A negative category based approach for Wikipedia document classification

Meenakshi Sundaram Murugeshan*,
K. Lakshmi and Saswati Mukherjee

Department of Computer Science and Engineering,
College of Engineering, Guindy,
Anna University, Chennai-600025, India
E-mail: msundar_26@yahoo.com
E-mail: lakshmi_tamil@hotmail.com
E-mail: msaswati@yahoo.com
*Corresponding author

Abstract: Profile based methods have been successfully used for the classification of unstructured texts. This paper presents a profile based method for Wikipedia XML document classification. We have used profiles built using negative category information. Our approach exploits the structure of Wikipedia documents to build profiles. Two class profiles are built; one based on the whole content and the other based on the initial description of the Wikipedia documents. In addition, we have also explored the option of using the terms in the section and subsection titles. The effectiveness of cosine and fractional similarity measures in classifying XML documents is compared. The importance of combining two profile based classifiers is experimentally shown to have worked better than individual classifiers.

Keywords: XML classification; profile creation; negative categories; similarity measures; feature selection; initial descriptions.

Reference to this paper should be made as follows: Murugeshan, M.S., Lakshmi, K. and Mukherjee, S. (2010) 'A negative category based approach for Wikipedia document classification', *Int. J. Knowledge Engineering and Data Mining,* Vol. 1, No. 1, pp.84–97.

Biographical notes: Meenakshi Sundaram Murugeshan is a Research Scholar in the Department of Computer Science and Engineering, Anna University, Chennai. His areas of interests are in information retrieval and natural language processing.

K. Lakshmi received her PhD from Anna University, Chennai. Her research interests include text mining and information retrieval. She has published many papers on text classification and topic detection and tracking.

Saswati Mukherjee is Assistant Professor in the Department of Computer Science and Engineering, Anna University, Chennai. She works in the areas of information retrieval, natural language processing and distributed systems.

1 Introduction

Research on exploiting the structure of XML documents is gaining enormous popularity in recent years. Text classification techniques, which was earlier applied on plain text documents, is now taking a new dimension exploring different methods of classification that can take advantage of the structural information available in the XML documents. The task in classification is to classify the documents into predefined categories. INEX, which is a forerunner in the field of exploiting semi-structured data, promotes the research on XML documents using the Wikipedia corpus.

In this paper, we perform the task of classification as is defined in INEX 2007 XML mining track. To this end we propose to use some of the existing techniques along with the exploitation of the structural peculiarity of the Wikipedia corpus. We have made use of the Wikipedia corpus compiled by the INEX organisers and used by the XML mining track (Denoyer and Gallinari, 2007) of INEX 2007. This research aims on exploiting profile-based classification. The focus of the work is on improving the profile creation thereby improving the performance of classification.

Each Wikipedia document is an article about a topic. The article starts with an appropriate title related to the topic followed by a short overview of the topic. The main body of the article comes after this. The main body usually contains several sections, subsections regarding the topic and some links and references to other relevant documents. The information available in each of the structure is of varied importance. For example, the first paragraph can be taken as a summary of the article since it presents an overview of the content of the article. This we mention as initial description (IDES) in our paper. Further the links may be used to exploit any relevant information related to the article and so on.

The challenge in any Wikipedia document related research is to maximise the utilisation of the structural information available in such documents. In the task of classification of these documents into non-overlapping categories, such information may certainly be more helpful than only taking the documents as a whole. This research uses a combination of the IDES of the Wikipedia document along with the titles of the sections and subsections for classification of these documents. A classifier is built which utilises these information in conjunction with the whole document information and classifies a new document into a predefined class.

In all classifications, finding an effective way of document representation plays a very vital role. The method that we have adopted for the XML classification in this research is profile-based classification. Obviously, there exists many methods of appropriate representation of a document and hence a class in profile based classification (Salton and Buckley, 1988). Profiles are vectors, containing features and their weights, which are used as representation of documents in a category. In profile-based method, the profiles of the documents in training set are used to form profiles for each category. Since the total number of terms collected from the training set per category is large, dimensionality reduction is a necessary requirement. The features or terms that are to be included in the profiles are selected using feature-weighting schemes. Once the profiles of the categories are built, these can be used to classify a new document using a similarity metric to identify the correct category for a new, hitherto unseen document. This is the process of classification in a profile based approach. A document in the test set, which is

different from the training set, is used for the purpose of testing the efficiency of the classification. The profile of such a test document is built which is a vector of terms and weights as well. The similarity values between the test document's profile and the profile of each category are calculated. The document is classified to the category whose similarity value is the highest.

