Open-Set Classification for Automated Genre Identification

Dimitrios A. Pritsos and Efstathios Stamatatos

University of the Aegean
Karlovassi, Samos – 83200, Greece
{dpritsos,stamatatos}@aegean.gr

Abstract. *Automated Genre Identification* (AGI) of web pages is a problem of increasing importance since web genre (e.g. blog, news, e-shops, etc.) information can enhance modern Information Retrieval (IR) systems. The state-of-the-art in this field considers AGI as a closed-set classification problem where a variety of web page representation and machine learning models have intensively studied. In this paper, we study AGI as an open-set classification problem which better formulates the real world conditions of exploiting AGI in practice. Focusing on the use of content information, different text representation methods (words and character n-grams) are tested. Moreover, two classification methods are examined, one-class SVM learners, used as a baseline, and an ensemble of classifiers based on random feature subspacing, originally proposed for *author identification*. It is demonstrated that very high *precision* can be achieved in open-set AGI while *recall* remains relatively high.

Keywords: Automated Genre Identification, Classifier Ensembles, One Class SVM.

1 Introduction

Genre is widely acknowledged as a significant factor for characterizing a document. Information about genre of web-pages (e.g. blogs, news, e-shops, etc.) could significantly enhance information retrieval systems by suggesting queries that better describe the user's information need or by facilitating intuitive navigation through search results [1,2]. During the last decade, automated genre identification (AGI) of web-pages has been thoroughly studied. The state-of-the-art has been focused on appropriate *web-page representation* techniques (textual content, HTML tags, etc.), *text representation* approaches (Character n-grams, Words, Part-of-speeches, etc.), *feature selection* methods (Chi-square, mutual information, etc.), term weighting schemes (Terms Frequency, Binary etc.), and classification methodologies (SVM, neural networks, etc.) [3,4,5,6,7].

So far, most published studies in this field consider AGI as a closed-set classification task (that is, each document should be assigned to at least one predefined genre label). However, it is clear that in large scale information retrieval systems, AGI can only be defined as an open-set task (a document may not be assigned to

P. Serdyukov et al. (Eds.): ECIR 2013, LNCS 7814, pp. 207–217, 2013.

any genre label) since it is quite likely that the predefined genre palette could not cover all the genres existing in a very large corpus. Moreover, web page genres are still evolving. So, it is not possible to define a complete set of genres and use it for a long period. On the other hand, while we potentially can have a great amount of *positive examples* for a given genre, it is difficult or impossible to compose a set of negative samples that provides a comprehensive characterization of everything that does not belong to the *target concept.*

In this study we are approaching AGI of web-pages as an *open-set classification* task, which better formulates the real world conditions of a constantly increasing *web-graph* and emerging *web-genres (or cybergenres).* We compare two different open-set classification methods; (i) *One-class SVM (OC-SVM)*, which builds one model per genre using only the positive examples and (ii) a *Random Feature Subspacing Ensemble (RFSE)* method, originally proposed for author identification [8], a task with many similarities with AGI. Two of the most popular web-genre collections used in previous studies are used to evaluate these methods: the *7-genre* and *KI-04* corpora [9]. Results suggest that RFSE performs significantly better than OC-SVM.

The rest of this paper is organized as follows. The next section comprises related work. Sections 3 and 4 describe in detail the open-set classification methods used in this study. Section 5 comprises the experimenal set-up and the evaluation results. Finally, Section 6 summarizes the conclusions drawn and discusses future work directions.

2 Related Work

Overcoming the lack of consensus about the definition of the genre itself or the genre palette, at least in the context of web-pages, a significant amount of work has been done on AGI during the last decade. Several aspects of this task have been studied thoroughly, including *document representation* (e.g. character n-grams, words, part-of-speeches etc.), *term weighting* (e.g. TF, TF-IDF, Binary, etc.) *feature selection* (e.g. frequency-based, chi-square, information gain, mutual information) and the *classification model* (e.g., SVM, decision trees, neural networks, etc.) [3,5,6,10,11,9,12,7]. To the best of our knowledge, all published studies consider AGI as a closed-set classification approach.

