Understanding Temporal Aspects in Document Classification

Fernando Mourão, Leonardo Rocha, Renata Araújo,
Thierson Couto, Marcos Gonçalves, and Wagner Meira Jr.
Federal University of Minas Gerais
Computer Science Department
Av. Antônio Carlos 6627 - ICEx - 31270-010
Belo Horizonte, Brazil
{fhmourao, lcrocha, renata, thierson, mgoncalv, meira}@dcc.ufmg.br

ABSTRACT

Due to the increasing amount of information present on the Web, Automatic Document Classification (ADC) has become an important research topic. ADC usually follows a standard supervised learning strategy, where we first build a model using pre-classified documents and then use it to classify new unseen documents. One major challenge for ADC in many scenarios is that the characteristics of the documents and the classes to which they belong may change over time. However, most of the current techniques for ADC are applied without taking into account the temporal evolution of the collection of documents.

In this work, we perform a detailed study of the temporal evolution in the ADC, introducing an analysis methodology. We discuss that temporal evolution may be explained by three factors: 1) class distribution; 2) term distribution; and 3) class similarity. We employ metrics and experimental strategies capable of isolating each of these factors in order to analyze them separately, using two very different document collections: the ACM Digital Library and the Medline medical collections. Moreover, we present some preliminary results of potential gains that could be obtained by varying the training set to find the ideal size that minimizes the time effects. We show that by using just 69% of the ACM database, we are able to have an accuracy of 89.76%, and with only 25% of the Medline, an accuracy of 87.57%, which means gains of up to 20% in accuracy with much smaller training sets.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Information Search and Retrieval;; I.5.2 [**Design Methodology**]: Pattern analysis;

General Terms

Algorithms, Experimentation

Keywords

Text Classification, Temporal Analysis, Digital Libraries

WSDM'08, February 11–12, 2008, Palo Alto, California, USA.

1. INTRODUCTION

The widespread use of the Internet has been increasing the amount of information being stored and accessible through the Web. This information is frequently organized as textual documents and has been the main target of search engines and other retrieval tools, which perform tasks such as searching and filtering. A common strategy to handle such amount of information is to associate it with semantically meaningful categories, a process known as Automatic Document Classification (ADC) [4]. The class assignment may support and improve several tasks such as automated topic tagging (i.e., assigning labels to documents), building topic directory, identifying document writing styles, creating digital libraries, improving the precision of Web searching, and even helping users to interact with search engines. Other useful application scenarios include spam filtering, detection of adult content, and plagiarism.

ADC usually follows a supervised learning strategy, where we first build a classification model using pre-classified documents and then use the model to classify new unseen documents. Constructing text classification models usually means finding the set of characteristics that better identify classes of documents. One major challenge in several scenarios is that the characteristics of the documents and the classes to which they belong may change over time, since new information is created, new terms are introduced, new fields emerge, and large fields are divided into more specialized sub-fields. Vocabulary patterns that characterize a given document category evolve over time, distorting differences among areas which once were useful for separating them apart.

Recently, we were able to observe an interesting example of the impact of class evolution. Besides being the god of hell in Roman mythology, Pluto was also considered to be a planet until mid-2006. Up to this date, documents with the term Pluto had a higher probability of being classified in the astrophysics class, due to the great amount of references that mention Pluto as a planet. From this date on, since Pluto is not considered to be a planet anymore, there has been a significant reduction in the number of documents refering to it in this context. In mythology, however, the reference of Pluto did not suffer any sensible variation. Therefore, classification models created after this change must have a consequent increase in the probability that a document with this term is classified in the mythology class. Nevertheless, such models are only possible when the temporal aspects of documents are taken into account, what is not the case in most strategies for building classifiers.

Despite the potential quality reduction in classification models associated with temporal-related changes, most of the current techniques for ADC, such as nearest-neighbor, Bayesian, support vector machines, and association-based classification [22] are applied

without taking into account the temporal evolution of the collection of documents. Although it is recognized that the temporal evolution is a relevant factor in the context of ADC, it is still not clear how the temporal evolution influences the performance of the classifiers, what are the temporal-related characteristics that affect the classification's quality, and how to exploit such characteristics to improve the classification's accuracy. Answering these questions is one of the goals of this work.

There are some works on ADC that consider the temporal evolution of the collections, but they focus just on the challenge of treating the class frequency imbalance and how it changes over time, which we mention here as class distribution. In this work, we go beyond by considering two other factors that are associated with the temporal evolution. The second factor is how the relationship between terms and classes changes over time, as a consequence of terms appearing, disappearing, and having variable discriminative power across classes. The third factor is how the similarity among classes, as a function of the terms that occur in their documents, varies over time, for instance, two classes may be similar at a given moment, and not similar later in the future. In order to understand and characterize these three factors, we evaluate the temporal evolution and its effects on two datasets: the digital library of the ACM (ACM-DL) containing documents in computing, and Medline, a digital library related to medicine. We propose and apply a novel methodology for assessing the impact of the temporal evolution on the classifier's performance. Our methodology allows us to analyze separately each of the factors, pointing out the causes that result in the classifier's performance degradation. Finally, we present some results from an exhaustive empirical analysis that demonstrates the potential improvements by accounting for temporal aspects.

The remainder of this paper is organized as follows. Section 2 describes the related work. Section 3 details each of the factors in ADC and how the temporal evolution is related to these factors and to the sampling effect. In section 4 we present our methodology for understanding the temporal evolution and apply it to two digital collections. In section 5, we present some preliminary results that show how the understanding of the time effects can be used to improve the classification process. Finally, in Section 6, we conclude and discuss future work.

2. RELATED WORK

Although document classification is a widely studied subject, the analysis of temporal aspects in this class of algorithms is quite recent - it has been studied just in the last decade. We distinguish three broad areas where there has been significant efforts: adaptive document classification, adaptive information filtering, and concept drift. We discuss each of them in this section.

One area where temporal aspects has been studied in more detail is Adaptive Document Classification [7], which encompasses a set of techniques related to temporal and other aspects with the goal of improving the effectiveness and precision of document classifiers through their incremental and efficient adaptation [19]. Adaptive Document Classification brings a variety of challenges to text mining, discussed in the next paragraphs.

The first challenge is the notion of context and how it may be exploited towards better classification models. A context is a semantic meaningful partition of the data space, as a strategy to reduce the complexity associated with generating a classifier model. Previous research in document classification identified two essential forms of context: neighbor terms that are close to a certain keyword [17] and terms that indicate the scope and semantics of the document [6]. Most of the the studies focus on the first kind of context. However, determining semantically meaningful terms for each class has a great applicability. These terms may be used to define categories' contexts. Unlike the neighbor terms, which are extracted based on their proximity to a single term, context terms must be determined through the analysis of multiple documents from multiple categories.

