on, then weighting schemes for weighting terms and term signals and finally wavelet transforms that are used to transform term signals from the frequency domain to the wavelet domain.

2.1 Term signal

A term signal, introduced by Park et al. (Park et al. 2004, 2002a,b, Park, Palaniswami & Ramamohanarao

2005, Park & Ramamohanarao 2004, Park, Ramamohanarao & Palaniswami 2005), is a vector representation of terms that describes frequencies of term occurrences in particular partitions within a document. To construct a term signal, a document is first divided into a user-defined B number of sections. Then, a term signal t in a document d can be represented as a vector of physical sections by Equation 1.

$$s(t, d) = [f_{t,1,d}, f_{t,2,d}, \dots, f_{t,B,d}], \quad (1)$$

where $f_{t,b,d}$ is frequency of term t in section b of document d for $0 < b \leq B$. $f_{t,b,d}$ can also be considered as the b^{th} signal component of a term signal $s(t, d)$. For example, suppose that a document consisting of a sequence of 32 words is divided into 8 partitions. There will be 4 words per partition or bin. The document d can be graphically represented, see Figure. 1.

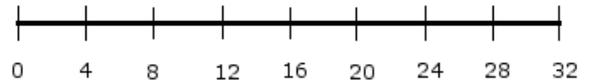


Figure 1: Document d with 8 partitions.

Accordingly, if a term t occurs once in each of the 2^{nd} , 3^{rd} , 10^{th} and 24^{th} positions in a document d , its term signal $s(t, d)$ can be represented by Equation 2 and depicted by Figure. 2.



Figure 2: Term signal t in document d .

$$s(t, d) = [2, 0, 1, 0, 0, 1, 0, 0]. \quad (2)$$

As shown in Figure 2, term t occurs two times in the 1^{st} bin, one time in the 3^{rd} bin and one time in the 6^{th} bin, respectively.

2.2 Weighting schemes

Weighting scheme is an assignment of numerical weights to terms in a vector space model. Weight of a term can be computed using a number of parameters such as term frequency, document frequency, document length, number of documents in a data set, etc. One of the most commonly used term-weighting schemes for document classification is $TF \cdot IDF$, which stands for term frequency multiplied by the inverse of document frequency. It can be formulated by Equation 3.

$$TF \cdot IDF = TF \times \lg(N/DF), \quad (3)$$

where TF is term frequency or frequency of term occurrences within a document, N is the total number of documents in a data set and DF is document frequency or frequency of term occurrences across a data set. The underlying assumption of $TF \cdot IDF$ weighting scheme is that terms that occur in many documents, high DF , are common and do not represent documents well. In contrast, terms that occur very often in a document, high TF , are considered

important features of the document. $TF \cdot IDF$ compensates between term frequency and document frequency.

For document representation using term signal, variations of $TF \cdot IDF$ weighting schemes can be used for weighting a term signal as described in (Park et al. 2002b). One of the variations, $PTF \cdot IDF$, is formulated by Equation 4.

$$PTF \cdot IDF = (1 + \lg(f_{t,d})) \left(\frac{f_{t,b,d}}{f_{t,d}} \right) \times \lg(N/DF), \quad (4)$$

where $f_{t,d}$ is frequency of occurrences of term t in document d and $f_{t,b,d}$ is frequency of occurrences of term t in partition b of document d .

2.3 Discrete Wavelet Transforms

A wavelet is a mathematical function in time/space domain, which can be expressed by Equation 5.

$$\psi_{s,l}(t) = 2^{s/2} \psi(2^s t - l), \quad (5)$$

where s is a dilation or scaling parameter, l is a translation or time-/space-location parameter and $s, l \in Z$. For any function $f(t) \in L^2(R)$ and for which $\psi_{s,l}(t)$ forms an orthonormal basis for the space of signals of interest (in this case, $f(t)$), a wavelet transform of $f(t)$ can be computed by Equation 6.

$$\Psi(s, l) = \langle f(t), \psi_{s,l}(t) \rangle = \int_{-\infty}^{\infty} f(t) \psi_{s,l}^*(t) dt, \quad (6)$$

where $\psi_{s,l}^*(t)$ is a complex conjugate of $\psi_{s,l}(t)$. Wavelet transform can be described by a concept of multi-resolution analysis. Multi-resolution analysis is the decomposition of a signal into sub-signals of different resolutions or scales. In each step of multi-resolution analysis, a signal is decomposed into two sub-signals, approximation and detail. The approximation sub-signal is expressed by a linear combination of scaling functions $\varphi(t)$, and the detail sub-signal is represented by a linear combination of wavelet functions $\psi(t)$.