Many studies underline the effectiveness of the character n-grams for this task [5,12,3]. This type of feature has been used in combination with classification methods able to handle very high number of features such as SVM as well as similarity-based methods that construct one representation vector per genre [12]. The combination of variable length n-grams, where a *LocalMaxs* algorithm was selecting the proper mixture of n-gram lengths, has also been proposed [5]. Character n-gram features seem to be very effective in combination with the binary term weighting scheme [5,3].

In addition to the textual content, structural information (e.g. HTML tags) of web-pages is usually exploited in AGI. It seems that structural information is useful as a complement to textual features. Combining structural with textual information usually improves the classifiers performance [5].

One-class classification or novelty-detection handles data where only positive examples are available and has been applied to several domains[13]. One-class SVM (OC-SVM) is perhaps the most popular method. The key concept of OC-SVM is based on the ν-SVM model proposed by Scholkopf et al.[14] cosniders the origin as the only negative example. OC-SVM is discussed in section 3 in more detail. A variation of this method, called *Outliers-SVM,* considers as outliers a few examples from the original positive sample space and use them as negative examples additionally to the origin *[15]*. *Outliers-SVM* together with several other one-class classification methods such as *One Class Neural Networks, One Class Naive Bayes Classifier, One Class Nearest Neighbor etc.,* have been tested in the text categorization domain where they achieve relatively low performance in comparison to *closed-set classification* methods *[15].*

In a recent paper, Anderka et al. [16] present a method to build an artificial negative class and then use a random forest classifier ensemble to distinguish it from the positive class in applying one class learning to the problem of detecting text quality. Another interesting idea is to use PU learning to form samples of the negative class [17].

3 One-Class SVM

One-class SVM is actually an ν-SVM for the case we want to find the contour which is prescribing the positive samples of the training set given for a single class, while there are *no negative samples.* nu-SVM (ν-SVM) is providing an alternative *trade-off control method of misclassification*, proposed from Scholkopf et al. [14].

In ν-SVM we are minimizing eq.1 with the constraints of eq.2, eq.3, and eq.4.

$$arg \min_{w,b} \left\{ \frac{1}{\nu\lambda} \sum_{n=1}^{N} (\xi_n - \rho) + \frac{1}{2}\|w\|^2 \right\} \tag{1}$$

$$0 \leqslant a_n \leqslant 1/N, \qquad n = 1, ..., N \tag{2}$$

$$\nu \leqslant \sum_{n=1}^{N} a_n \tag{3}$$

$$\sum_{n=1}^{N} a_n t_n = 0 \tag{4}$$

Following the logic from the conventional SVM, thoroughly analyzed in [18], the Lagrange multipliers for solving the optimization problem of eq.1 under eq.2, eq.3 and eq.4 constraints are used. Equation 5 is then derived, i.e. a Lagrangian function to be maximized as subject to the constraints eq.2, eq.3 and eq.4:

$$\widetilde{L}(a) = -\frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{M} a_n a_m t_n t_m k(x_{n,} x_m) \tag{5}$$

It should be noted that ν in ν-SVM has the flowing properties:

- ν is an upper bound on the fraction of *Outliers.*
- ν is a lower bound on the fraction of *Support Vectors.*
- ν values cannot exceed 1 (see eq.2).

In practice different values of v defines different proportion of the training sample as outliers. For example in Scholkopf et al. [14] is showed that in their experiments when using $\nu = 0.05$, 1.4% of the training set has been classified as outliers while using $\nu = 0.5$, 47.4% is classified as outliers and 51.2% is kept as SVs.

In the prediction phase in order for an SVM model to decide wether a document is belonging to the target genre-class or not a *decision function* is returned. The decision function indicates the distance of the document, positive or negative, to the hyperplane separating the classes. In the case of OC-SVM we usually only interested whether the decision function is positive or negative for deciding if an arbitrary document belonging or not to the target class.

In our case, where multiple genres are given, a number of *one-class learners* is build, one for each genre available in the training corpus. In the prediction phase, the predicted genre/class is the one for which its learner has the highest positive distance from the hyperplane (or the contour for OC-SVM). If all the classifiers return a negative distance (i.e. the web-page does not belong to this genre) the final answer is "Don't Know". We used the scikit-learn python package to implement this method [1].