The second challenge is creating the models incrementally, due to the fact that both, the content and the vocabulary of the collection, may evolve as new documents are added. As the collection evolves, the context that characterizes each class evolves too. Obviously, since the vocabulary may evolve, no context may be presumed. Even the context may evolve by covering all terms currently seen in the documents while inappropriate terms may introduce inefficiency and errors in the document classifier. Thus, the construction of an "optimum" context (if any) consists of a series of tuning processes, which may be re-triggered by the addition of a new document. Re-triggering the whole process of building a classifier for each document, however, is clearly computationally impractical.

The third challenge is the efficiency of the document classifiers. The use of a classifier is, most of the time, more frequently than its construction. The computational efficiency of the usage is thus essential to the process of managing information and knowledge, that is, adapting classifiers incrementally and efficiently is a key factor for improving the precision of classifiers.

Although there has been several efforts to identify new contexts and to deal with the challenges previously mentioned, none of these efforts is concerned about performing document classification in its original temporal context, that is, categorize the document according to the period of time to which it belongs. Besides, no works exploited the relation between this context evolution and the time itself, and the studies focus just on identifying the rising of a new context, and not relating these contexts to their chronological time.

Another field in which there is a concern about the temporal issues is Adaptive Information Filtering [11]. Information Filtering [3] is described as a binary classification problem in which documents are classified as relevant or not according to the user's interest. According to [8], the modeling of classifiers in static domains is sufficiently well controlled. Diverse applications assume that the training data distribution and the new data distribution are similar. This may be true for applications that last for a short period of time, but it becomes invalid to applications lasting long periods. In this case, it is unavoidable to adapt the classifier to new situations in order to maintain the quality of the classification.

In this context, Klinkenberg did some interesting research about Adaptive Filtering [15], where they explore methods to recognize concept changes and to determine windows of interest on the training data, whose size is either fixed or automatically adapted to the current extent of concept change. However, the effectiveness of the approach varied according to changes in the users' interest across time. Another relevant work is presented in [1]. In that work, the authors verified the incremental relevance of the users' answers to the information filtering process also considering changes in the users' interest. J. Allan also evaluated dynamic aspects in topic detection and tracking [2] In [23], the authors developed a new technique, called *Margin-based Local Regression*, which automatically optimizes the classifier accuracy across time, based on how important a partial answer is in the document retrieval system in chronological order.

In [14], the authors show that taking temporal aspects into consideration when generating results for queries is very important and may improve the quality of the results given. By analyzing the timeline of a query result set, it is possible to characterize how temporally dependent the topic is, and how relevant the results are

likely to be. They conclude that meta-features associated with a query may be combined with text retrieval techniques to improve the understanding and treatment of text search on documents with timestamp. This work is somehow similar to ours in the sense that it investigates how the temporal aspect can be used to improve the quality of a certain proccess. However, while we concentrate in the classification process, they explore it for information retrieval.

Considering the research that addresses text changes, [16] bases their work on the idea of guaranteeing the precision of the classifier. Basicaly, there are two strategies for adapting a classifier. The classifier may be completely retrained according to the current training set, or an existing classifier may be updated only based on some examples. In the article, the difficulties associated with each technique are discussed and, due to this fact, the work was restricted to only detecting text changes, through precision metrics, not dealing with this adaptation phase. Lewis [18] did some work on autonomous systems for text classification, where he suggests behavior monitoring in order to adapt to changes, if necessary.

Concept or topic drift comprises another relevant set of efforts [10, 21]. In [10], the Daily Classification Task (DCT) is presented, a framework that deals with document classification in order to skirt the temporal aspects. In that work, time is discretized into periods, e.g. days. For each time period, a limited size random sample is provided as a labeled training set. A performance metric, such as classification accuracy or F-measure, is calculated for each day of the benchmark dataset, and the average is reported over all days. In [21], the authors propose a boosting-like method to train a classifier ensemble from data streams. It naturally adapts to concept drift and allows to quantify the drift in terms of its base learners. The algorithm is empirically shown to outperform learning algorithms that ignore concept drift. In both works, the authors consider the existence of temporal effects without exactly understanding what these effects are. They discuss that despite the fact that some classes present popularity explosion in some periods, their true identity does not modify. We, on the other hand, present a detailed explanation of the temporal aspects based on empirical results showing that there are other effects beyond the explosion of popularity of certain classes. Moreover, we present an exaustive empirical analysis that demonstrates how we can improve the classifiers by using these temporal aspects.

3. TEMPORAL EFFECTS

ADC usually follows a supervised learning strategy, where we first build a classificiation model using pre-classified documents and then use it to classify new unseen documents. Constructing text classification models usually means finding the set of discriminative terms that better identify classes of documents, during which several challenges, that are not necessarily temporal-related, must be addressed. We discuss some of them next.

The first challenge is to deal with the class distribution, that is, some classes are much more frequent than others, making it more difficult to avoid the generation of biased models. The second challenge is related to the discriminative power of the terms in a specific class, which we call term distribution. In this case, we can see that some terms are more discriminative in some classes than others. The third challenge is the overlap among classes with respect to the terms that are associated with more than one class, which we call class similarity. The larger is the overlap, the harder is to generate a good model that distinguishes them. It is important to note that the intensity of each challenge is usually negatively correlated with the accuracy and generality of the model.

As we discuss in the next section, the temporal evolution may worsen each of the aforementioned challenges, in the sense that models based on the whole collection may not work well, and accounting for its evolution is key for an effective classification. We also show that the temporal evolution may be a clear and definite trend or (apparently) a random variation, and it increases the difficulty in building accurate classifiers, affecting the classification task significantly. Next we revisit each of challenges from the point of view of the occurrence of temporal evolution.

The impact of the temporal evolution on the class distribution is its variation across time, which should be quantified to avoid bias while addressing the other challenges. Notice that classes may appear and disappear, which may happen as a consequence of splits and joins, respectively, of existing classes. For example, the sub-classes Information Retrieval and Artificial Intelligence in the ACM-DL computing classification scheme from 1964, belonged to the same class: Applications. In the new ACM-DL classification schema, each of them belongs to a diferent class, Information Systems and Computing Methodologies, respectively. The effect of the temporal evolution on the class distribution should be qualified and quantified, so that we are able to account for it in the model.

From the perspective of the temporal evolution, the second challenge has to do with the evolution of the term distribution. We qualify and quantify when terms appear, disappear, migrate among classes, or simply lose or gain discriminative power. In order to illustrate this, we can use the same example of section 1 related to Pluto. In this case, the term Pluto lost discriminative power in class astrophysics and gained in class mythology. Intuitively, we may state that evolution usually happens gradually, so that time periods that are close timewise are also more similar.

The third challenge concerns the evolution of the similarity among classes, without considering the reasons behind the intensity of the degree of similarity. For example, some time ago the classes crime and biology were not very similar but recently they became more related once DNA analysis started to be used in criminal investigation. Class similarity quantifies how different or related two classes are. We distinguish two types of class similarity: global and local, which arise when considering temporal evolution. The global similarity considers the whole collection, while the local similarity considers just a portion of it. Notice that dealing with local similarity may be hard because of not only its variation, but also the variable granularity it may assume. For instance, two classes may be similar considering a 10-year period, but may have become more different in the last two years of that interval. It is important to note that the last two challenges use the differences in the occurrence of terms within a class, that is, term distribution considers the differences within the same class (intra-class characteristics), while class similarity considers the differences between distinct classes (inter-class characteristics).