The most popular feature weighting schemes are the term frequency (TF) and average TF*IDF feature selection methods. The TF captures the importance of a term in a particular document, whereas, the IDF compensates for the appearance of the term across all categories. IDF provides the global weight of a term. The concept of IDF is based on the assumption that the lesser number of times a term appears in the whole collection, its contribution is more towards distinguishing the class to which it belongs.

Creating profiles using average TF and average TF*IDF have the drawback of failing to consider the distribution of terms over positive and negative categories. Various researchers have explored the possibilities of exploiting the category-wise information. Terms are given weights based on their presence in positive and negative categories in the relevance frequency (RF) method (Yang et al., 2002). Authors (Lakshmi and Mukherjee, 2007) have shown that the presence of the term in large number of negative categories is undesirable, a fact that can be only partially captured using IDF or RF. Additionally, when such negative documents are clustered in lesser number of negative categories, the power of contribution of the term to the positive category reduces considerably. This factor has been taken care of in the method of negative category document frequency (NCD). In their work, authors have shown the effectiveness of using NCD in profile creation for non-overlapping categories in an unstructured text corpus. Since our research deals with non-overlapping categories (i.e., a document belongs to only one category), we are motivated to use NCD based method for our profile creation.

Our research is divided into two parts. In the first part, we consider a Wikipedia document as a normal text document and build a profile and proceed with classification. The second part exploits the structure of a Wikipedia document. Since the IDES contains important aspects of the topic discussed in the document and acts as a kind of summary, our second approach relies on this description for classification. We have built a separate profile of a document and hence a class based on IDES and classified a new document. We have also explored the use of the titles given in a document. However, the numbers of terms in such titles are low and hence can not be used as such for classification. Hence we have explored the option of combining the titles profile along with the IDES profile.

This research, therefore, uses two profiles for classification. The whole content based profile uses NCD as feature weighting scheme while IDES profile uses TF and DF based feature weighting scheme.

Similarity measure is another important aspect in classification. The similarity measure compares the test document's profile with each category profile and gives a score which implies the relatedness of the test document with that particular category. The higher the score, the better is the relatedness. Since fractional similarity measure (Lakshmi and Mukherjee, 2006) has shown better performance in text classification on various datasets, we chose to use fractional similarity measure. The most widely used similarity measure is cosine similarity which is used as a benchmark for comparing our similarity measure score in this paper.

This paper is organised as follows. In Section 2, we describe the related work. Section 3 explains our approach. In Section 4, we show and discuss our experimental results and Section 5 concludes the paper.

2 Related work

Feature weighting schemes are divided into local and global feature weights according to the impact of the terms in the overall collection (Salton and Buckley, 1988). One of the fundamental methods of finding local weights is using the raw term frequency (TF). It identifies the number of times a term appears in a document. Global weight is typically calculated using inverse document frequency (IDF). The most popular method that is frequently used to capture both the local and global weights is TF*IDF weighting scheme which is shown in equation (1).

$$tfidf(t_i, d_j) = tf_{ij} * \log(N / n) \tag{1}$$

tf_{ij} term frequency of the i^{th} term (t_i) in j^{th} document (d_j)

N total number of document in the collection

n document frequency of the i^{th} term

The high TF terms in a particular document affects the overall average TF. To overcome this problem normalised TF is used. Normalised TF is shown in equation (2) below.

$$\log(1 + tf) \tag{2}$$

A term is deemed to contribute much to a category if it appears more in positive documents and less in negative documents. This intuition is captured using the relevance frequency (RF) (Lan et al., 2005; Yang et al., 2002). Equation (3) below shows the calculation of RF.

$$RF(t) = \log(1 + n_{i+} / n_{i-}) \tag{3}$$

n_{i+} number of positive documents that contains the term

n_{i-} number of negative documents that contains the term

Use of distributional features *viz.* compactness and position of the first appearance of the word for text classification is explored (Xue and Zhou, 2009). Dimensionality reduction for multi-label categorisation that selects features from multiple linear methods is shown in Park (2008).

Similarity measures play an important role in classification approaches. They are used for comparing two profiles; profiles of two documents or profile of a class with that of a document. Popular measures such as Dice, Jaccard and cosine similarity depend on the overlap of terms between two vectors (Van Rijsbergen, 1979; Lewis et al., 2006). Euclidean measure gives the distance between the given two vectors.