4 Ensemble-Based Algorithm

Our ensemble-based algorithm is a variation of the method presented by Koppel et al. [8] for the task of *author identification.* In the original approach, there is only one training example for each author and a number of simple classifiers is learned based on random feature subspacing. Each classifier uses the cosine distance to estimate the most likely author. The key idea is that it is more likely for the true author to be selected by the majority of the classifiers since the used features will still be able to reveal that high similarity. That is the style of the author is captured by many different features so a subset of them will also contain enough stylistic information. Since AGI is also a style-based text categorization task, this idea should also work for it.

In our study, there are multiple training examples for each available genre. To maintain simplicity of classifiers, we have used a *centroid vector* for each genre. Each centroid vector is formed by averaging all the TF vectors of the training examples of web pages for each genre. Our ensemble-based algorithm is described in *Algorithm* 1.

This algorithm is based on three important parameters: the number of iterations (*k1*), the number of features used in each iteration (*k2*) and the proportion of times a genre has to win to be given as the final answer (*sigma*). The latter

[1] `http://scikit-learn.org/stable/`

Algorithm 1. The *Random Feature Subpacing Ensemble* algorithm.

```
Given: a set of known web-pages for each of G genres,
     an unknown web-page, k1, k2, sigma

For each genre in G
  a.Average known web-pages of genre G
  to build one centroid vector.
Repeat k1 times
  a.Randomly choose some fraction k2 of the full feature set.
  b.Find top match of unknown page in centroid vectors
  using cosine similarity.
For each genre g in G
  a.Score(g) = proportion of times g is the top match.

Output: arg max g Score(g) if max Score(g) > sigma
    ; else "Don't Know"
```

is crucial for the performance of the algorithm. The larger *sigma*, the larger the precision and the lower the recall. It is possible to use this algorithm in combination with any text similarity measure. The cosine distance has provided good results in the experiments of [8] and is also used in this study.

5 Experiments

Our main objective is to compare the two open-set classification methods presented in the previous sections using various evaluation corpora and text representation schemes. In the following, we first describe the experimental set-up and then present the evaluation results.

5.1 Experimental Set-Up

We use two of the most popular corpora in AGI:

- *7-genre* [10]: This is a collection of 1400 English web-pages evenly distributed into 7 genres.
- *KI-04* [9]: This is a collection of 1205 English web-pages categorized into 8 genres. It is unbalanced.

The genre palettes of these corpora have some similarities (e.g. e-shops and personal home pages are included in both) and differences (e.g. 7-genre comprises blogs while KI-04 doesn't). They have been extensively used in many AGI studies but following the closed-set classification scenario *[4,3]*. Hence, the results if these studies are not directly comparable to the our results. To obtain more reliable results, we followed the practice or previous studies and performed 10-fold cross-validation with these corpora.

We are using only textual information from the web pages. All HTML tags and other non-textual information is removed. Two well-known text representation methods are then used:

- *character n-grams.* Based on the results reported by Sharoff et al. [3]*character 4-grams* were tested.
- *words*

In both cases, we use the TF weighting scheme and the vocabulary comprises all the terms in the training corpus.

As concerns OC-SVM, two feature set sizes were examined, one based on the 1,000 most frequent terms of the vocabulary and one based on the 5,000 most frequent terms of the vocabulary. Following the reports of previous studies [14,18] and some preliminary experiments, we examined the parameter values ν=*{0.05, 0.1, 0.5, 0.8}.*

With respect to RFSE, following the suggestion of Koppel et al. [8] we used *k1*=100 in each experiment. Using more than 100 iterations to build the ensemble does not improve significantly the results. We examine several values of *k2* (i.e. 1000, 5000, 10000, 70000) to estimate how this affects the performance of the tested methods. In each case, the frequency of features is not used to select the subset of features from the vocabulary (i.e. random selection of features).

5.2 Performance of OC-SVM

In figure 1 the performance (precision values in 11 standard recall levels) of OC-SVM is depicted with respect to different values of the ν parameter. In more detail, we show the best performance we achieved with OC-SVM models for 7-genre and KI-04 in figures 1(a) and 1(b), respectively. In both cases, character 4-grams were the most effective features.