One clear trade-off while accounting for the temporal evolution is to find the proper granularity in terms of the time interval while building the models. The first issue associated with this trade-off is the sampling effect in the context of the temporal evolution of the collection. The sampling effect arises as a consequence of the popular strategy of building the model based on a sample of the document collection, and the fact that this sample may not have enough discriminants or distort the classes or the term distribution. For instance, if we look at the overall class distribution, it may seem to be balanced, but as we check the distribution in specific time periods, the sampling effect may arise. The occurrence of sampling effect suggests that we should increase the size of the sample, while the temporal evolution induces the reverse, since using long time periods may result in conflicting evidence. Determining the proper time granularity, that is, the time period during which the classification concepts are quite stable and the sample is large enough, is a

hard task and is a function of the collection characteristics. Our approach here is to first understand the trade-off by factoring out each of the factors that make part of it, and then try to investigate it towards a more accurate model. We are not aware of any other work that performs a deep investigation of this trade-off, as we discussed in Section 2.

In the next section we present our methodology for understanding the temporal evolution and apply it to two digital collections that span through decades.

4. CHARACTERIZING SAMPLING AND TEMPORAL EFFECTS

This section describes the analysis of the two factors that contribute to the trade-off previously discussed: the sampling effect and the temporal evolution. First, we briefly discuss the sampling effect in order to acknowledge its existence and its influence in ADC. Moreover, this analysis is relevant because it shows that it is necessary, for studying the temporal evolution, to isolate the sampling effect in the experiments, so that we remove its influence and are able to analyze just the temporal evolution. Therefore, after studying the sampling effect, we characterize and present a methodology for qualifying and quantifying the temporal evolution that may influence the performance of automatic classifiers. In order to do this, we analyze each of the challenges previously presented from the temporal evolution perspective, which we call time effects.

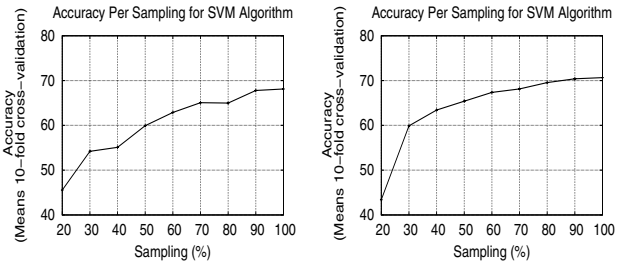
To demonstrate the existence of the temporal effects that may influence the performance of automatic classifiers, we performed a series of experiments using a Support Vector Machine (SVM) [12, 13], a state of the art classifier, using two document collections. The first document collection is a set of 30,000 documents of the ACM digital library containing articles from year 1980 to 2002. In this collection, classes of the ACM classification scheme are assigned to documents by their own authors. We use only the first level of the taxonomy, with 11 categories, which has not changed over the time period considered. All documents are assigned to a unique class. The second document collection is MedLine, which consists of 4,112,069 documents, with articles from 1970 to 2006, classified into 7 distinct classes.

4.1 Sampling Effect

Although the sampling effect is a well studied phenomenon, for completeness reasons, we have conducted a set of experiments to illustrate the impact of this effect on classifiers. Specifically in our case, we need to isolate it from the time effect. In order to do so, we built a document set C with documents belonging to just one year, more specifically 1999 for ACM-DL and 1970 for Medline.

From this new set C, we got a subset S, whose size is $X\%$ of C, to be used in the classification process. Then, we divided this subset S into ten parts, each part containing the same number of documents. Using this partition strategy, we performed a 10-fold cross-validation [5], varying the value of X from 20% to 100%, since all data belong to the same year, in order to analyze how the accuracy in the classification process is affected by the size of the sample. Figure 1 shows the results of these experiments. We can see that, as we increase the size of the database, the performance of the classifier gets better, as expected.

In both collections, the sampling effect is also influenced by the variability of documents throughout the years. In the ACM collection, the average number of documents per year, in the interval from 1980 to 2002, is 1,086.5, with a standard deviation of 550.8 documents. The year with the smallest number of documents is 1980, with 441 documents, whereas the year 1999 is the one with



(a) ACM Collection (b) Medline Collection

Figure 1: Sample Effect

the greatest number of documents, with a total of 2,728 documents. In Medline, the average number of documents per year is 100,789, with a standard deviation of 51,649.3 documents. The minimum number of documents per year is 34755 and it occurs in year 1970, whereas the maximum number is 208,878 and it occurs in year 2005.

These results demonstrate that it is impossible to analyze the time effect without isolating the sampling effect, which plays an important role in ADC. Moreover, when constructing a classification model, the sampling effect must be taken into account.

4.2 Time Effect

This section is divided into two parts. First, we present a set of experiments that were performed in order to demonstrate the existence of the time effect and its influence. Then, we analyze each of the challenges mentioned in the previous section from the temporal evolution perspective. It is important to notice that, for each of the challenges, we isolate the effects of other characteristics that might influence the classification task as well.

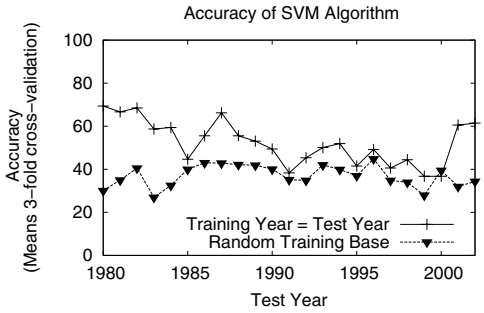
4.2.1 Quantifying the time effect

In order to demonstrate and quantify the temporal effect, we conducted two sets of experiments. In the first one, we divided the original collection in a per year basis using the same number of documents, selected randomly within each single year. We used the minimum number of documents found in any given year, i.e., 441 for ACM-DL and 34755 for Medline. By using the same number of documents for all years, our intention was to isolate the time effect from the sampling effect so that we could analyze the former in isolation. Next, we applied a 3-fold cross-validation classification process for each year.

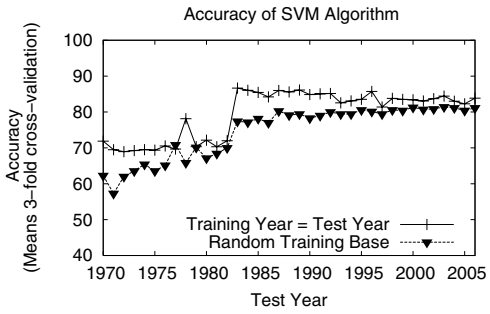
To draw a contrast with this result, we randomly selected from the entire collection the same number of documents (441 and 34755) in both collections to form another set of documents. Then, we randomly divided this set into three parts for performing a 3-fold cross-validation, where we created a classification model for each of the possible 2-part combinations among the three parts generated. It is important to point out that while in the first experiment the training set consisted of documents of specific years, in the second experiment there are documents from any given year in the training set. After, we tested each model built with a set of documents belonging to a specific year A_i created in the first experiment. We varied i to cover all years in the collections and calculated the average of the performance of the three models obtained with cross-validation for each year.