The scaling function is a finite energy function in $L^2(R)$ and is defined by Equation 7.

$$\varphi_{s,l}(t) = 2^{s/2} \varphi(2^s t - l), \text{ where } s, l \in Z. \quad (7)$$

In multi-resolution analysis, the subspace spanned by the scaling function must satisfy the following properties.

$$V_s \subset V_{s+1} \text{ for all } s \in Z,$$

$$V_{-\infty} = \{0\} \text{ and } V_{\infty} = L^2$$

and

$$\{0\} \leftarrow \dots \subset V_{-1} \subset V_0 \subset V_1 \subset \dots \rightarrow L^2.$$

The wavelet function is also a finite energy function in $L^2(R)$ and is defined by Equation 8.

$$\psi_{s,l}(t) = 2^{s/2} \psi(2^s t - l), \text{ where } s, l \in Z. \quad (8)$$

The subspace spanned by the wavelet function, W_s is an orthogonal complement of V_s in V_{s+1} . In other words, $V_s \perp W_s$ and $V_{s+1} = V_s \oplus W_s$, which leads to the following properties.

$$L^2 = \dots \oplus W_{-1} \oplus W_0 \oplus W_1 \oplus \dots$$

and

$$W_{-\infty} \oplus \dots \oplus W_{-1} = V_0.$$

Let V_s be a subspace spanned by scaling functions, W_s be a subspace spanned by wavelet functions, $V_0 \subset V_1 \subset \dots \subset L^2$ and $V_s = V_{s-1} \oplus W_{s-1}$, any function $f(t) \in L^2(R)$, where $L^2(R) = V_0 \oplus W_0 \oplus W_1 \oplus W_2 \oplus \dots$, can be mathematically expressed by Equation 9.

$$f(t) = \sum_{l=-\infty}^{\infty} a_l \varphi_l(t) + \sum_{s=0}^{\infty} \sum_{l=-\infty}^{\infty} d_{s,l} \psi_{s,l}(t). \quad (9)$$

The first term is mapped to the approximation sub-signal, and the second term is referred as the detail sub-signal. The coefficients a_l and $d_{s,l}$ are called discrete wavelet transforms, which can be computed by Equations 10 and 11.

$$a_l = \langle f(t), \varphi_{s,l}(t) \rangle = \int f(t) \varphi_{s,l}^*(t). \quad (10)$$

$$d_{s,l} = \langle f(t), \psi_{s,l}(t) \rangle = \int f(t) \psi_{s,l}^*(t). \quad (11)$$

Discrete wavelet transforms can be efficiently calculated by using the filter bank tree-structured algorithm. The filter bank tree of discrete wavelet transforms of a signal $f(t)$ can be recursively expressed by Equation 12.

$$\begin{aligned} f & \xrightarrow{DWT^1} A^1 + D^1 & (12) \\ & \xrightarrow{DWT^2} A^2 + D^2 + D^1 \\ & \xrightarrow{DWT^3} A^3 + D^3 + D^2 + D^1 \\ & \dots \\ & \xrightarrow{DWT^s} A^s + D^s + D^{s-1} + \dots + D^1, \end{aligned}$$

where A^s represents an approximation sub-signal and D^s corresponds to a detail sub-signal in the s^{th} level of transforms of discrete wavelet transforms of the signal $f(t)$, respectively.

As described by Equations 10 and 11, discrete wavelet transform coefficients are computed by inner products between the signal itself and the scaling/wavelet functions. For example, the 1-levels of Haar scaling and Haar wavelet signals are defined by Equations 13 and 14.

$$V_1^1 = \left(\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}, 0, 0, \dots, 0 \right), \quad (13)$$

$$V_2^1 = \left(0, 0, \frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}, 0, 0, \dots, 0 \right),$$

$$V_3^1 = \left(0, 0, 0, 0, \frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}, 0, 0, \dots, 0 \right),$$

...

$$V_{N/2}^1 = \left(0, 0, \dots, 0, \frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}} \right).$$

$$W_1^1 = \left(\frac{1}{\sqrt{2}}, \frac{-1}{\sqrt{2}}, 0, 0, \dots, 0 \right), \quad (14)$$

$$\begin{aligned}
W_2^1 &= (0, 0, \frac{1}{\sqrt{2}}, \frac{-1}{\sqrt{2}}, 0, 0, \dots, 0), \\
W_3^1 &= (0, 0, 0, 0, \frac{1}{\sqrt{2}}, \frac{-1}{\sqrt{2}}, 0, 0, \dots, 0), \\
&\dots \\
W_{N/2}^1 &= (0, 0, \dots, 0, \frac{1}{\sqrt{2}}, \frac{-1}{\sqrt{2}}).
\end{aligned}$$

