# Relaxing Feature Selection in Spam Filtering by Using Case-Based Reasoning Systems

J.R. Méndez[1], F. Fdez-Riverola[1], D. Glez-Peña[1], F. Díaz[2], and J.M. Corchado[3]

[1] Dept. Informática, University of Vigo, Escuela Superior de Ingeniería Informática, Edificio Politécnico, Campus Universitario As Lagoas s/n, 32004, Ourense, Spain
{moncho.mendez,riverola,dgpena}@uvigo.es
[2] Dept. Informática, University of Valladolid, Escuela Universitaria de Informática, Plaza Santa Eulalia, 9-11, 40005, Segovia, Spain
fdiaz@infor.uva.es
[3] Dept. Informática y Automática, University of Salamanca, Plaza de la Merced s/n, 37008, Salamanca, Spain
corchado@usal.es

**Abstract.** This paper presents a comparison between two alternative strategies for addressing feature selection on a well known case-based reasoning spam filtering system called SPAMHUNTING. We present the usage of the *k* more predictive features and a percentage-based strategy for the exploitation of our amount of information measure. Finally, we confirm the idea that the percentage feature selection method is more adequate for spam filtering domain.

## 1 Introduction and Motivation

With the boom of Internet, some advanced communication utilities have been introduced in our society in order to improve our quality of life. Nowadays, there is no doubt of the large utility of some services like the World Wide Web (WWW), the Instant Messaging (IM) and Internet Relay Chat (IRC) tools as well as the e-mail possibilities. However, due to the great audience of these recent technologies and the lack of international legislation for their regulation, Internet has also been used as the basis for some illegal activities.

In this context, spam can be viewed as a generic concept referred to the usage of Internet technologies in order to promote illegal/fraudulent products and/or disturb Internet users. The low cost associated to these technologies is the main attraction for the spammers (the senders of spam messages). The most common forms of spam are (*i*) the insertion of advertisements in blog comments, (*ii*) the delivery of announcement mobile messages (SMS), (*iii*) the usage of advertising bots in IM or IRC channels, (*iv*) the delivery of advertisement messages on newsgroups and (*v*) the distribution of spam e-mail.

Since the most common and oldest form of spam is the usage of e-mail for disturbing Internet users, most of the efforts on spam filtering have been focused in this direction. In the same way, as the most extensive form of spam is the delivery of spam messages, this work has also been bounded to this area.

Previous research work on spam filtering have shown the advantages of disjoint feature selection in order to affectively capture the information carried by e-mails

J. Neves, M. Santos, and J. Machado (Eds.): EPIA 2007, LNAI 4874, pp. 53–62, 2007.

[1, 2, 3]. These methods are based on representing each message using only the words that better summarize its content. This kind of message representation has been successfully used in combination with some latest technologies in order to build a Case Based Reasoning (CBR) spam filtering system called SPAMHUNTING [3]. The usage of a poor feature selection method can significantly reduce the performance of any good classifier [1].

Traditionally, the feature selection stage has been carried out from the training dataset by using measures of global performance for each possible feature (term). The most well-known metrics used are Information Gain (IG), Mutual Information (MI), Document Frequency (DF) and Chi square test ($\chi^2$) [1]. Features can be selected either using a threshold over the metric or a best $k$ number of features with $k$ established before training the model [4]. The latest one is the most common in spam filtering domain [1, 4, 5, 6, 7].

As showed in [1, 2], when a new message is received SPAMHUNTING extracts its tokens and computes a measure of the Achieved Information for each term included into the message, $AI(t)$. Then, terms are sorted descending by its $AI(t)$ score. Finally, starting from the $n$ terms extracted from the message, only the first $k$ terms having an amount of information greater than a percentage $p$ of the total information achieved by using all terms. During our past experiments, we had found 60% as a good value for the percentage $p$.

Taking into consideration the SPAMHUNTING feature selection technique, a static number $k$ of features could be used instead of a percentage selection of attributes. In this paper we test the benefits of using a percentage feature selection approach instead of selecting a pre-established number of features for every message. As these alternatives have not been compared yet, we are interested in a deep analysis of performance in order to select the most appropriate way for guaranteeing the most accurate results.

