Combining Content and Structure Similarity for XML Document Classification using Composite SVM Kernels

Saptarshi Ghosh and Pabitra Mitra
Computer Science and Engineering
IIT Kharagpur, India
saptarshi_bec@yahoo.com, pabitra@gmail.com

Abstract

Combination of structure and content features is necessary for effective retrieval and classification of XML documents. Composite kernels provide a way for fusion of content and structure information. In this paper, we demonstrate that a linear combination of simple and low cost kernels such as cosine similarity on terms and selective paths provide a good classification performance. We also propose a corpus-driven entropy-based heuristic for determining the optimal combination weights. Classification experiments performed on the INEX 1.3 XML corpus, demonstrate that the composite kernel classifier achieves significantly better performance as compared to complex and time consuming approaches.

1 Introduction

Algorithms for retrieval and classification of semi-structured documents can be broadly divided into three categories - (i) those that use general classifiers on an unstructured flat text view of the documents, without assigning any special significance to the tags or the structural information, (ii) those which take only the structure of the semi-structured documents into account, and (iii) the algorithms that take both the content and the structure into account.

A number of approaches exist for classification of unstructured text. Naive-Bayes and Support Vector Machine (SVM) based classifiers like [6] have proved to be quite effective for this purpose.

A number of measures have been proposed for measuring structural similarity between semi-structured documents like XML [1]. Some of the structure-only approaches for XML classification based on tree edit distances are of quadratic complexity and thus are infeasible for large corpus. Other approaches include rule-based approaches like XRules [9]. It is based on the premise that the presence of one or more particular type(s) of structural pattern(s) in an XML document can be used to predict the likelihood of its belonging to a particular category. In this approach, a set of 'structural rules' are learned during the training phase, by identifying a representative set of structural patterns for each category. During classification, the rules relevant to a new document are identified, and the statistics of the matched rules are used to predict the category of the new document. Flesca et. al.[3] uses the Fourier Transform technique to compute the structural similarity between two XML documents. They extract the sequence of start tags and end tags from the documents, and convert the tag sequence to a sequence of numbers to represent the structure of the documents. The number sequence is then viewed as a time series and Fourier transform is applied to convert the data into a set of frequencies. The similarity between two documents is computed in the frequency domain by taking the difference in magnitudes of the two signals. Kashima & Koyanagi [7] proposed a structure-only classification technique for semi-structured documents using Support Vector Machines, however their proposed kernel is of quadratic complexity w.r.t. the number of nodes in the DOM trees, hence it is impractical for use with large corpuses.

The recent trend in XML retrieval and classification, as exemplified by the INEX 2006 challenge [2], is to utilise both structure and content information. Content-structure classification techniques proposed over the years can be said to have two main approaches from a modeling perspective - one is to use different flat-text classifiers operating on distinct document elements (which are identified by the document structure i.e. the tags) and then combine these base classifiers to classify the entire document. The second family attempts to design new types of classifiers adapted for structured and

semi-structured documents.

The first approach has mostly been used for classification of HTML documents, using a combination of flat-text classifiers, each operating on the contents of a particular HTML tag. The second family of approaches include the HyREX system [4] which uses document substructures as indexing elements. A XML subtree indexing based approach has been adopted by Grabs et. al. [5].

In this paper, we propose the combination of structure and content information using composite support vector machine (SVM) kernels for supervised XML classification. The motivation is to isolate text and structure components, use simple and computationally low-cost content and structure similarity measures and then combine them using the flexibility offered by composite SVM kernels. The composite kernel is constructed by a linear combination of text and structure kernels, the combination weights are determined by a corpus-driven heuristic based on entropy of the document frequency distribution of the indexing units. We observe that this method is quite effective and achieves superior performance compared to other techniques.

We adopt the popular text kernel based on cosine similarity of term frequency vectors [6] to measure content similarity (on the extracted text). For structural similarity we use a selective path kernel having linear computational complexity. It is to be noted that other structural similarity measures may also be used to define composite kernels. However, we have observed that the selective path kernel provides comparable performance as compared to quadratic complexity algorithms e.g., the Kashima-Koyanagi kernel [7] for semi structured documents, while requiring several order less time even on moderate size collections.

Extensive experiments were performed on the INEX 1.3 corpus. The proposed algorithm was found to significantly outperform the reported results in the INEX 2006 challenge [2].

The remainder of the paper is structured as follows. The content and structure kernels used in the present work, and the technique used to combine the content and structure kernels to generate the composite kernel are discussed in Sect. 2. Section 3 briefly describes our experimental setup and the results achieved by the proposed classification technique. Section 4 summarizes the contributions of the current work and points out directions of future work.

2 Content and Structure Kernels

A linear combination of the content and structure kernel was used to construct a composite content-structure kernel for designing a supervised SVM classifier. Content similarity is measured by the common text kernel which uses cosine similarity between term frequency vectors.