Hence, for a discrete signal $f(t) = (4, 6, 10, 12, 8, 6, 5, 5)$, its 1-level Haar transform can be computed by Equations 15 and 16, and is conclusively represented by Equation 17.

$$\begin{aligned}
A_1^1 &= \langle f(t), V_1^1 \rangle = 5\sqrt{2}, \\
A_2^1 &= \langle f(t), V_2^1 \rangle = 11\sqrt{2}, \\
A_3^1 &= \langle f(t), V_3^1 \rangle = 7\sqrt{2}, \\
A_4^1 &= \langle f(t), V_4^1 \rangle = 5\sqrt{2}.
\end{aligned} \tag{15}$$

$$\begin{aligned}
D_1^1 &= \langle f(t), W_1^1 \rangle = -\sqrt{2}, \\
D_2^1 &= \langle f(t), W_2^1 \rangle = -\sqrt{2}, \\
D_3^1 &= \langle f(t), W_3^1 \rangle = \sqrt{2}, \\
D_4^1 &= \langle f(t), W_4^1 \rangle = 0.
\end{aligned} \tag{16}$$

$$f(t) \xrightarrow{H^1} (5\sqrt{2}, 11\sqrt{2}, 7\sqrt{2}, 5\sqrt{2} | -\sqrt{2}, -\sqrt{2}, \sqrt{2}, 0). \tag{17}$$

The numerical example above was excerpted from (Walker 2008). Interested readers can further study wavelets from several useful resources such as (Burrus et al. 1998), (Goswami & Chan 1999), (Mix & Olejniczak 2003) and (Walker 2008).

3 Data and methods

3.1 Data

WebKB 4-Universities. The WebKB 4-Universities data set² is a collection of 8282 web pages collected from computer science departments of several universities by the World Wide Knowledge Base project of the Carnegie Mellon text learning group in January 1997. The web pages in the data set are manually classified into the following 7 categories, student, faculty, staff, course, project, department and other. Each category contains web pages gathered from 4 main universities, Texas, Washington, Wisconsin and Cornell, and the remaining pages collected from other universities. Four most populous categories, student, faculty, course and project, are used in this paper, which accounts for 4199 web pages.

TREC Genomics 2005. The TREC Genomics 2005 data set is a corpus of full-text documents in SGML format. The data set is a collection of mouse genome articles from three journals, Journal of Biological Chemistry (JBC), Journal of Cell Biology (JCB), and Proceedings of the National Academy of Science (PNAS), over a two-year (2002-2003) period. There are four major types of articles in the data set – Alleles of mutant phenotypes, Embryologic gene expression, Gene Ontology and Tumor biology, which are corresponding to the following four class labels, A, E, G and T, respectively. TREC Genomics 2005 data set consists of 5837 training documents and

6403 testing documents. Among the 5837 training documents, 338 documents are related to Alleles (A), 81 documents to Gene Expression (E), 462 documents to Gene Ontology (G) and 36 documents to Tumor (T). In 6403 testing documents, 332 documents are assigned to class A, 105 documents to class E, 518 documents to class G and 20 documents to class T. The remaining documents do not have any class labels associated with them.

3.2 The proposed SDMDWT model and its preprocessing framework

The SDMDWT document model is an enhancement of the term signal proposed by Park et al. (Park et al. 2004, 2002a,b, Park, Palaniswami & Ramamohanarao 2005, Park & Ramamohanarao 2004, Park, Ramamohanarao & Palaniswami 2005) such that each bin is mapped to a document component in a document based on the captured document structure rather than is derived from computation. We call our term signal based on document structure the *structure-based term signal*. The structure-based term signal is defined in Definition 1.

Definition 1. *Structure-based term signal.*

Structure-based term signal is a vector of frequencies of term occurrences in different components of a document, where components are derived from the actual document structure. The structure-based term signal of a term t in a document d is defined by Equation 18.

$$st(t, d) = [f_{t,c_1,d}, f_{t,c_2,d}, \dots, f_{t,c_n,d}], \tag{18}$$

where $f_{t,c_1,d}, f_{t,c_2,d}, \dots, f_{t,c_n,d}$ are corresponding to the frequencies of term t in document components c_1, c_2, \dots, c_n of document d , respectively.

The key differences between our proposed structure-based term signal (Equation 18) and the existing term signal (Equation 1) are (i) the number of components of our model is derived from the actual document structure such as the number of sections, but that of the existing term signal is defined by users and (ii) the length of each component in our model is based on the actual length of document components. However, the length of each component in the existing term signal model is computed from the total number of terms in a document divided by the user-defined number of components.