As it has been observed in previous AGI studies on this corpora, KI-04 is harder than 7-genre [5,3]. It should be noted that in most of the examined cases ν=0.1 provided the best results for the *7-genre* corpus. On the other hand, for the *KI-04* corpus ν=0.8 was the most appropriate value. Note that the higher the ν, the more strict the boundary of the area that includes the positive class. This probably means that *KI-04* classes are more vague in comparison to *7-genre.*

5.3 Performance of RFSE

Figures 2, 3,4 and 5 show the performance of RFSE models on *7-genre* and *KI-04*, respectively. Character 4-gram and word features are examined for several values of *k2*. In addition, the performance of the best OC-SVM model on each corpus based on the corresponding feature type is used as a baseline. It is clear that RFSE significantly outperforms the baseline for all *k2* values. Moreover, large *k2* values (70,000) seem to be better choices for maintaining precision on top level for low and middle recall values. On the other hand, for high recall values, lower *k2* values (5,000 or 10,000) provide more robust solutions.

The larger marks on each curve of figures 2, 3,4 and 5 correspond to *sigma*=0.9. As can be seen, the larger the *k2* value, the higher the corresponding recall measure. For very high values of *k2* most of the web-pages are covered by this condition. This means we have a clear decision about the genre of the majority of the web-pages.

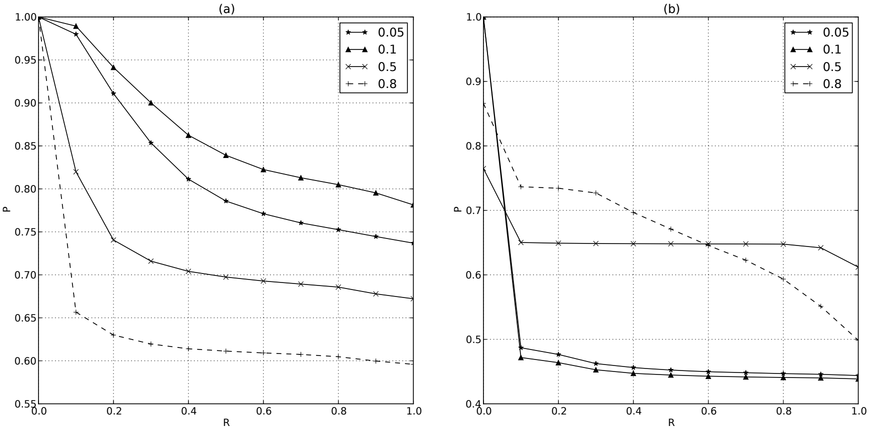

Fig. 1. Performance of the best OC-SVM models on (a) *7-genre* corpus and (b) *KI-04* corpus

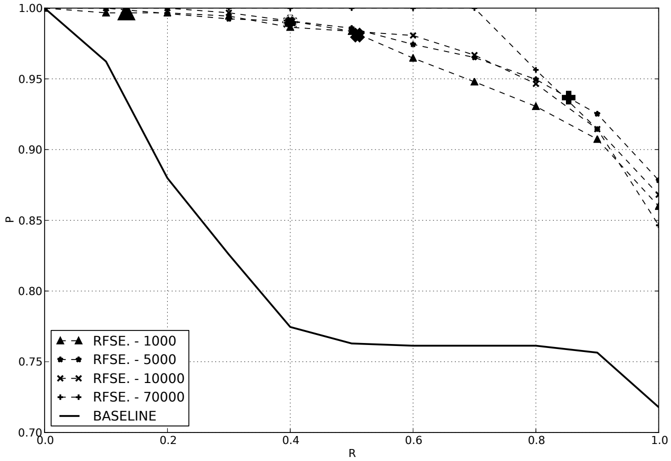

Fig. 2. Performance of RFSE models on *7-genre* corpus, for character 4-grams. The baseline refers to the best OC-SVM model for the same corpus.

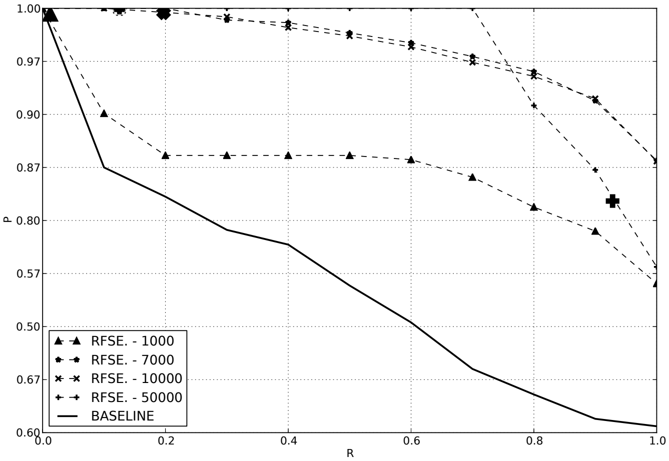

Fig. 3. Performance of RFSE models on *7-genre* corpus, for words. The baseline refers to the best OC-SVM model for the same corpus.