Figure 2 shows the comparison of the results obtained from these experiments. We can notice that, for most of the years in both collections, the accuracy of the classification task considering only

documents of the same year of the documents being tested is better than the accuracy of the classification task considering documents chosen randomly. Notice that the latter is the experiment normally used in classification experiments.



(a) ACM Collection



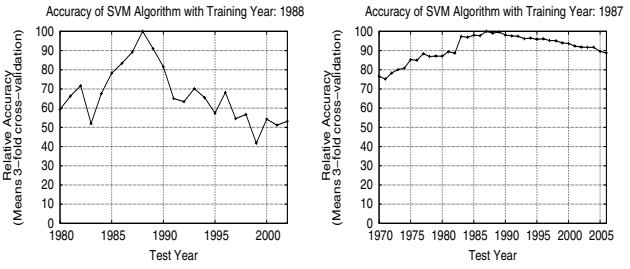
(b) MedLine Collection

Figure 2: Temporal Effect Means

We further characterize the time effect using a different strategy. In this second set of experiments, as it was done before, we divided the original collection in a per year basis and we used the same number of documents for each year. Again, we randomly divided each year A_i into three parts of same size. We used two of the three parts as our training set to generate the classification model. This model is evaluated over all the documents from each year A_j of the collections, in which $j \neq i$. Besides, we also tested the model with the remaining portion of A_i that was not used as the training set. We repeated this process for each of the three combinations using a 3-fold cross-validation. After the execution of the three models generated from A_i, we calculated the average of the performance of the models used in each year being tested. Then we ranked the resulting accuracy values so that the highest value obtained becomes 100 and the other values are calculated relative to it.

Figure 3 shows the results when using 1988 and 1987 as the training years, respectively in the ACM and Medline collections. We can see that there is an accuracy peak in the year in which the dataset was trained or years temporally close to it. There is also a visible degradation in performance when the test year becomes more distant from the training one.

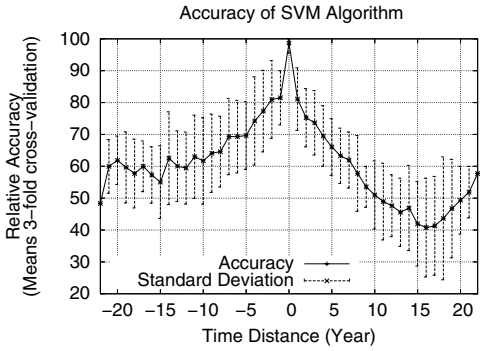
Figure 4 summarizes the described experiments for both ACM and Medline. For each year of training, we obtained the test results after cross-validation and ranked them as before. We then computed for each distance in time (relative to the training year) the mean performance, which corresponds to the dark points in the



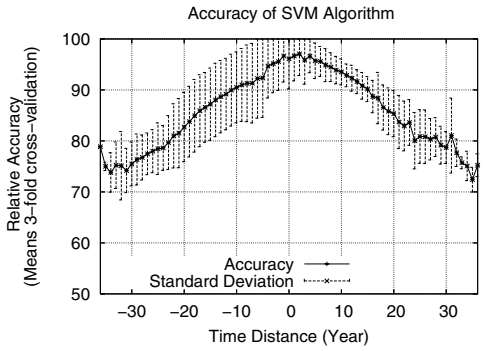
(a) ACM Collection (b) Medline Collection

Figure 3: Temporal Locality

graph. We also plotted for each mean the interval determined by the standard deviation. A positive distance in time, say 10, in the x axis, means that the test documents used are 10 years newer than the training documents. As can be seen, in most of the cases, the best scenario is found when the training and the test documents belong to the same year (interval zero). Only in very few cases, slightly better performance was achieved using the training and the test documents of distinct years. However, even in these cases the best performance was achieved using the training and the test documents close in time.



(a) ACM Collection



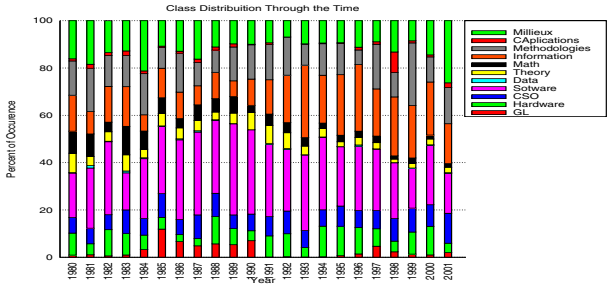
(b) MedLine Collection

Figure 4: Temporal Locality Variation

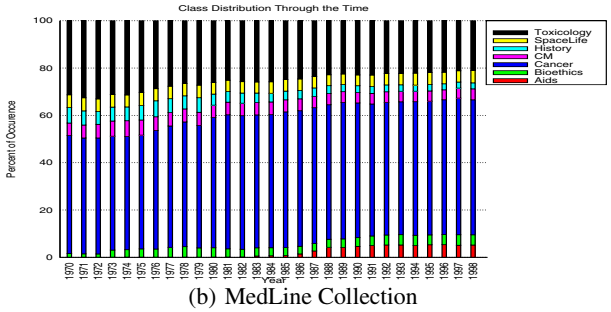
The described experiments present strong evidence of the existence of the time effect as a factor that contributes to the degradation of the performance of automatic classifiers. Following, we present and analyze the three challenges that may contribute to this degradation, as mentioned before, namely: class distribution, term distribution, and class similarity.

4.2.2 Class Distribution

In this section, we analyze in more detail the temporal effect in the class distribution. Figure 5, where we plot, for each year, the class probability distribution, illustrates the variation in terms of the representativeness of the classes across time for the ACM and the MedLine collections. As can be seen, the oscillation of the occurrence of the classes over the years is a frequent phenomenon. For example, there are documents in class *Aids* dating from 1963, but the class only became significant after the year 1985. Other classes, such as the class *Mathematics* of ACM, became less frequent as time went by.