Definition 2. *Structure-based document model.*

According to the structure-based term signal in Definition 1, a document can be represented by a vector of structure-based term signals, which is defined by Equation 19.

$$d = [st(t_1, d), st(t_2, d), \dots, st(t_n, d)], \tag{19}$$

where t_1, t_2, \dots, t_n are term features selected in the feature selection step, and $st(t_1, d), st(t_2, d), \dots, st(t_n, d)$ are the structure-based term signals of terms t_1, t_2, \dots, t_n in document d , respectively.

The SDMDWT preprocessing framework. The preprocessing framework for constructing SDMDWT model is summarized as follows.

1. *Capture the document structure of a data set:* The first step for constructing the proposed SDMDWT model is to analyze documents in the data set and to capture the common characteristics of their document structure. As mentioned by Bratko and Filipič (Bratko & Filipič 2006),

²<http://www.cs.cmu.edu/afs/cs.cmu.edu/project/theo-20/www/data/>

how the structure of a document is captured depends on particular semantics of document structure and its perceived relevance to the text mining task that is document classification in this paper. For WebKB 4-Universities data set, we divided each document into the following 2 components, heading text (H) and non-heading text (N). We used the `<h#>` and `</h#>` HTML tags for distinguishing between the two components. The heading text is text enclosed by the `<h#>` and `</h#>` HTML tags, and the non-heading text is text that is not. The “#” inside each `<h#>` tag represents the level number of the `<h>` tag. We also included text enclosed by `<title>` and `</title>` into the heading text component. For TREC Genomics 2005 biomedical data set, we partitioned each document into the following 7 components based on the actual logical organization of documents, *Title* (T), *Abstract* (A), *Introduction* (I), *Method* (M), *Result* (R), *Conclusion* (C) and *Other* (O). Note that these document divisions are by no means definite. They depend on the data set used and how the document structure is captured. For example, for full-text biomedical documents, *Title* and *Abstract* may be combined into one component instead of being independent components.

2. *Collect variants of component labels and construct a mapping table:*

Although documents in the same domain tend to have similar structure, it is uncommon that their component labels are different. For example, in the WebKB 4-Universities HTML pages, the HTML heading `<h>` tags have several levels such as `<h1>`, `<h2>`, `<h3>`, `<h4>`, etc. Moreover, in TREC Genomics 2005 data set, the *Method* components of various papers are labeled as “Methods”, “Patients and methods”, “Experimental procedures”, etc. Therefore, after collecting all variants of component labels in the data set, a mapping table was constructed for mapping variants of component labels to their corresponding user-defined component names. Table 1 gives an example of name variants of the *Method* section in TREC Genomics 2005 data set.

3. *Pre-process documents and perform feature selection:*

For WebKB 4-Universities data set, we utilized a combination of JTidy³ and Java regular expression to clean up and parse the original HTML documents into the semi-structured XML documents. We used simple words as features and did not perform word stemming and stop-word removal on this data set. For TREC Genomics 2005 data set, we implemented an SGML Java parser to parse original SGML documents into the semi-structured XML documents. We utilized Lingpipe⁴ to break texts into sentences and used Genia Tagger⁵ to perform part-of-speech tagging and to detect terms, phrases and biological entities. We used phrases as features and removed those that are in the PubMed stop word list⁶. Word stemming using Porter stemmer (Porter 1997) was also carried out. For both data sets, we ranked terms using Information Gain based on (Yang & Pedersen 1997), which is formulated by Equation 20, and then selected the

Table 1: An example of the mapping tables.

Name variations of the <i>Method</i> component	
	methods
	method
	experimental procedures
	experimental procedure
	experiemntal procedures
	experimantal procedures
	experimental proceduures
	experimental procedoures
	experimental approach
	experimental approaches
	experimental results and interpretation
	materials and methods
	material and methods
	materials
	mateials and methods
	matelials and methods
	materials methods
	metrials and methods
	methods and materials
	methods and methods
	patients and methods
	subjects and methods
	computational methods
	model and methods
	research design and methods
	media and materials

top n terms to be used in experiments. Note that we cleaned and parsed texts in the original documents into XML documents with organized components as intermediate representation in order to facilitate the construction of structure-based term signals and document models in the next steps.

$$\begin{aligned}
 IG(t) = & - \sum_{i=1}^m P_r(c_i) \log P_r(c_i) \\
 & + P_r(t) \sum_{i=1}^m P_r(c_i|t) \log P_r(c_i|t) \\
 & + P_r(\bar{t}) \sum_{i=1}^m P_r(c_i|\bar{t}) \log P_r(c_i|\bar{t}),
 \end{aligned} \tag{20}$$

where $P_r(c_i)$ is the probability of class c_i , $P_r(t)$ is the probability of term t , $P_r(c_i|t)$ is the probability of a document that contains term t and has class c_i , and finally $P_r(c_i|\bar{t})$ is the probability of a document that does not contain term t and has class c_i .