As concerns the feature types, character 4-grams and words provide competitive results in both corpora. In average, character 4-grams are slightly more effective. However, low values of *k2* (1,000) seem to be particularly harmful for representation models based on words. Character n-grams are not considerably affected by decreasing the dimensionality of the base classifiers.

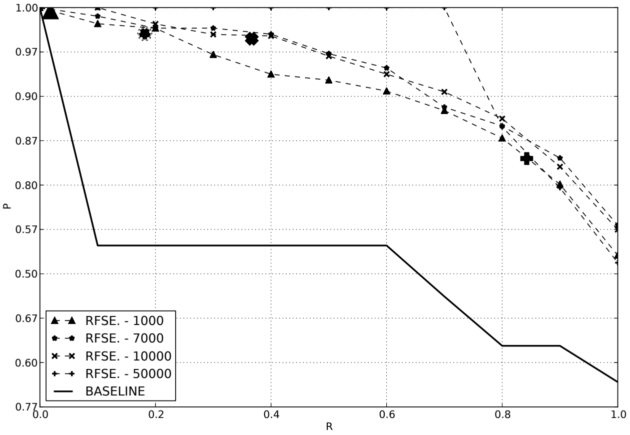

Fig. 4. Performance of RFSE models on *KI-04* corpus, for character 4-grams. The baseline refers to the best OC-SVM model for the same corpus.

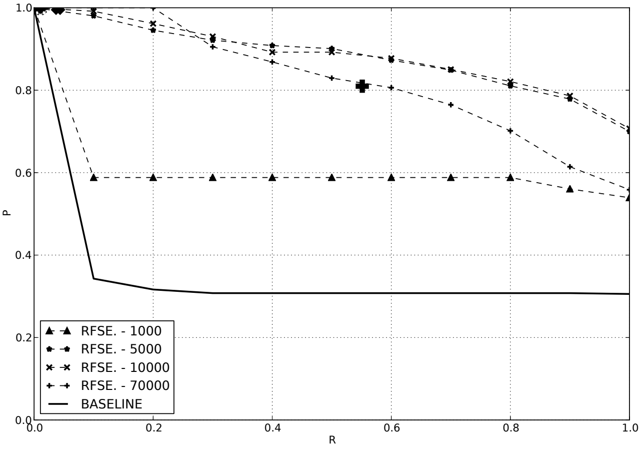

Fig. 5. Performance of RFSE models on *KI-04* corpus, for words. The baseline refers to the best OC-SVM model for the same corpus.

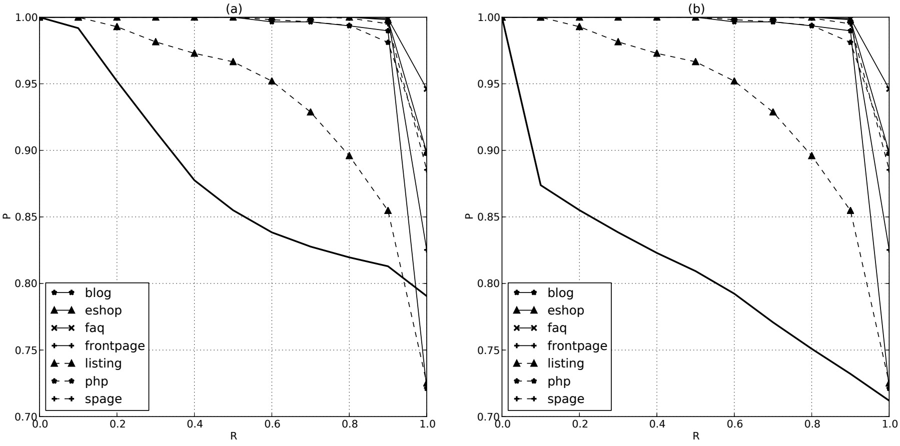

Fig. 6. Results of the RFSE models per genre for 7-genre using (a) character 4-grams and (b) word features. Baseline refers to the best OC-SVM model on the same corpus.