Cosine similarity between profiles of a category and that of a test document is calculated as given in equation (4).

$$\text{Cosine similarity } (P_i, d) = \frac{P_i \bullet d}{|P_i||d|} \tag{4}$$

where P_i is the profile of the i^{th} category and the numerator is the dot product of the category profile and test document's profile.

The structural information such as textual context of the links when used as features perform better than the features derived from the text of the web page itself which is

experimentally shown in Fürnkranz (1999). Larkey and Croft (1996) have shown that combination of classifiers produce better results than individual classifiers. The classifiers they used are K-nearest neighbour classifier, relevance feedback and bayesian independence classifier. All the two-way and three-way combinations showed better results than the individual classifiers.

3 Classification of Wikipedia documents

Profile-based classification pivots on the ability of creating a good representative profile for documents and hence for a class. Usually, profile of a document is built from the whole document. However, since Wikipedia documents have characteristics of their own, we have treated a document from two aspects, *viz*. the whole document and only the IDES. We have utilised the whole document for creating a classifier and eventually improved its performance by IDES classifier. We have also exploited the section and subsection titles and we have combined these with the IDES classifier to show the effectiveness of their use. The classifier that uses terms in IDES only for profile creation is called 'IDES alone' classifier and that which combines IDES terms along with title terms for profile creation is called 'enhanced IDES' classifier.

In profile based classification, we are required to create a profile (P_x) for a category X. This is done from the profiles of the documents belonging to category X in the training set. P_x serves as a representation of category X. This is repeated for all the categories in the whole corpus. For building profiles from the whole document and from the IDES, we have employed two different methods. For creating profiles from the whole document, NCD based feature-weighting scheme has been employed. Since TF appeals intuitively as an obvious method of building a profile, we have used a variant of this feature weighting scheme for the 'IDES alone' profiles. Since IDES contains domain dependent terms, we argue that if a term is present in the IDES part of most documents within that category in high frequency, it is most likely a representative term for that category.

For the creation of a profile of a new incoming document, we have used the popular TF*IDF feature selection method. Similarity measures are applied between the profile of an incoming document and category profiles. As is shown in the experimental section, a weighted average of the similarity measures of the two profiles gives the best result. The next subsections describe the methods used for building and comparing profiles.

3.1 Whole content based approach

3.1.1 Negative category document frequency

While creating profiles that represent a category, the contribution of a particular term on negative categories plays a vital role. If there are n categories, the $(n - 1)$ categories other than the correct category can be considered as negative categories. Reducing the weight of terms according to their presence in more negative categories, particularly if they are clustered in a few negative categories helps to form better profiles. NCD of a term t, which is shown in equation (5), reduces the weight of a term according to its distribution over negative categories.

$$ncd(t) = \log(1 + ncf / ndf) \text{ if } t \in \text{negative document}$$
$$= 1 \quad \text{if } t \notin \text{negative document} \tag{5}$$

where

ncf no. of negative categories the term appears

ndf no. of negative documents the term appears.

We have applied NCD along with average TF for profile creation. The feature weighting scheme used is TF*NCD.

3.2 Initial description based approach

Our next category profile has been created based on IDES of the documents belonging to the category.

'IDES alone' category profiles are created using those terms that have appeared at least in a minimum number of documents in the class. Average TF of all terms in the IDES of the documents to a category are retrieved and are filtered using a minimum document frequency (DF) to select the representative terms as given in equation (6). The threshold DF values have been empirically decided. We have used 10% of the total number of the documents present in the category as threshold. The profile thus created represents the common features present in a particular category.

$$IDES(t) = TF / n \text{ if } DF \geq \Theta \tag{6}$$

where

IDES(*t*) The term that belongs to the 'IDES alone' profile

TF The total number of times the term appears in all the IDES-s in that category

DF Total number of documents in the category the term appears

n Total number of documents in the category

Θ Threshold.

Titles of sections and subsections are unique to Wikipedia articles. However unlike IDES terms which are domain specific, these titles contain more general terms. So we relied on the TF*NCD method for creating profiles from this part. The number of terms in such parts is very low and even some articles do not contain any sections. Hence dimensionality reduction was not performed while creating profiles from Wikipedia titles. The titles profiles were combined with the IDES profiles and this 'enhanced IDES' profiles are used in conjunction with the overall whole content based profile for classification.

3.3 Similarity measures

The purpose of similarity measure in our work is to measure how similar the profile of a new document is with the profiles of each of the categories so that we can determine the category to which the new document belongs. The general form of the category score calculation that uses similarity score is given below.