2.1 Structure Similarity: Selective Path Kernels

We extract some of the paths from the DOM trees to represent the structure of an XML document. Only those linear paths are selected which extend from the root of the tree to a leaf node containing textual content; henceforth such paths are referred to as '*TPaths*'. The TPaths are treated as indexing elements for representing structure of the document. We use both Boolean and Cosine similarity models to measure the structural similarity among XML documents. In the cosine similarity model, the structure of an XML document is represented by a *structure-vector* whose elements give the $tf * idf$ score of the TPaths in the document. The structural similarity between two XML documents is computed by the cosine similarity between the structure-vectors of the two documents. This kernel method is henceforth referred to as the 'Cosine TPath' kernel. In the boolean similarity model (referred to as the 'Boolean TPath' kernel), the elements of the structure-vector are binary values indicating the presence or absence of TPaths in the documents.

2.2 Composite Kernel

A single numerical similarity score (in the range $[0, 1]$) between two XML documents is computed by a linear combination of the content similarity score and the structural similarity score. This is used as a composite kernel for the SVM. The composite kernels obtained by combining the 'Cosine TPath' kernel and the 'Boolean TPath' kernel with the text kernel are respectively referred to as 'Cosine TPath & Cosine Text' kernel and 'Boolean TPath & Cosine Text' kernel hereafter.

An entropy-based heuristic has been used to determine the relative weights in the linear combination. The method is described below. Let $t_1, t_2, ., t_i, ., t_k$ be the indexing units (terms or TPaths for text and structure kernel respectively). The document frequencies of these indexing units (i.e. the number of documents containing t_i-s) are represented by $df_1, df_2, ., df_i, ., df_k$. We normalize the df_i-s such that they have values in $[0, 1]$ and their sum is unity. Let these normalized document frequencies be denoted by p_i-s.

The entropy of the document frequency distribution

in the corpus is defined by

$$H = -\sum_i p_i \log(p_i) \quad (1)$$

Let H_s be the entropy of the distribution of document frequency of TPaths in the training corpus and H_c be the entropy of the distribution of document frequency of text terms. Then the composite kernel is defined as

$$\kappa_{\text{composite}}(x_i, x_j) = \frac{H_s}{H_s + H_c}\kappa_{\text{structure}}(x_i, x_j) + \frac{H_c}{H_s + H_c}\kappa_{\text{content}}(x_i, x_j) \quad (2)$$

where $\kappa_{\text{structure}}$ and κ_{content} are the structure and the content kernel respectively.

It may be noted that the entropy value will be higher for heterogeneous corpus with a varying distribution of indexing elements. Thus, if the structural variation among the documents of the corpus is large as compared to the content variation, more weightage is given to the structure kernel.

3 Experimental Results

Experiments were performed on the standard INEX 1.3 single-label categorization corpus (obtained from the website *http://xmlmining.lip6.fr/Corpus*). This corpus consists of 12107 XML documents, of which 6053 are used for training and rest as the testing corpus. The documents are the full-texts, marked up in XML, of articles of the IEEE Computer Society's publications from 12 journals and 6 transactions. The goal of the classification task is to classify the documents according to the source of the documents. The INEX 1.3 corpus is typically a text-rich and structurally homogeneous corpus. The documents show limited variations in the set of tags used - in the course of our experiments, the 6053 XML documents in the training corpus have been found to contain only 163 unique tags.

The text and set of TPaths extracted from each XML document are indexed using the SMART IR tool[8]. A standard set of stop-words are ignored during the indexing of the text. The kernel matrices are constructed from the index. Classification is then carried out using the popular LIBSVM implementation of Support Vector Machines, using the pre-computed kernel type. The results of the classification experiment have been summarized in Table 1. The performance measures considered are: micro-average and macro-average values for the precision, recall and F1-measure. The macro-average values are the mean values of the corresponding measure over all the classes in the corpus, and the micro-average values are the mean values weighted by the size of every class.

To compare the composite kernels, we have used the Cosine Similarity kernel on an unstructured bag-of-words model of the XML documents (without isolating the text and structural tags). This kernel is referred as the 'Cosine Bag-of-words' kernel. The results achieved have been shown in Table 1.

As can be seen from Table 1, the content-and-structure kernels, achieve higher values of precision and recall than that obtained using the 'Cosine Bag-of-words' kernel (i.e. using cosine similarity measure on tf-idf vector representations constructed using the entire XML document as an unstructured bag-of-words). This shows that measuring the structure similarity and the content similarity individually and then combining the two to define a composite similarity measure can produce better classification of semi-structured XML documents (than considering the structure information as part of the contents as in the unstructured bag-of-words model), and our approach using SVMs and kernel methods can utilize this property.