(a) ACM Collection

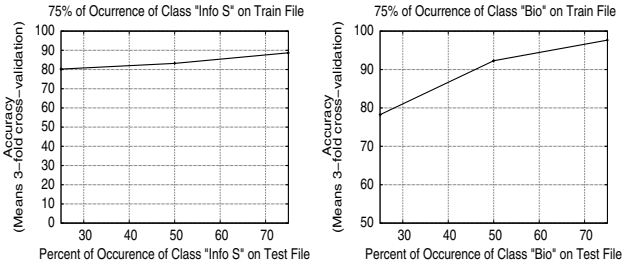


(b) MedLine Collection

Figure 5: Class Occurrence Variation Through the Time

To show the effect of this oscillation (i.e., the class distribution) in isolation, we conducted the following experiment. We randomly selected the same number of documents from two different classes (C_k and C_l) in a certain year for each collection to create other sub-collections. Using these new sub-collections, we generated subsets T_X of the same size. These subsets were created using the following procedure: $X\%$ of the documents of T_X are randomly selected from a certain class C_k and the remaining documents of T_X are randomly selected from the other class (C_l). We considered the possible values of X to be 25, 50 and 75. Then, each of the subsets T_X was randomly divided into three different parts. For each possible combination of two of the three parts, we generated a classification model. Each model created was tested with the other subsets T_X, besides being tested with the remaining part of the same subset T_X from which it was created.

We conducted this experiment in both collections and the graphs in Figure 6 show how the accuracy increases as the document distribution of a certain class in the test phase is closer to the document distribution of the training set. Therefore, it is clear that the document distribution among classes influences the performance of the classifiers.



(a) ACM Collection (b) MedLine Collection

Figure 6: Class Distribution

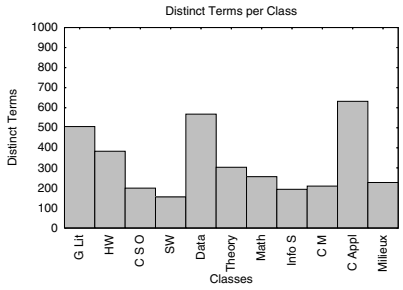
These results show us that we have to be very careful when creating classification models, because we can generate a biased model that may not be accurate for the data set to be tested. When we consider that the frequency of the classes are constantly changing over time, this becomes an even bigger problem that must be taken into account. Therefore, this challenge is very important in the process of constructing a classification model.

4.2.3 Terms Distribution

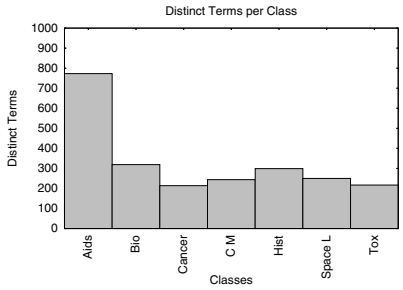
In this section, we analyze the second challenge: the evolution of the term distribution. This challenge occurs due to the movement of terms in a certain class over time and is associated with intra-class characteristics. Terms appear, disappear, migrate among classes and become more or less discriminative for a class in distinct periods, and that should be quantified and qualified so that we are able to account for it in the model.

In order to measure this effect, we created a vocabulary $V_{k,i}$ containing the t words with the highest info-gain [9] in class C_k in a given year A_i. They can be considered the most discriminative terms of the class for a given year. We represent this vocabulary in a vector space where each position in the vector represents a term and contains the weight of that term based on its frequency in class C_k in year A_i. Next we compute the union of all vectors of the same class C_k for all years. This way, we can evaluate how constant the vocabulary of each class is over the years. In theory, for a given class C_k, if all of its $V_{k,i}$ vocabulary vectors were constant (i.e., they did not change over time), the union vector for C_k would contain exactly t words. On the other hand, if all the $V_{k,i}$ for C_k were completely different from each other, then the size of the union vector would be t times the total number of years. For ACM, we used $t = 50$ and for the Medline collection we used $t = 100$. The union vectors for each class in both collections are shown in Figure 7. As can be seen, the movement of terms is distinct for each class, which means that there are classes that are more dynamic than others in both collections.

To better characterize the term distribution phenomenon, we performed the following experiment: we divided the database in a per year basis and, for each year, we separated the documents according to their respective classes. For each class of a certain year, we created the vocabulary $V_{k,i}$, which is formed by the t words that have the highest values of info-gain [9] in the class C_k in the year A_i. Then, we represent this vocabulary as a vector and verified the cosine similarity [20] between vocabulary vectors $V_{k,i}$ and $V_{k,j}$ of a same class over all years in the collection. Figure 8 shows that the vocabularies from years close to each other tend to be more similar



(a) ACM Collection



(b) MedLine Collection

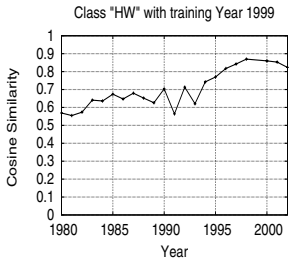
Figure 7: Analysis of the movement of terms in each class

in both collections[1]. The more distant in time two documents of a certain class are, the smaller is the probability of having common terms between them. Besides, this result is empirically recurrent, From these graphs, we can see that even when a certain class continues to exist (that is, neither the class disappears, nor it is divided, nor it aggregates other classes), its subject or main characteristics may change over time. For instance, although the artificial intelligence class is present in the ACM classification scheme since its inception, probably over a certain period of time the most widely studied subject in this area was neural networks while a long time ago it was first-order logic. It is also interesting to notice that the classes $Hardware$ in the ACM collection and the class $Aids$ in MedLine have quite different behaviors. While the cosine similarity curve of class $Hardware$ ("HW") presents a smooth decline, the same curve for class $Aids$ presents a strong decline. That is, the vocabulary of the class $Hardware$ changes slowly over time whereas the vocabulary of the class $Aids$ is much more dynamic. As a result, time has greater impact on the classification of documents of the class $Aids$ than in the classification of documents of the class $Hardware$.

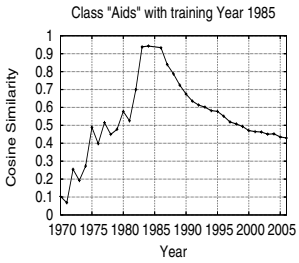
Figure 9 shows the mean cosine similarity of classes when we vary the distance in time among the distinct vocabulary vectors of the classes in both collections. Distance zero means that we are comparing a vocabulary to itself, which obviously corresponds to the maximum similarity. We can observe that the greater the distance in time is between the vocabularies, the less similar they are, which demonstrates an evolution of the vocabularies of each class over the years. An interesting example is the class $Aids$ of the Medline collection. Different from the other classes, the difference in similarity among its vocabulary vectors declines very fast, showing that the class $Aids$ is very dynamic.

As we demonstrate that the vocabulary evolves, it becomes obvious that a classification model generated considering documents

[1]The graphs of Figure 8 do not contain the value of the cosine similarity for the training year, since this value is always equal to 1.



(a) ACM Collection



(b) MedLine Collection

Figure 8: Terms Distribution

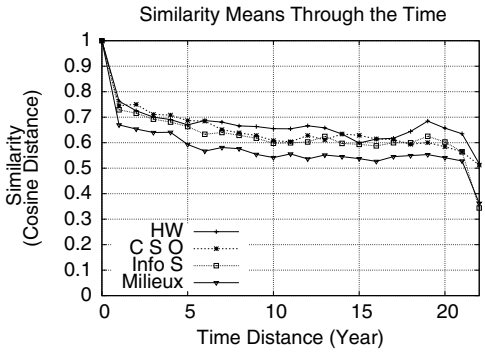
of a certain period of time may be less effective when test using documents from another period of time, since the vocabulary may have evolved in a way that the premises constructed are no longer true, that is, the discriminative terms may not be the same. That makes it a very interesting challenge as well.