Category score = Similarity measure (Category profile, Test document profile) (7)

Fractional similarity is a measure that has the advantage that it punishes a document if it contains terms that are not present in the category profile under consideration. In fractional similarity method, the higher the presence of terms in the document alone and not in the category profile, the lower is the similarity score for the document with respect to the category profile. Equation (8) below shows the calculation of fractional similarity measure between profile (P_i) and the document (d).

$$\begin{aligned} \text{Fraction } (P_i, d) &= \frac{\alpha}{\gamma} \quad \text{if } \{d\} - \{P_i\} \neq \phi \\ &= \alpha \quad \text{if } \{d\} - \{P_i\} = \phi \end{aligned} \tag{8}$$

where

$$\alpha = \sum_{k=1}^{n} W_k * V_k \quad \text{if } term_k \in P_i \text{ and } d$$

$$\gamma = \sum_{k=1}^{n} V_k \quad \text{if } term_k \notin P_i \text{ and } term_k \in d$$

W_k weight of $term_k$ in P_i

V_k weight of $term_k$ in document d

n number of terms in the P_i and document

3.4 Combination of two classifiers

The general form of the classification method that includes the scores obtained for a Test document by whole content based classifier along with 'IDES alone' or 'enhanced IDES' classifier is shown in Equation (9) below.

$$\text{Category score} = W_1 * AP \text{ score} + W_2 * IDES \text{ score} \tag{9}$$

where

W_1	Percentage weight given to whole content based TF*NCD classifier
W_2	Percentage weight given to IDES classifier
AP score	Similarity measure (TF*NCD profile, test document profile)
IDES score	Similarity measure ('IDES alone' or 'enhanced IDES' profile, test document profile)
Category score	The similarity score for a particular category

In equation (9) 'AP score' is the similarity score obtained while comparing the test document's profile with TF*NCD. Whole content based profiles and 'IDES score' is the similarity score obtained while comparing either the test document's profile built from IDES of the test document alone with the 'IDES alone' profiles of each category or the

test document's profile built from IDES and title parts of the test document with the 'enhanced IDES' profiles of each category. Category score determines the category to which the test document should be classified.

The values of the percentage weights (W_1 and W_2) are empirically decided. In spite of the descriptions in IDES being short, their effectiveness is quite high for the overall classification, which justifies our choice of IDES in combination with the whole content based profile for the overall classification.

Similarity score plays a key role in identifying the proper category of the test document. In our approach we have applied fractional similarity for comparing the profiles of test document and categories. We have used cosine similarity as our baseline method and compared our result of fractional similarity with those of the cosine similarity.

4 Experimental results and evaluation

The INEX corpus used for our experimentations is a subset of the Wikipedia corpus. The INEX corpus contains 96,611 documents. These documents are divided in 21 categories. Half of the documents are used for training and the other half for testing. We have used three forms of feature weighting schemes. The test documents are represented using TF*IDF weighting scheme, the IDES are represented using average TF, while the whole document is represented using TF*NCD scheme. Authors (Lan et al., 2006) have shown that profiles built using TF*RF feature weighting scheme have given consistent results over the TF or TF*IDF on various datasets. Hence we have chosen TF*RF as our baseline method and opted to compare our result of TF*NCD with this.

As is given in Equation (9) in the previous section, we have experimented with various values of W_1 and W_2. Of the various combinations we have tried out, the best result comes when we have combined the results of whole content based classifier and 'enhanced IDES' classifier.

The similarity scores of the whole content based TF*NCD profiles and the 'IDES alone' or 'enhanced IDES' profiles in the overall classification have been combined using these weights and the results were compared. Since dimensionality reduction is necessary, for the whole content based methods, we set a threshold of top 3% for profile creation (Rogati and Yang, 2002). We initially experimented with giving equal percentage weights for both classifiers. Of the various combinations we have tried, we have concluded that 90% W_1 and 10% W_2 provides the best results.

We found that for all ratios the combination of classifiers performed better than TF*RF method. Finally we also experimented with enhancing the IDES profile by combining the title features with it.

We have used the F-measure and accuracy values for the evaluation of our method since these are commonly used for evaluating the performance of text classifiers. We compared each test document against each of the 21 category profiles and classified the document to the category whose similarity value is the highest.

Table 1 below shows the F-measure and accuracy values achieved by applying TF*NCD profiles created from the whole content for classification. The comparison shows the performance of cosine and fractional similarities. From the results, we observe that fractional similarity performs marginally better than the cosine similarity.