The rest of the paper is organized as follows: Section 2 introduces previous work on feature selection used in conjunction with CBR systems. Section 3 shows the available corpus for empirical evaluation as well as some miscellaneous configuration details used during our experimental stage. Section 4 presents the experimental protocol and the results of the experiments carried out, discussing the major findings. Finally, Section 5 exposes the main conclusions reached by the analysis of the experiments carried out.

## 2 Feature Selection on Spam Filtering CBR Systems

Recent research works have shown that CBR systems are able to outperform some classical techniques in spam filtering domain [1, 2, 3, 6, 7]. Moreover, some previous works state that CBR systems work well for disjoint concepts as spam (spam about *porn* has little in common with spam offering *rolex*) whereas older techniques try to learn a unified concept description [6]. Another important advantage of this approach is the ease with which it can be updated to tackle the *concept drift* problem and the changing environment in the anti-spam domain [7]. Due to the relevance of these works, subsection 2.1 presents different feature selections strategies followed by some well-known CBR systems used for spam filtering. Subsection 2.2 explains the approaches for addressing feature selection that will be compared in this work.

### 2.1 Previous Work on Feature Selection

Motivated for the relevance of some previous work on spam filtering CBR systems [1, 2, 3, 6, 7], we had analyzed the weaknesses and strengths of several feature selection methods implemented by these successful classifiers.

In the work of [7] a new CBR spam filter was introduced. One of the most relevant features included in ECUE, (*E-mail Classification Using Examples*), was the ability for dynamically updating the knowledge. The feature selection approach used in this system has been designed as a classical spam filtering technique, by using the 700 word-attributes from training corpus having the highest IG score. The IG measure for a term, *t*, is defined by Expression (1).

$$IG(t) = \sum_{c \in \{spam, legitimate\}} P(t \wedge c) \cdot \log \frac{P(t \wedge c)}{P(t) \cdot P(c)} \tag{1}$$

where $P(t \wedge c)$ is the frequency of the documents belonging to the category *c* (legitimate or spam) that contains the term *t*, $P(t)$ represents the amount of documents having the term *t* and, finally, $P(c)$ is the frequency of documents belonging to the class *c*. Moreover, ECUE uses a similarity retrieval algorithm based on Case Retrieval Net*s* (CRN) [8], an efficient implementation of a *k*-nearest neighbourhood strategy. Finally, the system uses a unanimous voting strategy to determine whether a new e-mail is spam or not. It also defines a combination of two Case-Based Editing techniques known as Blame Based Noise Reduction (BBNR) and Conservative Redundancy Reduction (CRR) [9].

The ECUE system represents the evolution from a previous Odds Ratio (OR) based filter [6]. The main difference between both systems is the feature selection method used. This fact supports the idea of the relevance about feature selection methods in spam filtering domain. The feature selection method used in the oldest version of ECUE is based on selecting the 30 terms having the highest OR measure for each category. The OR measure for a term, *t*, in the class *c* (spam or legitimate) is computed as showed in Expression (2).

$$OR(t,c) = \frac{P(t \wedge c) \cdot \left[1 - P(t \wedge \overline{c})\right]}{P(t \wedge \overline{c}) \cdot \left[1 - P(t \wedge c)\right]} \tag{2}$$

where $P(t \wedge c)$ represents the frequency of the documents belonging to category *c* (legitimate or spam) that contains the term *t*, and $P(t \wedge \overline{c})$ stands for the frequency of documents containing the term *t* that are not included in category *c*.