4.2.4 Class Similarity

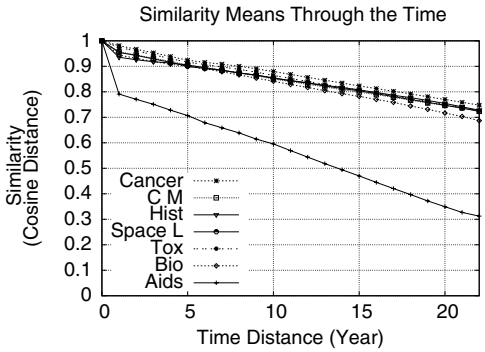
Finally, in this section we analyze the last challenge: the evolution of the class similarity. This challenge is interesting due to the migration and variation of the frequency of the terms in the vocabularies of the classes over time. It is important to note that it is correlated to the inter-class characteristics.

To analyze this phenomenon in more details, we conducted the following experiment. As we did in the previous experiments, we divided the collection in a per year basis and we randomly selected the same number of documents per class in each year. Also, as done before, we created a vocabulary $V_{k,i}$. To simplify the analysis of the experiment, we considered just two classes and we calculated the cosine similarity between their vocabulary vectors in a certain year A_i. Varying A_i for all years in the collections, we obtain the graph shown in Figure 10. As can be seen, the vocabulary vectors of these two classes are more or less similar in different periods of time. For instance, the classes $Data$ and $Mileux$ of the ACM collection are less similar to each other in the period from 1996 to 1998. This indicates that if we were going to classify documents within this interval using as our training set just documents from this period, the classifier performance must be better than when using random documents. This is due to the fact that if the vocabularies used to train the classifiers for the two class are different of the ones in this period, the two classes tend to be considered more similar than they really are in this time-period. Similar observations are valid for classes $Aids$ and $Toxicology$ of the Medline collection. Figure 10 shows that these classes are becoming more similar to each other since the 70's. Consequently the classification of documents of both classes is easier to be done for documents of the 70's and the 80's than for documents of year 2000.

Tables 1 and 2 show the variation over time in the similarity between each pair of classes for MedLine and ACM, respectively. A value in the table represents the standard deviation from the mean of the similarities between the corresponding pairs of classes in all years. As we can observe, for some pairs of classes the variation of similarity is very high. For example, the standard deviation for the mean similarity between class *Complementary Medicine* (CM) and class *History* (Hist) is 21%, very high. It means that the two classes may have been very similar in some periods, and loosely related in others. Consequently, the difficulty in separating the two classes also varies considerably over time. Another interesting observation is that the class *Cancer* of MedLine differs from others since

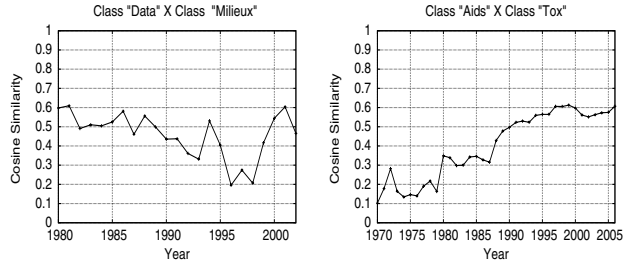


(a) ACM Collection



(b) MedLine Collection

Figure 9: Terms Distribution Means



(a) ACM Collection (b) MedLine Collection

Figure 10: Cosine Similarity Through the Time

it maintains low variability of its similarity to all other classes over the years. This indicates that time affects less this class than the other classes of MedLine.

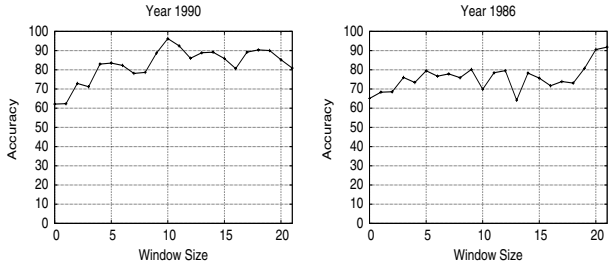
In summary, based on the characterizations presented above, it was possible to provide evidence of the class and term variation over time. We also checked that the similarity among these classes varies as well. Moreover, we found that these phenomena have great influence in the performance of the classifiers.

5. EXPLORING THE TEMPORAL EFFECTS

In this section, we present some preliminary results that show that the appropriate analysis and understanding of the time effects may lead to improvements in the classification process. The idea is not to propose a method for exploring temporal information but to show that there is room for the development of such methods as a way to increase the classifier performance.

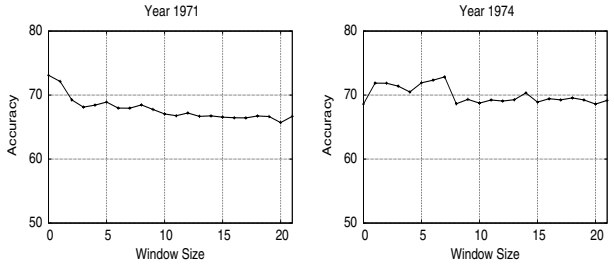
Improvements can be obtained by considering both the sampling and the time effects. The challenge, though, is to deal with these two effects simultaneously, because they demand opposite strategies. Simply increasing the size of the training set with no criteria may not produce improved performance since it normally implies in increasing the amount of data from years not temporally close to the test documents which, as discussed in the previous section, may introduce noise. On the other hand, reducing the training set to only documents close to the test ones in time will certainly reduce the training sample size. Therefore, there is a trade-off between these two effects.

Our strategy to optimize this trade-off and consequently the classifer performance is to use a time-sensitive selection of the documents for training. That is, we select as our training set, for each document being tested, only the documents that are closer in time to it. The proximity is defined by a window that can grow symmetrically in both directions, past and future, based on the year of the test document. To also consider the sampling effect, we varied the size of this window from 0 (that is, the year A_i of the tested document belongs to) to N, in which N represents the number of years before and after A_i. The value of N is defined as the value that maximizes the performance of the classifier to the documents that belong to the year A_i. Figure 11 shows the accuracy of the classifier as we varied the size of the window for test documents from years 1986 and 1990 (randomly chosen) in the ACM collection. And, Figure 12 shows the accuracy of the classifier as we varied the size of the window for test documents from years 1971 and 1974 (also randomly chosen) in the MedLine collection.