Table 1 Evaluation results of whole content based classifier using NCD based approach

Category	*F-measure*		*Accuracy*	
	*TF*NCD cosine*	*TF*NCD fractional*	*TF*NCD cosine*	*TF*NCD fractional*
Music	0.6219	0.6119	0.9968	0.9967
Art	0.5984	0.5904	0.9357	0.9341
Sports	0.8434	0.8632	0.9528	0.9588
Astronomy	0.6816	0.6985	0.9926	0.9930
Aviation	0.7692	0.7603	0.9941	0.9939
Literature	0.7395	0.7670	0.9092	0.9189
Archaeology	0.7748	0.7781	0.9938	0.9939
History	0.5847	0.5620	0.9726	0.9710
Physics	0.6277	0.6480	0.9590	0.9613
Trains	0.7399	0.7668	0.9948	0.9953
Law	0.6842	0.7431	0.8419	0.8709
Pornography	0.8357	0.8198	0.9983	0.9982
Writing	0.8173	0.8130	0.9982	0.9982
University	0.4517	0.4581	0.9937	0.9938
Formula one	0.9000	0.8997	0.9975	0.9975
Comics	0.5460	0.5016	0.9942	0.9937
Chemistry	0.8489	0.8598	0.9855	0.9865
War	0.6781	0.6893	0.9849	0.9854
Sexuality	0.5403	0.5412	0.9768	0.9768
Spirituality	0.6351	0.6844	0.9798	0.9826
Christianity	0.8201	0.8196	0.9834	0.9833
Average	*0.7018*	*0.7083*	*0.9731*	*0.9754*

Table 2 shows the results of TF*RF method for the whole content based classification using cosine and fractional similarity measures. It can be seen that fractional similarity performs slightly better than cosine similarity. These results justify clearly the appropriateness of our choice of fractional similarity over cosine similarity.

Table 2 Results achieved by TF*RF classifier using cosine and fractional similarities

Category	*F-measure*		*Accuracy*	
	*TF*RF cosine*	*TF*RF fractional*	*TF*RF cosine*	*TF*RF fractional*
Music	0.6035	0.5835	0.9967	0.9965
Art	0.4508	0.4505	0.9120	0.9120
Sports	0.8287	0.8345	0.9485	0.9503
Astronomy	0.7161	0.7363	0.9935	0.9939
Aviation	0.7070	0.7005	0.9924	0.9923
Literature	0.7120	0.7140	0.8997	0.8998

Table 2 Results achieved by TF*RF classifier using cosine and fractional similarities (continued)

Category	*F-measure*		*Accuracy*	
	*TF*RF cosine*	*TF*RF fractional*	*TF*RF cosine*	*TF*RF fractional*
Archaeology	0.7391	0.7390	0.9929	0.9929
History	0.5467	0.5437	0.9702	0.9700
Physics	0.5511	0.5549	0.9505	0.9511
Trains	0.6171	0.6217	0.9923	0.9924
Law	0.5742	0.5743	0.8008	0.8009
Pornography	0.7577	0.7651	0.9975	0.9977
Writing	0.7118	0.7292	0.9972	0.9974
University	0.17297	0.1745	0.9905	0.9906
Formula one	0.8817	0.8736	0.9971	0.9969
Comics	0.5499	0.5443	0.9943	0.9942
Chemistry	0.8074	0.8112	0.9815	0.9819
War	0.5412	0.5533	0.9785	0.9792
Sexuality	0.4567	0.4554	0.9724	0.9724
Spirituality	0.5910	0.6008	0.9774	0.9780
Christianity	0.7863	0.7857	0.9801	0.9801
Average	*0.6334*	*0.6355*	*0.9674*	*0.9676*

Figure 1 combines the above two results and serves as a comparison amongst all the methods. The efficacy of the NCD based approach over the TF*RF method can be seen in this figure.

Figure 1 Comparison of evaluation results for methods based on NCD and the TF*RF method for whole content based classification (see online version for colours)