Another relevant system can be found in [2], where the authors present a lazy learning hybrid system for accurately solving the problem of spam filtering. The model, known as SPAMHUNTING, follows an Instance-Based Reasoning (IBR) approach. The main difference between SPAMHUNTING and other approaches can be found during the feature selection stage. Instead of a global feature selection, SPAMHUNTING addresses this phase as an independent process for each email. Therefore, when a new message *e* arrives, SPAMHUNTING extracts its terms $\{t_i \in e\}$ and computes for each of them, the Amount of Information (AI) achieved when used for

the representation of the target e-mail $e$ [1]. This estimation is made by means of using Expression (3).

$$AI(t,e) = P(t \wedge e) \cdot \left[ \frac{\left| P(spam) \cdot P(t \wedge spam) - P(legitimate) \cdot P(t \wedge legitimate) \right|}{P(t)} \right] \quad (3)$$

where $P(t\ ,\ e)$ represents the frequency of the term $t$ in the message $e$, $P(spam)$ and $P(legitimate)$ are in that order, the frequency of spam and legitimate messages stored in the system memory, $P(t\ ,\ spam)$ and $P(t\ ,\ legitimate)$ stand for the frequency of spam and legitimate messages stored the system memory having the term $t$ and, finally, $P(t)$ represents the frequency of instances from CBR knowledge containing the term $t$. The $AI$ estimation summarizes the information of the relevance of the term $t$ in the message $e$ (by using $P(t\ ,\ e)$) as well as the global relevance of the term in the whole corpus (the remaining expression).

Each message is represented using a set of attributes that better describe its content as a set of keywords describe a research paper. For this purpose we had defined a measure for the total amount of information of a message, $AI(e)$, computed as the sum of the $AI$ measure for each of the terms extracted from $e$. This measure can be computed as Expression (4) shows.

$$AI(e) = \sum_{t_i \in e} AI(t_i, e) \quad (4)$$

where $AI(t_i,\ e)$ stands for the amount of information of the term $t_i$ in the message $e$ defined in Expression (3). The selection of features for the message $e$ has been defined in [2] as the list of terms $t_i$ with the highest $AI(t_i\ ,\ e)$ rate having an amount of achieved information greater than a certain percentage $p$ of the total amount of information $AI(e)$. Expression (5) demonstrates how to carry out the feature selection process for a given message $e$, $FS(e)$.

$$FS(e) = \left\{ t_i \in e \left| \left\{ \begin{array}{l} \left\{ AI\left(t_j, e\right) \geq AI\left(t_k, e\right) \wedge t_k \in FS(e) \right\} \rightarrow t_j \in FS(e) \\ \\ \nexists FS' \subset FS\left(e\right) \left| \sum_{t_j \in FS'} AI\left(t_j, e\right) > \frac{p}{100} \cdot \sum_{t_k \in e} AI\left(t_k, e\right) \right. \end{array} \right. \right\} \right\} \quad (5)$$

where $AI(t_k,\ e)$ stands for the achieved information score of the term $t_k$ in the message $e$ (showed in Expression (3)). Previous work using this approach has shown good performance results when $p$=60 [1, 2].

According to the disjoint feature selection method used by SPAMHUNTING, it comprises an Enhanced Instance Retrieval Network known as EIRN as a primary way of managing knowledge [10]. This indexing structure guarantees the representation of disjoint information and implements an efficient similarity metric defined as the number of relevant features found in a set of given messages. Similarly with ECUE, the reuse of retrieved messages has been defined as a simple unanimous voting strategy

[7]. Finally, the revision stage comprises the usage of a measure of the quality of the available information for classifying the target message [11].

Once the ground works has been exposed, next subsection presents a different approach for carrying out the feature selection using the background ideas of our successful SPAMHUNTING system process.

### 2.2 Best *k* Feature Selection on SPAMHUNTING

This subsection presents a new approach for feature selection in our previous successful SPAMHUNTING system. This proposal for feature selection will be compared with the original (showed in Expression (5)) during the experimentation stage.

The SPAMHUNTING feature selection stage has been addressed as an independent process executed every time a message is received. The main instrument for this computation is a percentage $p$ of the global $AI$ from the target message $e$, $AI(e)$. Despite this approach is reasonable and has led to obtain good results, authors do not justify the requirement of the percentages.