4. *Represent each term using the proposed structure-based term signal:*

After pre-processing documents and selecting features, we constructed the structured-based term signal for each selected term. For WebKB 4-Universities data set, since each WWW page is divided into 2 components, the heading text component (H) and non-heading text component (N). As a result, each selected term could be constructed by Equation 21.

$$st(t, d) = [f_{t,CH,d}, f_{t,CN,d}], \tag{21}$$

For TREC Genomics 2005 data set, each selected term was represented by Equation 22. Since

³<http://jtidy.sourceforge.net>

⁴<http://alias-i.com/lingpipe>

⁵<http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/tagger/>

⁶<http://www.ncbi.nlm.nih.gov/books/bv.fcgi?highlight=\&stopwords&rid=helppubmed.table.pubmedhelp.T43>

Discrete Wavelet Transforms requires a signal length to be a power of two, we added an additional zero-value component to the structure-based term signal.

$$st(t, d) = [f_{t,c_T,d}, f_{t,c_A,d}, f_{t,c_I,d}, f_{t,c_M,d}, f_{t,c_R,d}, f_{t,c_C,d}, f_{t,c_O,d}, 0], \quad (22)$$

5. *Apply pre-weighting and Discrete Wavelet Transforms to each structured-based term signal:*

In this step, we applied the pre-weighting in Equation 4 and Haar Discrete Wavelet Transforms explained in the technical background section to each structured-based term signal in each document.

6. *Construct the structured-based document model:* Finally, for both data sets, each document was constructed as a vector of structured-based term signals as defined by Equation 19.

4 Experiments and results

WebKB 4-Universities. We evaluated our SDMDWT model using an open source LIBSVM (Chang & Lin 2001) in Weka (Witten & Frank 2005), with C-SVC, linear kernel and 0.01 tolerance as SVM parameter values. We generated a sub data set for each class with one-against-all strategy. For example, if documents in the category “student” is specified as positive documents, all other documents in the data set will be labeled as negative documents. For each sub data set, we performed 10-cross validation with binary classification and then collected results.

TREC Genomics 2005. We evaluated our SDMDWT model using an open source LIBSVM library (Chang & Lin 2001) with C-SVC, linear kernel and 0.001 tolerance as parameter values. We also used one-against-all strategy. For documents that belong to more than one class, if one of their classes is the positive class under consideration, then we assign positive labels to them. For each class, we performed a binary classification with train/test sub data sets, and then collected results.

For both data sets, we utilized F-measure, micro-averaged F-measure and macro-averaged F-measure as performance measures.

F-measure or F1-measure is a combination of Precision and Recall, defined by Equation 23.

$$F\text{-measure} = \frac{2 * Precision * Recall}{Precision + Recall}. \quad (23)$$

Precision and Recall can be calculated by Equations 24 and 25.

$$Precision = \frac{TP}{TP + FP}, \quad (24)$$

$$Recall = \frac{TP}{TP + FN}, \quad (25)$$

where TP, TN, FN and FP are the number of true positives, true negatives, false negatives and false positives, respectively.

The micro-average and macro-average are performance averages across multiple categories. As described in (Yang 1999), the micro-averaged performance is viewed as a per-document average because it gives equal weight to every document. To compute a micro-averaged performance, a global contingency

Table 2: WebKB 4-Universities: Micro-averaged F-measure.

F ^a /M ^b	SDMDWT	SPSMDWT	VSM
500	0.928794	0.924501	0.887868
1000	0.926999	0.923317	0.894601
1500	0.929770	0.927415	0.900745
2000	0.928939	0.924868	0.900231
2500	0.929213	0.921566	0.896142
5000	0.924592	0.919918	0.892918
7500	0.922485	0.921965	0.884373

^aNumber of features

^bDocument model

Table 3: WebKB 4-Universities: Macro-averaged F-measure.

F ^a /M ^b	SDMDWT	SPSMDWT	VSM
500	0.923765	0.917390	0.867331
1000	0.922871	0.916421	0.874013
1500	0.926470	0.921714	0.884561
2000	0.923818	0.918427	0.883605
2500	0.923653	0.914590	0.880300
5000	0.918141	0.914226	0.873226
7500	0.915144	0.914865	0.863115

^aNumber of features

^bDocument model

table is constructed, whose cell value is the sum of the corresponding cell in each contingency table of each class. For example, the number of true positives in the global contingency table is the sum of the number of true positives from all contingency tables of all classes. Then, a micro-averaged performance such as micro-averaged Precision or micro-averaged Recall is computed from this global contingency table. In this paper, we use the micro-averaged F-measure, which can be computed by Equation 26.

$$\text{Micro-averaged F-measure} = \quad (26)$$

$$\frac{2 * \text{Micro-averaged Precision} * \text{Micro-averaged Recall}}{\text{Micro-averaged Precision} + \text{Micro-averaged Recall}}.$$

The macro-averaged performance is considered per-category average because it gives equal weight to every class. It can be computed by the sum of performance from each class divided by the total number of classes. The macro-averaged F-measure can be computed by Equation 27.

$$\text{Macro-averaged F-measure} = \frac{\sum_{i=1}^c F\text{-measure}_i}{c}, \quad (27)$$

where F-measure_{*i*} is the F-measure of class *i*, and *c* is the number of classes.