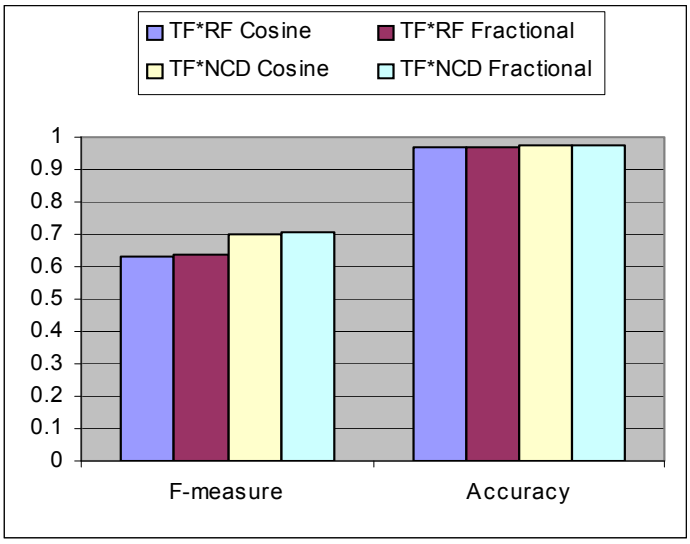

Table 3 below shows the comparison of F-measure and accuracy values for method based on combining the two profiles. It is evident that the use of 'IDES alone' or 'enhanced IDES' classifier with the whole content based classifier has helped to improve the F-measure and accuracy values. The 'enhanced IDES' profiles that include terms from titles have marginally improved the overall classification.

We observe that profiles created using IDES and title terms has helped in improving the overall classification performance, but if we increase the percentage weight given to the 'IDES alone' or 'enhanced IDES' classifier score (W_2) the F-measure and accuracy values decrease. We have concluded that 90% weight for whole content based classifier and 10% weight for 'IDES alone' or 'enhanced IDES' classifier as the best combination.

Table 3 Results achieved by the combination of two classifiers using fractional similarity

Category	*F-measure*			*Accuracy*		
	*Whole content (TF*NCD) 50% IDES alone 50%*	*Whole content (TF*NCD) 90% IDES alone 10%*	*Whole content (TF*NCD) 90% enhanced IDES 10%*	*Whole content (TF*NCD) 50% IDES alone 50%*	*Whole content (TF*NCD) 90% IDES alone 10%*	*Whole content (TF*NCD) 90% enhanced IDES 10%*
Music	0.6049	0.6434	0.6169	0.9967	0.9970	0.9968
Art	0.4413	0.5714	0.5820	0.9092	0.9313	0.9322
Sports	0.7905	0.8618	0.8564	0.9369	0.9586	0.9565
Astronomy	0.7154	0.7374	0.7844	0.9935	0.9939	0.9951
Aviation	0.6353	0.7399	0.7469	0.9907	0.9933	0.9935
Literature	0.6602	0.7510	0.7395	0.8812	0.9127	0.9026
Archaeology	0.6485	0.7614	0.7625	0.9903	0.9935	0.9935
History	0.3886	0.5455	0.5527	0.9593	0.9698	0.9704
Physics	0.5434	0.6421	0.6512	0.9493	0.9604	0.9613
Trains	0.7293	0.7863	0.7781	0.9946	0.9957	0.9955
Law	0.6328	0.7326	0.7336	0.8152	0.8658	0.8667
Pornography	0.7307	0.8208	0.8371	0.9973	0.9982	0.9983
Writing	0.7214	0.8268	0.8253	0.9973	0.9983	0.9983
University	0.4734	0.5036	0.5073	0.9941	0.9943	0.9944
Formula one	0.8571	0.9030	0.9117	0.9965	0.9976	0.9978
Comics	0.4455	0.5155	0.5355	0.9930	0.9939	0.9942
Chemistry	0.7511	0.8486	0.8534	0.9762	0.9854	0.9859
War	0.5992	0.6864	0.6938	0.9812	0.9853	0.9856
Sexuality	0.3813	0.5463	0.5421	0.9687	0.9771	0.9768
Spirituality	0.5169	0.6795	0.6831	0.9733	0.9824	0.9825
Christianity	0.7351	0.8153	0.8126	0.9756	0.9830	0.9827
Average	*0.6191*	*0.7104*	*0.7145*	*0.9652*	*0.9746*	*0.9743*

Figure 2 Comparison of F-measure values for all categories using different profile creation methods (see online version for colours)