In this work we plain to use a fixed number of attributes ($k$) for representing each email. In order to do this, we will select these attributes holding the greatest $AI$ rate, $AI(t_i, e)$, for each given message $e$. In our new proposal, the selected terms for a message $e$ will be computed as showed in Expression (6)

$$FS(e) = \left\{ t_i \in e \left| \begin{Bmatrix} numTerms(e) \geq k \rightarrow size(FS(e)) = k \\ numTerms(e) \geq k \rightarrow size(FS(e)) = numTerms(e) \\ AI(t_i, e) \geq AI(t_j, e) \wedge t_j \in FS(e) \rightarrow t_i \in FS(e),\ \forall t_j, t_i \in e \end{Bmatrix} \right. \right\} \quad (6)$$

where $FS(e)$ represents the list of selected features for the message $e$, $size(x)$ is the length of $x$ and $numTerms(e)$ stands for the amount of terms extracted from the target message $e$.

This paper contains a brief empirical analysis of the analyzed approaches shown in Expressions (5) and (6). The final goal is to determine the best approach for spam filtering CBR systems or identify the circumstances that improve their performance. Next section presents the available corpora for spam filtering, the main criterions for the dataset selection and some miscellaneous configuration issues used during the experiments.

## 3 Available Corpus and Experimental Setup

Due to privacy issues (mainly from the usage of legitimate messages), the spam filtering domain presents some difficulties in order to guarantee the relevance of the results achieved. For addressing this difficulty, only publicly available datasets should be used. This section contains a summary of the available dataset repositories for spam filtering. Moreover, it includes a discussion about the decisions adopted for the experimental datasets and several SPAMHUNTING configuration details used during our experimentation.

There are several publicly available corpora of e-mails including LingSpam, Junk E-mail, PU corpuses, DivMod or SpamAssassin[1]. Most of them, according with the RFC (*Request for Comments*) 822, are distributed as they were sent through Internet. Although some researchers and students from our University had built a corpus called SING, most of the messages have been qualified as private by its owners and we are not authorized to publish it. This corpus can only be used for parameter optimization purposes while the results should be generated using publicly available datasets. Moreover, there are some corpuses shared after some preprocessing steps that can cause the loosing of some information. Table 1 shows a brief description of all of them focusing on preprocessing and distribution issues.

**Table 1.** Comparison of some publicly available datasets

| Name | Public | Message amount | Contains dates | Legitimate percentage | Spam percentage | Distribution - preprocessing |
|---|---|---|---|---|---|---|
| Ling-Spam | YES | 2893 | NO | 83.3 | 16.6 | tokens |
| PU1 | YES | 1099 | NO | 56.2 | 43.8 | token ids |
| PU2 | YES | 721 | NO | 80 | 20 | token ids |
| PU3 | YES | 4139 | NO | 51 | 49 | token ids |
| PUA | YES | 1142 | NO | 50 | 50 | token ids |
| SpamAssassin | YES | 9332 | YES | 84.9 | 15.1 | RFC 822 |
| Spambase | YES | 4601 | NO | 39.4 | 60.6 | feature vectors |
| Junk-Email | YES | 673 | YES | 0 | 100 | XML |
| Bruce Guenter | YES | 171000 | YES | 0 | 100 | RFC 822 |
| Judge | YES | 782 | YES | 0 | 100 | RFC 822 |
| Divmod | YES | 1247 | YES | 0 | 100 | RFC 822 |
| Grant Taylor | YES | 2400 | YES | 0 | 100 | RFC 822 |
| SING | NO | 20130 | YES | 69.7 | 39.3 | XML y RFC 822 |

As we can see from Table 1, most of the available corpus do not contain legitimate messages and should be discarded. Moreover, in order to use some preprocessing issues introduced in [12], an RFC 822 distribution of the corpus need to be selected. Due to the above mentioned reasons, we have selected the SpamAssassin corpus containing 9332 different messages from January 2002 up to and including December 2003.

Following the findings of [12], hyphens and punctuation marks are not removed during the tokenizing stage. We have used "\\S+" as the regular expression for determining a token on the message text. This expression means that we define tokens as character lists containing the greatest amount of non white space consecutive characters. Moreover, attending to [12], we have executed a stopword removal process over the identified tokens [13].

Finally, continuous updating strategies have been used during the experiments. Every time a message is classified, a new instance is generated using the e-mail content and the generated solution. Next section presents the experimental design and the benchmark results.