(a) Year 1990 (b) year 1986

Figure 11: Slide Window Classification - ACM Collection



(a) Year 1971 (b) Year 1974

Figure 12: Slide Window Classification - MedLine Collection

As we can see from the examples, there is not only one optimum size for the window for the entire database, in both collections, but an optimum size window may be found for each specific year. For instance in the ACM collection, for the documents from 1990, the

	Aids	Bio	Cancer	C M	Hist	Space L	Tox
Aids	0	0.19	0.16	0.18	0.19	0.18	0.19
Bio	-	0	0.04	0.20	0.17	0.19	0.12
Cancer	-	-	0	0.04	0.03	0.04	0.05
C M	-	-	-	0	0.21	0.08	0.05
Hist	-	-	-	-	0	0.20	0.11
SpaceL	-	-	-	-	-	0	0.05
Tox	-	-	-	-	-	-	0

Table 1: Similarity Std_Dev Matrix - MedLine

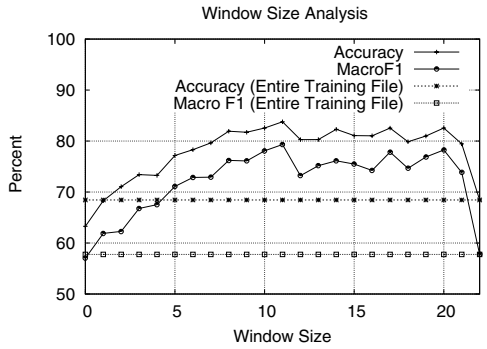
	Lit	HW	C S O	SW	Data	Theory	Math	Info S	C M	C App	Milieux
Lit	0	0.14	0.12	0.12	0.12	0.29	0.14	0.13	0.14	0.12	0.29
HW	-	0	0.08	0.13	0.11	0.12	0.11	0.12	0.10	0.10	0.13
C S O	-	-	0	0.10	0.09	0.10	0.08	0.07	0.08	0.10	0.13
SW	-	-	-	0	0.09	0.06	0.09	0.10	0.11	0.12	0.13
Data	-	-	-	-	0	0.05	0.08	0.09	0.10	0.13	0.13
Theory	-	-	-	-	-	0	0.14	0.13	0.07	0.06	0.29
Math	-	-	-	-	-	-	0	0.13	0.10	0.09	0.15
Info S	-	-	-	-	-	-	-	0	0.10	0.08	0.15
C M	-	-	-	-	-	-	-	-	0	0.11	0.13
C App	-	-	-	-	-	-	-	-	-	0	0.12
Milieux	-	-	-	-	-	-	-	-	-	-	0

Table 2: Similarity Std_Dev Matrix - ACM-DL

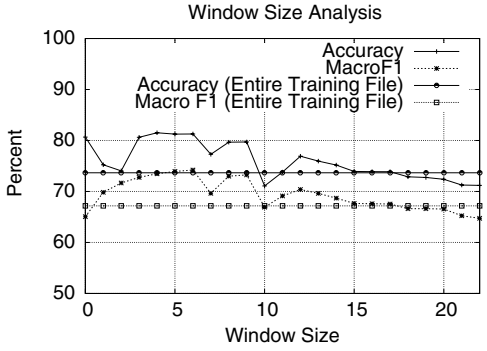
optimum size window is 10 years. For the documents of 1989, however, the optimum size is 20 years. Although the mechanism used to determine the temporal window is appropriate for a temporal evolution analysis, it is not feasible to real scenarios. However, as explained below, the accuracy of the classification process when we considered the temporal window was higher than using the entire training set, and this must be considered while building classifiers. A classifier that considers the temporal aspects, in addition to improving accuracy, could improve performance, since it could use a much smaller training set.

We first show that improvements can be obtained with the use of an average size window, i.e., a common window for all years. Figure 13 shows the optimum size of the window considering the entire training set using all years. Even when we do not consider an optimum window for each year, the use of an optimum global window is still better than the use of the whole training set. The graph in Figure 14 shows the size of the training set by the size of the window used. We can notice that using only 33% of the entire ACM training set (window size=5) we achieve a performance, in terms of accuracy, as good as using the whole training set. For the ACM collection, the optimum size of the global window is 11 years, which is 61% of the entire database. Analyzing the MedLine database, we can notice that the use of just 14% (window size=3) is enough to achieve a performance as good as using the whole collection. For this collection, the best result was found using the size of window as 4 years, which represents 22% of the entire training set.

A natural extension of the proposed procedure is to try to explore an optimum time window for each year. Figure 15 shows that the size of the optimum time window varies significantly for each year. Figure 16 shows the highest accuracy values found for each year in both collections. By analyzing this graph, we can easily see that the average of the classifier's performance is higher than the one we get when we use only one window size for all years, in which we have an average performance of 83.7% and 81.5% for ACM and MedLine, respectively, as shown in Figure 13. On the other hand, when we vary the size of the time window for each year, we get much higher values for both collections. By measuring all



(a) ACM Collection

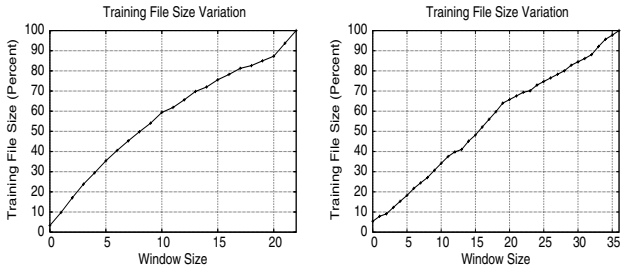


(b) MedLine Collection

Figure 13: Window Size Performance

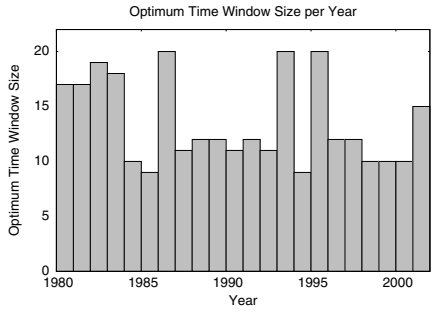
documents correctly classified in the whole test collection, we get an accuracy of 89.76% for ACM and 87.57% for MedLine.