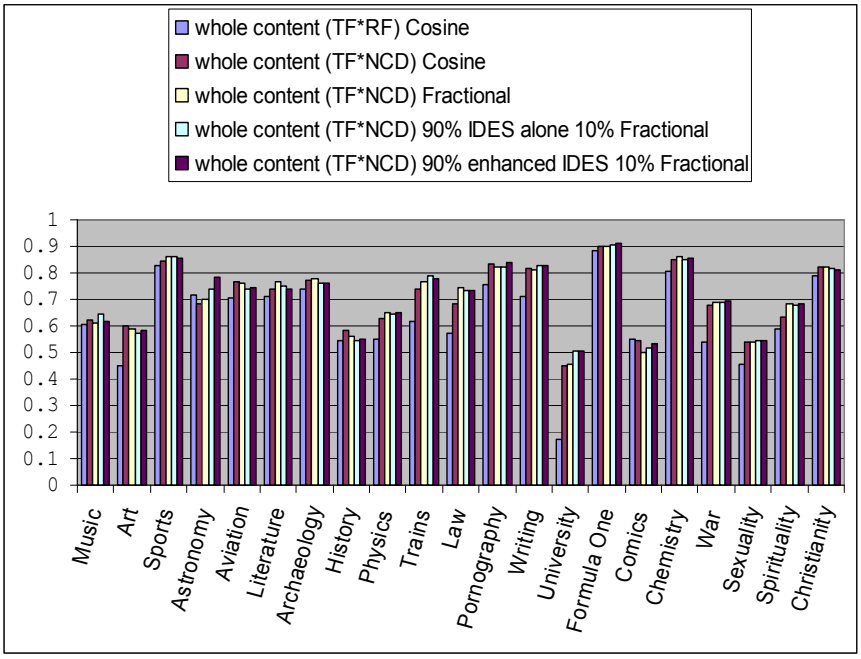

Figure 3 Comparison of Accuracy values for all categories using different profile creation methods (see online version for colours)

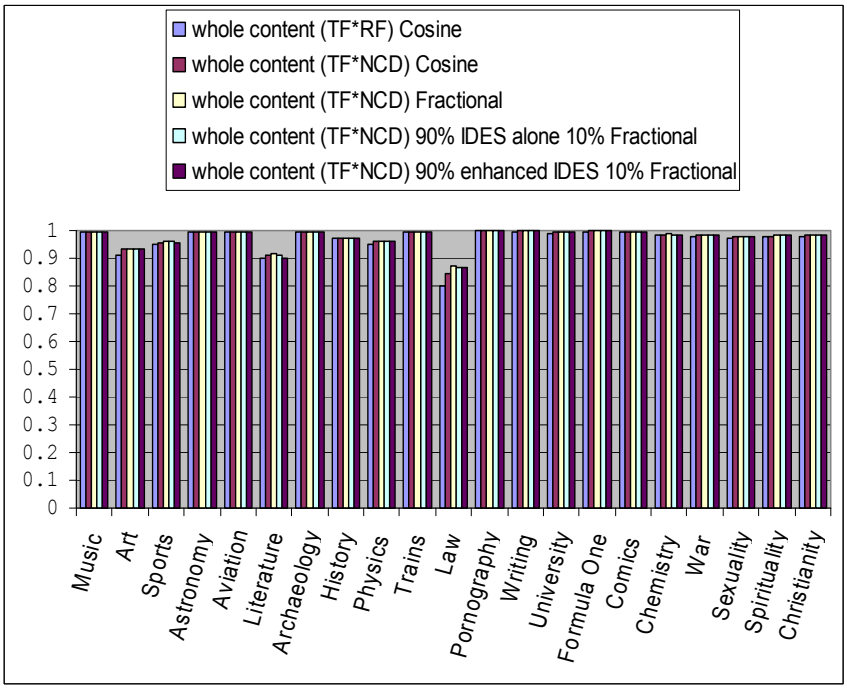

Figure 2 compares the F-measures for all categories using different profile creation methods. Out of 21 categories, the best results for 20 categories were obtained while using the NCD based methods. Only in the comics category the TF*RF method performed slightly better than TF*NCD method using the cosine similarity. In the 20 categories in which NCD based methods and combined classifiers have outperformed TF*RF results, the combination of classifiers showed the best results in ten categories, thereby improving the overall average F-measure which can be seen from Table 4. The whole content based TF*NCD methods which do not make use of the IDES information (i.e., TF*NCD 100%) obtained the best results in ten categories in which cosine and fractional similarity methods gave the best results in four and six categories respectively. TF*NCD based whole content classifier when combined with the 'enhanced IDES' classifier obtained better results in 13 categories over the method where it is combined with the 'IDES alone' classifier. However, the overall F-measure improved only marginally.

Figure 3 compares the accuracy values of TF*RF method with the NCD based methods and classifiers that use two profiles. The difference between all methods is marginal which is evident from the graph. Similar to the F-measure results, for the comics category the TF*RF results showed marginally better results than NCD based methods. We observe that NCD methods with fractional similarity performed better than cosine similarity.