[1] Available at http://www.spamassassin.org/publiccorpus/

## 4 Experimental Protocol and Results

The final goal of our experiments is to measure the differences between the feature selection strategies showed in Expressions (5) and (6). This section presents the experimental methodology and the results achieved from the tests carried out. All the experiments have been carried out using a 10-fold stratified cross-validation [14] in order to increase the confidence level of the results obtained.

We have executed our SPAMHUNTING system using Expression (5) applying the values 30, 40, 50, 60 and 70 for $p$ and Expression (6) using values from 4 to 14 stepping 2 for $k$. Then, we have selected the best of them using the area under the Receiver Operating Characteristic (ROC) curve [15]. ROC curves have been computed using the amount of spam votes as target measure for the discrimination between spam and legitimate messages. In order to provide information about the theoretical maximum performance reachable by using each analyzed configuration, we have also computed the sensitivity and specificity for the best classification criterion. Table 2 summarizes the area under ROC curves for the different model configurations and the theoretical maximum performances achieved.

As we can see from Table 2, the difference between the analyzed configurations of each feature selection strategy is very small. Although there are some configurations able to achieve some important values on singular measures, we have selected those achieving the highest global performance level (evaluated through the area under the ROC curve). In this sense, $p$=40 and $k$=12 have shown to be the best configurations for the approach.

After the preliminary analysis, we have compared the results achieved by using the best configuration for the two selected proposals. This task has been carried out using a complete ROC curve analysis [11, 15]. Then, we have carried out a comparison of the areas under ROC curves and executed a statistic test in order to determine the significance of the differences found. Figure 1 shows the ROC curves plot for the best configuration of the analyzed approaches.

As we can see from Figure 1, there is a difference between the analyzed strategies. As we have previously mentioned, we have executed a statistic test considering the

**Table 2.** Preliminary comparison of ROC curves achieved by the SPAMHUNTING system

| | Configuration | Area under ROC | Sensitivity | Specificity | Best criterion |
|---|---|---|---|---|---|
| PERCENTAGE SELECTION | 30% | 0.985 | 96.9 | 98.5 | >0,0714 |
| | 40% | 0.987 | 97.4 | 98.9 | >0.0417 |
| | 50% | 0.984 | 96.9 | 99.3 | >0.0426 |
| | 60% | 0.983 | 96.8 | 99.4 | >0.0417 |
| | 70% | 0.980 | 96.1 | 99.7 | >0.0741 |
| BEST $k$ SELECTION | 4 | 0.990 | 97.4 | 97.6 | >0.2 |
| | 6 | 0.989 | 97.7 | 98.3 | >0.1644 |
| | 8 | 0.992 | 97.8 | 98.9 | >0.1648 |
| | 10 | 0.992 | 97.9 | 99.2 | >0.1603 |
| | 12 | 0.994 | 98.7 | 98.7 | >0.0545 |
| | 14 | 0.993 | 98.4 | 99.1 | >0.0972 |

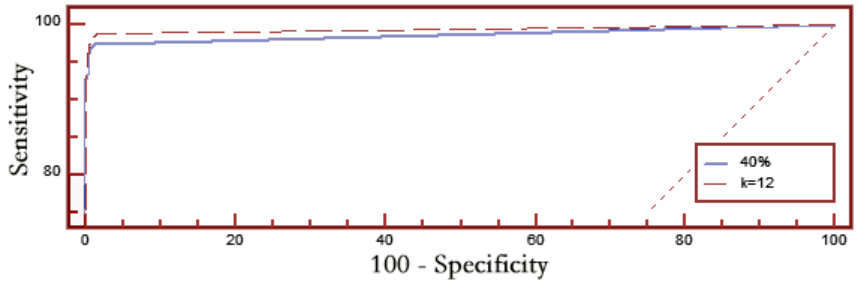


**Fig. 1.** ROC curves plot for the best configuration of the two feature selection approaches

equality of the areas under ROC curves showed as the null hypothesis. As the computed *p*-value is lower than 0.01 ($p$=0.001), the null hypothesis must be discarded and we can state that, from a statistically point of view, there is a very significant difference between both methods for feature selection.