4.1 WebKB 4-Universities

According to Tables 2 and 3, we can conclude that our SDMDWT model is better than SPSMDWT that is the document model based on the original term signal concept, and it distinguishably outperforms VSM for all different numbers of features based on micro-averaged and macro-averaged F-measures.

In addition, in the Faculty and Project categories, our SDMDWT model is clearly superior to SPSMDWT and VSM models based on F-measure. The performance comparisons of document models for these two classes are shown in Figures 3 and 4, accordingly.

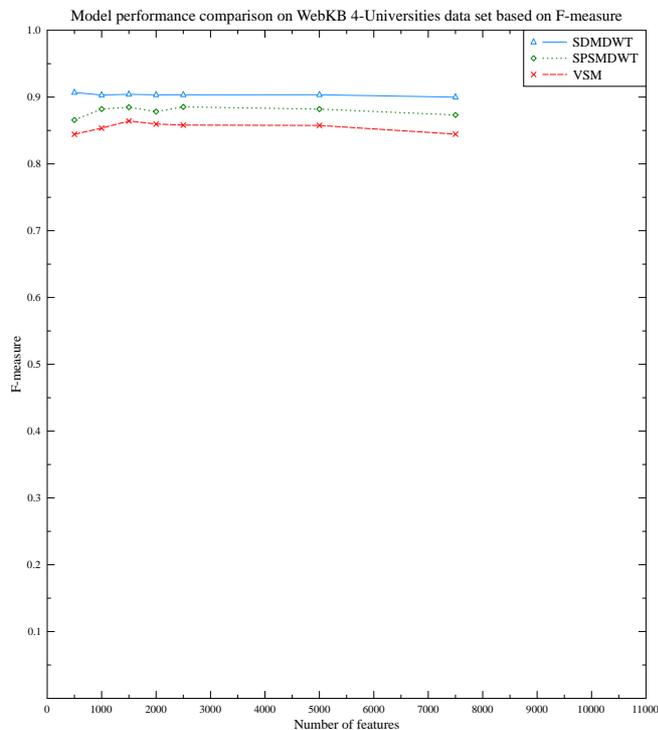


Figure 3: WebKB 4-Universities: Performance comparison based on F-measure when the Faculty category is positive category.

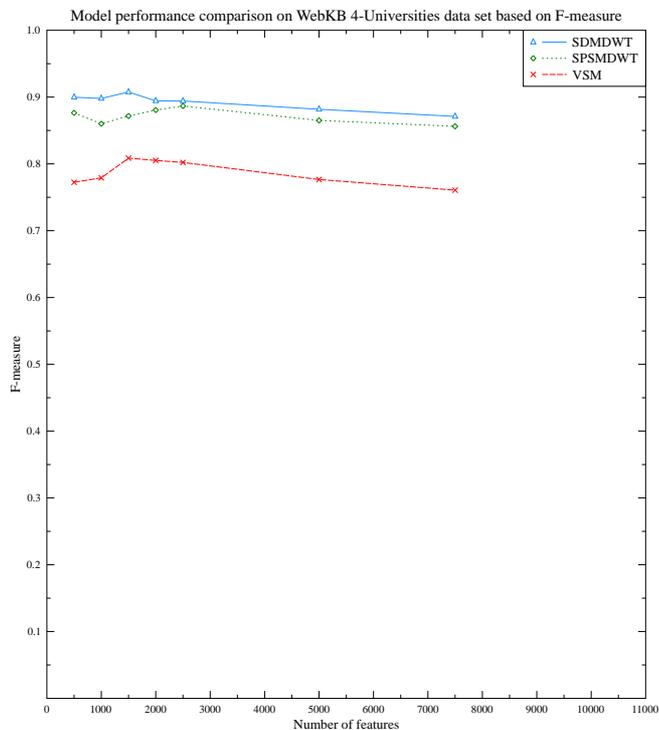


Figure 4: WebKB 4-Universities: Performance comparison based on F-measure when the Project category is positive category.

Table 4: TREC Genomics 2005: Micro-averaged F-measure.