The proposed classification technique using composite kernels achieves significantly improved results as compared with other approaches reported in the XML Document Mining Challenge [2]. As given in the Results page of this challenge, a supervised learning approach on structure and content using Graph Neural Networks has been reported to achieve a micro F1 value of 0.721 and a macro F1 value of 0.714, compared to which the approaches proposed in this paper achieve substantially better results.

3.1 Choice of Kernel Combination Weights

As mentioned earlier, a linear combination of the content and structure kernels was used to generate the composite kernel. The effect of varying the relative weight of these kernels is shown in Fig. 1 for the INEX 1.3 corpus. It is observed that the highest value of F-measures is obtained when the relative weight for the content kernel is in the range $0.62 - 0.65$ (i.e. the weight of the structure kernel is in the range $0.35 - 0.38$). The relative weight for the content kernel obtained using the entropy-based heuristic on the INEX 1.3 corpus is 0.624. This value can be seen to be in close agreement with the experimental results. Thus the heuristic is effective for content-structure based classification of XML documents.

4 Conclusion

The problem of content-structure based classification of XML documents has two dimensions - (i) choice

Table 1. Results of single-label classification on INEX 1.3 corpus using the proposed classification technique and other techniques for comparison.

SVM kernel used	Precision		Recall		F1	
	Micro	Macro	Micro	Macro	Micro	Macro
Structure only kernels						
Cosine TPath	0.597	0.629	0.584	0.516	0.562	0.532
Boolean TPath	0.567	0.578	0.568	0.499	0.543	0.512
Using SVM on unstructured bag-of-words model of documents						
Cosine Bag-of-words	0.832	0.819	0.821	0.752	0.819	0.774
Composite Structure-Content kernels						
Cosine TPath & Cosine Text	0.862	0.886	0.857	0.842	0.857	0.861
Boolean TPath & Cosine Text	0.867	0.891	0.862	0.846	0.861	0.864
Other structure-content classification approaches submitted to the XML Document Mining Challenge [2]						
Unsupervised: Contextual Self-Organizing Map for Structures					0.135	0.085
Supervised: Graph Neural Network					0.721	0.714

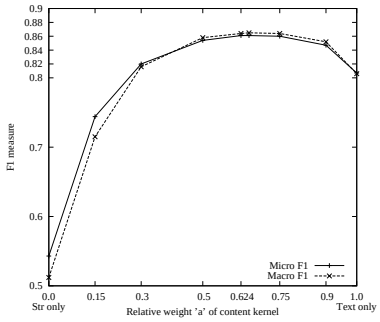

Figure 1. Variation of F1 measure of content-structure classification with a (the relative weight for the content similarity), composite similarity = a · content similarity + (1 - a) · structure similarity

of structure and content similarity measures or representation/ indexing schemes, and (ii) combination of these measures for retrieval/ classification. In this paper, we have used simple low-cost kernels which individually measure the structure and content similarities respectively. A linear combination of these kernels is used to generate a composite kernel finally used along with SVM for content-structure based classification. A corpus-driven entropy based technique was used to determine the relative weights in the linear combination. The values predicted by this technique was experimentally observed to be close to the values corresponding to the best classification performance.

Classification experiments performed on the INEX 1.3 XML corpus, demonstrate that the proposed classification technique achieves significantly better performance as compared to complex and time consuming approaches. In the future, we plan to further study algorithms for learning the optimal combination weights. One approach for this may be to use hyperkernels.

References

[1] D. Buttler. A short survey of document structure similarity algorithms. In *Proc. 5th Intl. Conf. Internet Comput.*, pages 3–9, 2004.

[2] L. Denoyer. XML document mining challenge. *http://xmlmining.lip6.fr/*.

[3] S. Flesca, G. Manco, E. Masciari, and L. Pontieri. Fast detection of XML structural similarity. *IEEE Trans. Knowledge Data Engg*, 17(2):160–175, 2005.

[4] N. Fuhr and G. Weikum. Classifi cation and intelligent search on information in XML. *Bulletin IEEE Tech. Com. Data Engineering*, 25(1):51–58, 2002.

[5] T. Grabs and H.-J. Schek. Flexible information retrieval on XML documents. In *Intelligent Search on XML Documents (LNCS 2818)*, pages 95–106, 2003.

[6] T. Joachims. Text categorization with support vector machines: Learning with many relevant features. In *ECML*, pages 137–142, 1998.

[7] H. Kashima and T. Koyanagi. Kernels for semi-structured data. In *ICML*, pages 291–298, 2002.

[8] G. Salton. *The SMART Retrieval System - Experiments in Automatic Document Processing*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 1971.

[9] M. J. Zaki and C. Aggarwal. Xrules: An effective structural classifi er for XML data. In *9th ACM SIGKDD*, pages 316–325, Washington, DC, 2003.