In order to provide a more detailed analysis, we have computed positive Likelihood Ratio (+LR), Negative Likelihood ratio (-LR) and $\delta$ index [16] for the best cut values of the analyzed proposals. These measures have been included in Table 3.

**Table 3.** ROC analysis for the selected feature selection strategies

| | *Percentage 40%* | *Best k=12* |
|---|---|---|
| **+LR** | 86.76 | 73.07 |
| **-LR** | 0.03 | 0.01 |
| **$\delta$ index** | 2.76 | 2.91 |

As we can realize from Table 3, the detection of a legitimate message has a great confidence level when best *k* feature selection is used, whereas the percentage method presents a great confidence level when detecting spam messages (reducing the false positive error rate). Moreover, analyzing the $\delta$ index we can see that the best *k* feature selection alternative can theoretically achieve a better amount of correct classifications.

Finally, six well-known metrics [17] have been used in order to evaluate the performance (efficacy) of both models: percentage of correct classifications (%OK), percentage of False Positives (%FP), percentage of False Negatives (%FN), spam *recall*, spam *precision* and *total cost ratio* (TCR) with three different cost values. In order to achieve a down-to-earth approximation of the performance level, these measures has been computed considering the results of the unanimous voting strategy instead of the best criterion identified by using the ROC curve analysis. Table 4 shows a comparison of both feature selection techniques using percentages of correct/fail classifications, precision, recall and TCR measures.

As we can see from Table 4, despite the differences between best *k* and percentage feature selection are small, they support the conclusions achieved during the ROC analysis. Percentage feature selection is suitable to address FP error reduction while best *k* feature selection can increase the percentage of correct classifications.

Finally, keeping in mind the relevance of false positive errors on the target domain, TCR measure achieved for the highest cost factor ($\lambda$=999) supports the convenience of the percentage feature selection.

**Table 4.** Performance measures for the analyzed feature selection strategies in SPAMHUNTING

| | *Classification percentages* | | | | | *TCR* | | |
|---|---|---|---|---|---|---|---|---|
| | *%OK* | *%FP* | *%FN* | *Recall* | *Precision* | *λ=1* | *λ=9* | *λ=999* |
| *Percentage 40%* | 97.51 | 0.09 | 2.39 | 0.90 | 0.99 | 10.54 | 8.70 | 4.76 |
| *Best k=12* | 97.82 | 0.10 | 2.09 | 0.92 | 0.99 | 12.22 | 8.84 | 2.25 |

## 5 Conclusions and Further Work

This work has presented, discussed and analyzed two different proposals for addressing the feature selection stage on a previous successful spam filtering CBR system. Moreover, we have discussed the main differences between several feature selection schemes applied in the context of the target domain. We had also analyzed the current SPAMHUNTING feature selection heuristic in order to find new improvements. Finally, we have executed a benchmark finding the percentage approach as the most suitable alternative for spam filtering domain.

The performance of compared feature selection strategies are based on an analysis of the discrimination capabilities of each term. When the most relevant terms have lower discerning capabilities, the percentage selection strategy is able to select a greater amount of terms in order to guarantee the existence of a minimum amount of information for accomplishing the classification. Moreover, the best *k* feature selection approach can not guarantee a minimum of information.

Spammers often include terms that can be understood as reliable signs for classifying their messages as legitimate. These messages present a large amount of information because they combine a greater number of legitimate and spam signs. Percentage selection is able to include both spam and legitimate signs for an adequately classification avoiding the possibilities of finding disagreement between the terms. Moreover, best *k* feature selection will include only the legitimate evidences storing these inconsistencies in the system memory. Although the global performance can increase, those inconsistencies could generate a greater amount of FP errors. These findings are supported by the +LR and –LR tests carried out and the TCR comparisons for $\lambda$=9.

We believe in the relevance of a disjoint and percentage feature selection strategy for the spam filtering domain. In this sense future works should be focused in improving our feature selection methods [1, 2], using semantic information and addressing the noise reduction during the feature selection stage.

Acknowledgments. This work has been supported by TIN-2006-14630-C03-03.