F^a/M^b	SDMDWT	SPSMDWT	VSM
7500	0.244156	0.168052	0.223663
8500	0.213144	0.156863	0.180879
10000	0.198377	0.138127	0.100000

^aNumber of features

^bDocument model

4.2 TREC Genomics 2005

Based on Tables 4 and 5, we can conclude that our SDMDWT model is superior to SPSMDWT and VSM models based on the micro-averaged and macro-averaged F-measures.

Moreover, our SDMDWT model gives distinct performance improvements based on F-measure compared to SPSMDWT and VSM model where class Alleles (A) is considered the positive class, which is shown by Figure 5.

5 Conclusions

In this paper, we proposed a novel document representation model, termed Structure-Based Document Model with Discrete Wavelet Transforms (SDMDWT), that exploits structural information of doc-

Table 5: TREC Genomics 2005: Macro-averaged F-measure.

F^a/M^b	SDMDWT	SPSMDWT	VSM
7500	0.141999	0.104290	0.129638
8500	0.126299	0.098373	0.110393
10000	0.119382	0.087978	0.065065

^aNumber of features

^bDocument model

uments and Discrete Wavelet Transforms for document representation. The proposed model is built on the existing term signal concept that represents a pattern of term occurrences in different partitions of a document as a vector. The main difference between our SDMDWT and SPSMDWT models lies on the fact that our SDMDWT model takes into consideration structural information when partitioning a document. Accordingly, a document division of our SDMDWT model is more coherent with the actual logical document structure than that of SPSMDWT model. This inclusion of structural information into document representation allows further improvement of text mining tasks where document structure is concerned such as weighting document components differently. The pre-processing framework of the proposed SDMDWT document model can be divided into the following steps: (i) capturing document structure, (ii) collecting all various names of document component headings and constructing a mapping table, (iii) pre-processing documents and performing feature selection, (iv) representing each selected term using the structure-based term signal, (v) applying pre-weighting and Discrete Wavelet Transforms to each structure-based term signal and (vi) constructing the structure-based document models.

According to the experimental results on both TREC Genomics 2005 and WebKB 4-Universities data sets, using our SDMDWT model for document representation gives better classification performances (F-measure, micro-averaged F-measure and macro-averaged F-measure) than using the traditional vector space model (VSM) and the document model that is based on the original term signal concept (SPSMDWT). The clear performance improvements based on F-measure occur on the Faculty and Project categories of WebKB 4-Universities dataset and on the Alleles (A) category of TREC Genomics 2005 dataset, which are shown by Figures 3, 4 and 5, respectively. Therefore, we can conclude that struc-

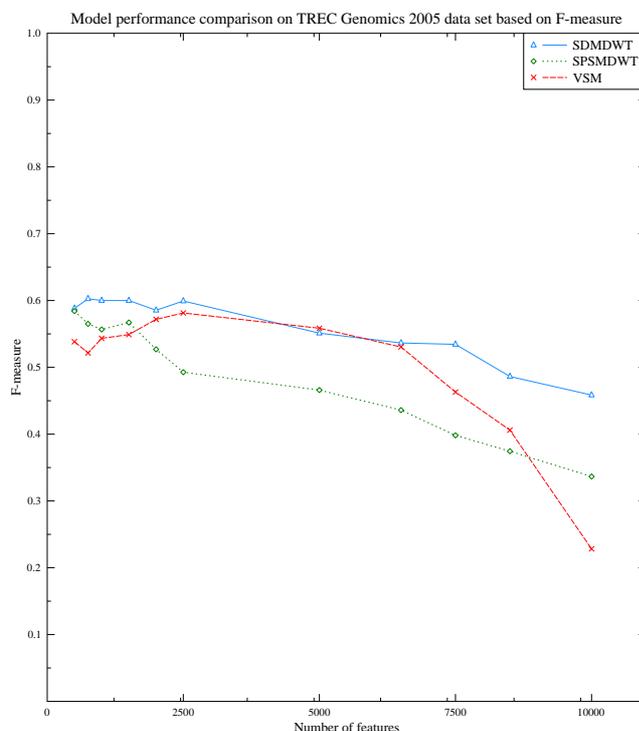


Figure 5: TREC 2005: Performance comparison based on F-measure when the Alleles (A) category is positive category.

tural information of documents can be incorporated into document representation to improve performance of document classification.

Lessons learned from this research study are that although several text documents in scientific or WWW domains are presented in the semi-structured formats such as XML, SGML and HTML, to be able to exploit structural information, a manual analysis is still required for capturing the common structural characteristics of documents. In addition, even with the same document structure, components in different documents are generally labeled with different names such as “Conclusions”, “Concluding remarks”, “Summary”, etc. As a result, to facilitate pre-processing, full-text documents (particularly in scientific domain) should be standardized on the labels and number of main components and should be presented in a semi-structured format such as XML with component labels used as element names.