These results are even better when compared to the results found when we use the whole training set for the classification proccess, in which we do not consider the temporal aspects. In the latter, we

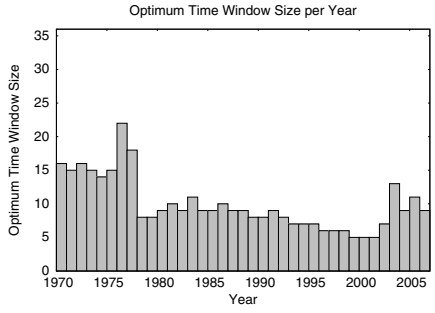


(a) ACM Collection (b) Medline Collection

Figure 14: Size of Collection

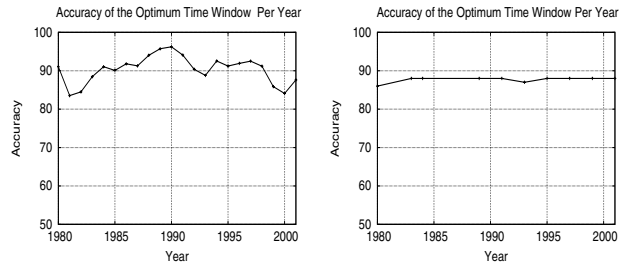


(a) ACM Collection



(b) MedLine Collection

Figure 15: Analisys of the Optimum Time Window Per Year



(a) ACM Collection (b) Medline Collection

Figure 16: Accuracy of the Optimum Time Window Per Year

had a performance of only 68.44% for ACM and 73.67% for MedLine, which corresponds to a gain of 31% in accuracy for ACM for the method that considers an optimum time window for each year, and a gain of approximattely 19% for MedLine. Gains in execution time are also obtained when we employ a different optimum time window size for each year. The optimum window produces a training set of approximately 69% of the ACM collection original training set and 25% for MedLine. Since the execution time of the SVM is directly related to the size of the training set, its reduction makes the execution time significantly shorter.

In summary, our empirical analysis shows that proper exploitation of the temporal aspects may lead to significant improvements in the classification process. Despite the exaustive technique being appropriate to analytical evaluation, as applied in the context of this work, certainly it is impracticable in actual scenarios. From this point of view, it emerges, like a major challenge and ongoing work, the need for a new approach to improve the quality of ADC using temporal aspects.

6. CONCLUSION AND FUTURE WORK

Using representative document collections (ACM and MedLine) and a state of the art classifier (SVM), we have shown evidence that time is indeed an important factor that should be taken into consideration in classification techniques and algorithms.

The temporal evolution may worsen the construction of text classification models, in the sense that models based on the whole collection do not work well, and accounting for its evolution is key for an effective classification. Therefore, there is a trade-off while handling the temporal evolution, becoming even more challenging the task of finding the proper granularity in terms of time interval for building the models. The first issue associated with this trade-off is the sampling effect in the context of the temporal evolution of the collection. The occurrence of sampling effect suggests that we should increase the size of the sample, while the temporal evolution, the second issue, induces the reverse, since using long time periods may generate conflicting evidence. In order to understand this trade-off, we introduced a methodology of analysis by factoring out each of the factors that are part of it. Moreover, we present some preliminary results that show that the appropriate analysis and understanding of the these factors may lead to improvements towards a more accurate model.

Our strategy for optimizing this trade-off and consequently the classifier performance was to use a time-sensitive selection of the documents for training. That is, for each document being tested, we select as our training set only the documents that are closer in time to it. The proximity is defined by a window that can grow symmetrically in both directions, past and future, based on the year of the test document. By using an optimum size window for each specific year that maximizes the accuracy of classifier, we have achieved an accuracy of 89.76% for ACM and 87.57% for MedLine, which means gains of up to 20% in accuracy with much less training data.

We intend to apply the methodology presented in this paper in diferent Web collections in order to show that the same temporal evolution effects occur. Despite the fact that an exhaustive technique is appropriate to analytical evaluation, as applied in the context of this work, it certainly is not feasible in actual scenarios. However, it shows that there is room for the development of such methods as a way to increase a classifier performance. Therefore, as future work, we intend to explore these time issues in the design of new lazy classification algorithms that use time information in the construction of classification models.

7. REFERENCES

[1] J. Allan. Incremental relevance feedback for information filtering. In *Proceedings of th 19th annual internacional ACM SIGIR conference on research and development in information retrieval*, Zurich, Switzerland, 1996. IEEE Computer Society.

[2] J. Allan, J. Carbonell, G. Doddington, J. Yamron, and Y. Yang. Topic detection and tracking pilot study final report. In *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop*, 1998.

[3] N. Belkin and W. Croft. Information filtering and information retrieval: two sides of the same coin? In *Communications of the ACM*. ACM, 1992.

[4] H. Borko and M. Bernick. Automatic document classification. *J. ACM*, 10(2):151–162, 1963.

[5] L. Brieman and P. Spector. Submodel selection and evaluation in regression: The x-random case. *International Statistical Review*, 60:291–319, 1992.

[6] N. H. M. Caldwell, P. J. Clarkson, P. A. Rodgers, and A. P. Huxor. Web-based knowledge management for distributed design. *IEEE Intelligent Systems*, 15(3):40–47, 2000.

[7] W. W. Cohen and Y. Singer. Context-sensitive learning methods for text categorization. *ACM Trans. Inf. Syst.*, 17(2):141–173, 1999.

[8] S. Dumain, J. Platt, D. Heckerman, and M. Sahami. Inductive learning algorithms and representation for text categorization. In *Proceedings of the Seventh Internacional Conference on Information and Knoledge Management*, 1998.

[9] G. Forman. An extensive empirical study of feature selection metrics for text classification. *Journal of Machine Learning Research*, 3:1289–1305, 2003.

[10] G. Forman. Tackling concept drift by temporal inductive transfer. In *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 252–259, New York, NY, USA, 2006. ACM Press.

[11] S. Haykin. Adaptive filters. In *Signal Processing Magazine*. IEEE Computer Society, 1999.

[12] T. Joachims. Making large-scale support vector machine learning practical. In A. S. B. SchÃűlkopf, C. Burges, editor, *Advances in Kernel Methods: Support Vector Machines*. MIT Press, Cambridge, MA, 1998.

[13] T. Joachims. Training linear svms in linear time. In *KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 217–226, New York, NY, USA, 2006. ACM Press.

[14] R. Jones and F. Diaz. Temporal profiles of queries. *ACM Trans. Inf. Syst.*, 25(3):14, 2007.

[15] R. Klinkenberg and I. Renz. Adaptive information filtering: Learning in the presence of concept drifts. In *Learning for Text Categorization*, pages 33–40, Menlo Park, California, USA, 1998. AAAI Press.

[16] C. Lanquillon and I. Renz. Adaptive information filtering: Detecting changes in text streams. In *Conference on Information and Knowledge Management (CIKM)*, Kansas City, USA, 1999. ACM.

[17] S. Lawrence and C. L. Giles. Context and page analysis for improved web search. *IEEE Internet Computing*, 2(4):38–46, 1998.

[18] D. D. Lewis. Evaluating and Optimizing Autonomous Text Classification Systems. In E. A. Fox, P. Ingwersen, and R. Fidel, editors, *Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 246–254, Seattle, Washington, 1995. ACM Press.

[19] R. Liu and Y. Lu. Incremental context mining for adaptive document classification, 2002.

[20] G. Salton and M. J. McGill. *Introduction to Modern Information Retrieval*. McGraw-Hill, 1983.

[21] M. Scholz and R. Klinkenberg. Boosting classifiers for drifting concepts.

[22] F. Sebastiani. Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1):1–47, 2002.

[23] Y. Yang and B. Kisiel. Margin-based local regression for adaptive filtering. In *Conference on Information and Knowledge Management (CIKM)*, New Orleans, USA, 2003. ACM.