## References

1. Méndez, J.R., Fdez-Riverola, F., Iglesias, E.L., Díaz, F., Corchado, J.M.: A Comparative Performance Study of Feature Selection Methods for the Anti-Spam Filtering Domain. In: Proc. of the 6th Industrial Conference on Data Mining, pp. 106–120 (2006)
2. Méndez, J.R., Fdez-Riverola, F., Díaz, F., Iglesias, E.L., Corchado, J.M.: Tracking Concept Drift at Feature Selection Stage in SpamHunting: an Anti-Spam Instance-Based Reasoning System. In: Proc. of the 8th European Conference on Case-Based Reasoning, pp. 504–518 (2006)
3. Fdez-Riverola, F., Iglesias, E.L., Díaz, F., Méndez, J.R., Corchado, J.M.: SpamHunting: An Instance-Based Reasoning System for Spam Labeling and Filtering. Decision Support Systems (in press, 2007) http://dx.doi.org/10.1016/j.dss.2006.11.012

4. Méndez, J.R., Corzo, B., Glez-Peña, D., Fdez-Riverola, F., Díaz, F.: Analyzing the Performance of Spam Filtering Methods when Dimensionality of Input Vector Changes. In: Proc. of the 5th International Conference on Data Mining and Machine Learning (to appear, 2007)
5. Metsis, V., Androutsopoulos, I., Paliouras, G.: Spam Filtering with Naive Bayes – Which Naive Bayes? In: Proc. of the 3rd Conference on Email and Anti-Spam, pp. 125–134 (2006), http://www.ceas.cc
6. Cunningham, P., Nowlan, N., Delany, S.J., Haahr, M.: A Case-Based Approach to Spam Filtering that Can Track Concept Drift. In: Ashley, K.D., Bridge, D.G. (eds.) ICCBR 2003. LNCS, vol. 2689, Springer, Heidelberg (2003)
7. Delany, S.J., Cunningham, P., Coyle, L.: An Assessment of Case-base Reasoning for Spam Filtering. In: AICS 2004. Proc. of Fifteenth Irish Conference on Artificial Intelligence and Cognitive Science, pp. 9–18 (2004)
8. Lenz, M., Burkhard, H.D.: Case Retrieval Nets: Foundations, properties, implementation and results. Technical Report: Humboldt University, Berlin (1996)
9. Delany, S.J., Cunningham, P.: An Analysis of Case-Based Editing in a Spam Filtering System. In: Proceedings of the 7th European Conference on Case-Based Reasoning, pp. 128–141 (2004)
10. Fdez-Riverola, F., Iglesias, E.L., Díaz, F., Méndez, J.R., Corchado, J.M.: Applying Lazy Learning Algorithms to Tackle Concept Drift in Spam Filtering. ESWA: Expert Systems With Applications 33(1), 36–48 (2007)
11. Méndez, J.R., González, C., Glez-Peña, G., Fdez-Riverola, F., Díaz, F., Corchado, J.M.: Assessing Classification Accuracy in the Revision Stage of a CBR Spam Filtering System. Lecture Notes on Artificial Intelligence (to appear, 2007)
12. Méndez, J.R., Iglesias, E.L., Fdez-Riverola, F., Díaz, F., Corchado, J.M.: Tokenising, Stemming and Stopword Removal on the Spam Filtering Domain. In: Proc. of the 11th Conference of the Spanish Association for Artificial Intelligence, pp. 449–458 (2005)
13. Baeza-Yates, R., Ribeiro-Neto, B.: Modern Information Retrieval. Addison-Wesley, Reading (1999)
14. Kohavi, R.: A study of cross-validation and bootstrap for accuracy estimation and model selection. In: IJCAI 1995. Proceedings of the 14th International Joint Conference on Artificial Intelligence, pp. 1137–1143 (1995)
15. Egan, J.P.: Signal Detection Theory and ROC Analysis. Academic Press, New York (1975)
16. Hasselband, V., Hedges, L.: Meta-analysis of diagnostics test. Psychological Bulletin 117, 167–178 (1995)
17. Androutsopoulos, I., Paliouras, G., Michelakis, E.: Learning to Filter Unsolicited Commercial E-Mail. Technical Report 2004/2, NCSR "Demokritos" (2004)