For future work, our proposed technique and idea could possibly be applied to other types of text mining tasks for performance improvement such as document clustering and feature selection that additionally take into account document structure.

6 Acknowledgments

The authors thank all researchers at the Center for Computational Pharmacology at the University of Colorado Denver. Particularly, William Baumgartner for a number of useful discussions at the beginning of this work, and Dr. Lawrence Hunter for providing the data and other resources for this research. We also thank Sam Wheeler, System Administrator and Lab Manager at the College of Engineering and Applied Science, University of Colorado Denver, for his technical assistance.

References

- Bratko, A. & Filipič, B. (2006), ‘Exploiting structural information for semi-structured document categorization’, *Inf. Process. Manage.* **42**(3), 679–694.
- Burrus, C. S., Gopinath, R. A. & Guo, H. (1998), *Introduction to Wavelets and Wavelet Transforms: A Primer*, Prentice Hall.
- Chang, C.-C. & Lin, C.-J. (2001), *LIBSVM: a library for support vector machines*. Software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>.
- Denoyer, L. & Gallinari, P. (2004), ‘Bayesian network model for semi-structured document classification’, *Inf. Process. Manage.* **40**(5), 807–827.
- Goswami, J. C. & Chan, A. K. (1999), *Fundamentals of Wavelets: Theory, Algorithms, and Applications*, John Wiley & Sons, Inc.
- Hakenberg, J., Rutsch, J. & Leser, U. (2005), Tuning text classification for hereditary diseases with section weighting, in ‘Proc International Symposium on Semantic Mining in Biomedicine, SMBM’, Hinxton, UK, pp. 34–37.
- Joachims, T. (1998), Text categorization with support vector machines: Learning with many relevant features, in ‘ECML ’98: Proceedings of the 10th European Conference on Machine Learning’, Springer-Verlag, London, UK, pp. 137–142.
- Mix, D. F. & Olejniczak, K. J. (2003), *Elements of wavelets for engineers and scientists*, Wiley-Interscience.
- Park, L. A. F., Palaniswami, M. & Ramamohanarao, K. (2002a), A new implementation technique for fast spectral based document retrieval systems, in ‘ICDM ’02: Proceedings of the 2002 IEEE International Conference on Data Mining (ICDM’02)’, IEEE Computer Society, Washington, DC, USA, p. 346.
- Park, L. A. F., Palaniswami, M. & Ramamohanarao, K. (2002b), A novel web text mining method using the discrete cosine transform, in ‘PKDD ’02: Proceedings of the 6th European Conference on Principles of Data Mining and Knowledge Discovery’, Springer-Verlag, London, UK, pp. 385–396.
- Park, L. A. F., Palaniswami, M. & Ramamohanarao, K. (2005), ‘A novel document ranking method using the discrete cosine transform’, *IEEE Trans. Pattern Anal. Mach. Intell.* **27**(1), 130–135. Member-Laurence A. F. Park and Sr. Member-Marimuthu Palaniswami and Member-Kotagiri Ramamohanarao.
- Park, L. A. F. & Ramamohanarao, K. (2004), Hybrid pre-query term expansion using latent semantic analysis, in ‘ICDM ’04: Proceedings of the Fourth IEEE International Conference on Data Mining (ICDM’04)’, IEEE Computer Society, Washington, DC, USA, pp. 178–185.
- Park, L. A. F., Ramamohanarao, K. & Palaniswami, M. (2005), ‘A novel document retrieval method using the discrete wavelet transform’, *ACM Trans. Inf. Syst.* **23**(3), 267–298.
- Park, L. A., Ramamohanarao, K. & Palaniswami, M. (2004), ‘Fourier domain scoring: A novel document ranking method’, *IEEE Transactions on Knowledge and Data Engineering* **16**(5), 529–539.

- Porter, M. F. (1997), ‘An algorithm for suffix stripping’, pp. 313–316.
- Pryczek, M. & Szczepaniak, P. S. (2006), ‘On textual documents classification using fourier domain scoring’, *wi* **0**, 773–777.
- Walker, J. S. (2008), *A Primer on Wavelets and Their Scientific Applications*, second edn, Chapman & Hall/CRC.
- Witten, I. H. & Frank, E. (2005), *Data Mining: Practical machine learning tools and techniques*, 2nd edn, Morgan Kaufmann, San Francisco.
- Yang, Y. (1999), ‘An evaluation of statistical approaches to text categorization’, *Inf. Retr.* **1**(1-2), 69–90.
- Yang, Y. & Pedersen, J. O. (1997), A comparative study on feature selection in text categorization, *in* ‘ICML ’97: Proceedings of the Fourteenth International Conference on Machine Learning’, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp. 412–420.