Table 4 shows the comparison of average F-measure and accuracy achieved by different methods when we have considered all 21 classes. From the results it is evident that giving higher percentage weight to the NCD based classifier and less percentage weight to the 'IDES alone' or 'enhanced IDES' classifier helps the XML document classification immensely. In addition, it can be seen that combining the profiles of IDES and titles improves the performance in terms of F-measure.

Table 4 Overall F-measure and accuracy values for different methods discussed in this paper

Methods	*Average F-measure*	*Average accuracy*
Whole content (TF*RF) cosine	0.6334	0.9674
Whole content TF*RF fractional	0.6355	0.9676
Whole content (TF*NCD) cosine	0.7018	0.9731
Whole content (TF*NCD) fractional	0.7084	0.9754
Whole content (TF*NCD) 50% IDES alone 50% fractional	0.6191	0.9652
Whole content (TF*NCD) 90% IDES alone 10% fractional	0.7104	0.9746
Whole content (TF*NCD) 90% enhanced IDES 10% fractional	0.7145	0.9743

5 Conclusions

This paper presents a method of Wikipedia classification. Since NCD based profile creation proved to perform well for non-overlapping categories, we have experimented with this method, coupled with the method that exploits IDES and title terms for profile creation. The IDES of the Wikipedia documents which contain domain specific terms

helped to improve the performance of overall classification. Combination of two classifiers has shown better results than any of the classifiers taken individually. We also plan to extend this method, by exploring more Wikipedia specific structures such as links in a document.

References

Denoyer, L. and Gallinari, P. (2007) 'Report on the XML mining track at INEX 2007 categorization and clustering of XML documents', *SIGIR Forum*, Vol. 42, No. 1.

Fürnkranz, J. (1999) 'Exploiting structural information for text classification on the WWW', *Proceedings of the Third international Symposium on Advances in intelligent Data Analysis*, pp.487–498.

Lakshmi, K. and Mukherjee, S. (2006) 'An improved feature selection using maximized signal to noise ratio technique for TC', *Proceedings of the third International Conference on Information Technology, New Generations (ITNG 2006)*, 10–12 April 2006, Las Vegas, Nevada, USA.

Lakshmi, K. and Mukherjee, S. (2007) 'Category based feature weighting for automatic text classification', *Proceedings of 3rd Indian International Conference on Artificial Intelligence (IICAI-07)*, 17–19 December 2007, Pune, India.

Lan, M., Tan, C., Low, H., and Sung, S. (2005) 'A comprehensive comparative study on term weighting schemes for text categorization with support vector machines', *Proceedings of the 14th International World Wide Web Conference (WWW 2005)*, 10–14 May 2005, Chiba, Japan.

Lan, M., Tan, C., and Low, H.B. (2006) 'Proposing a new term weighting scheme for text classification', *Proceedings of the 21st National Conference on Artificial Intelligence (AAAI-2006)*, 16–20 July 2006, Boston, Massachusetts, USA.

Larkey, L. and Croft, W.B. (1996) 'Combining classifiers in text classification', *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 18–22 August 1996, Zurich, Switzerland, pp.289–297.

Lewis, J., Ossowski, S., Hicks, J., Errami, M., and Garner, H.R. (2006) 'Text similarity: an alternative way to search MEDLINE', *Bioinformatics 2006*, Vol. 22, No. 18, pp.2298–2304.

Park, C.H. (2008) 'Dimension reduction using least squares regression in multi-labeled text categorization', *Proceedings of the 8th IEEE International Conference on Computer and Information Technology*, 8–11 July 2008. pp.71–76.

Rogati, M. and Yang, Y. (2002) 'High-performing and scalable feature selection for text classification', *Proceedings of the 11th International Conference on Information and Knowledge Management (CIKM 2002)*, 4–9 November 2002, McLean, VA, USA.

Salton, G. and Buckley, C. (1988) 'Term-weighting approaches in automatic text retrieval', *Information Processing Management*, Vol. 24, No. 5, pp.513–523.

Van Rijsbergen, C.J. (1979) *Information Retrieval,* Second Edition, Butterworths.

Xue, X. and Zhou, Z. (2009) 'Distributional features for text categorization', *IEEE Transactions on Knowledge and Data Engineering*, Vol. 21, No. 3, pp.428–442.

Yang, S.M., Wu, X-B., Deng, Z-H., Zhang, M. and Yang, D-Q. (2002) 'Relative-term-frequency based feature selection for text classification', *Proceedings of First International Conference on Machine Learning and Cybernetics*, Beijing, November 2002, pp.1432–1436.