5.4 Genre Analysis

So far, we examined the overall performance on the whole corpus. In this section, we discuss the behavior of the RFSE method for each genre separately. In particular, figure 6 depicts the precision-recall curves of each individual genre of the 7-genre corpus using the character 4-gram and word features. In addition, the best corpus-level performance of OC-SVM based on the corresponding text representation method is used as baseline.

It is clear that most of the genres have a similar performance curve with the exception of Listing which is the worst case. Note that this genre label can be decomposed into some consistent genres. Therefore it can be considered as a super-genre. Previous studies have also find it the most difficult to correctly identified in this collection [10,5]. On the other hand, the best performing genres are FAQs, Online newspaper frontpages, and e-shops.

As concerns the two types of text representation, character n-grams slightly outperform word features in most cases. Interestingly, words are significant better than character 4-grams in identifying the super-genre Listing pages and significantly worse in identifying Search pages.

6 Conclusion and Future Work

In this study, we focus on the AGI task but in contrast to the state-of-the-art in this field, we consider it as an open-set classification problem. This is particularly suitable to AGI applications since there is not agreement over the set of existing web genres and, moreover, they are constantly changing. In this framework, two algorithms are examined, one based on one-class SVM and the other a modification of a method applied to a similar problem of style-based text categorization, that is author identification.

Results on two small-size corpora show that RFSE is far more accurate than OC-SVM for this task. The main idea of this algorithm is that a random subset of the features is likely to be able to show the main stylistic properties of the document. When this procedure is repeated many times, the most likely genre will prevail. It has been demonstrated that for most web-pages there is a clear winner especially when the size of each feature subspace is relatively big (i.e. more than half of the feature set). Another advantage of this approach is that it does not depend on a small set of features and can handle high dimensional feature spaces. That way, it is more robust in case some features are under-represented in some web-pages.

The presented AGI methods were tested as single-class classifiers. However, it is easy to extend these models to provide multiple answers per web page. In particular, it is possible to assign a weight and rank these answers. That way, web pages where multiple genres co-exist can also be handled.

The presented experiments were based on content information of web pages. We tested two types of text representation features, character 4-grams and words. Although both are competitive, character 4-grams seem to provide more effective and robust models. Other types of information coming from the structure and presentation of the web page, the URL etc. can also be included to the proposed models.

Another interesting future work direction would be to compare the RFSE model with other one-class classifiers, especially the approach presented by [16]. Moreover, larger corpora including richer genre palettes suitable for certain applications of AGI are needed to evaluate the open-set classification models.

References

1. Rosso, M.: Using genre to improve web search. PhD thesis, University of North Carolina at Chapel Hill (2005)
2. Braslavski, P.: Combining relevance and genre-related rankings: An exploratory study. In: Proceedings of the International Workshop Towards Genreenabled Search Engines: The Impact of NLP, pp. 1–4 (2007)
3. Sharoff, S., Wu, Z., Markert, K.: The web library of babel: evaluating genre collections. In: Proceedings of the Seventh Conference on International Language Resources and Evaluation, pp. 3063–3070 (2010)
4. Santini, M., Sharoff, S.: Web genre benchmark under construction. Journal for Language Technology and Computational Linguistics 24(1), 129–145 (2009)
5. Kanaris, I., Stamatatos, E.: Learning to recognize webpage genres. Information Processing & Management 45(5), 499–512 (2009)
6. Dong, L., Watters, C., Duffy, J., Shepherd, M.: Binary cybergenre classification using theoretic feature measures (2006)
7. Feldman, S., Marin, M., Medero, J., Ostendorf, M.: Classifying factored genres with part-of-speech histograms. In: Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers, Association for Computational Linguistics, pp. 173–176 (2009)
8. Koppel, M., Schler, J., Argamon, S.: Authorship attribution in the wild. Language Resources and Evaluation 45(1), 83–94 (2011)
9. Meyer zu Eissen, S., Stein, B.: Genre Classification of Web Pages. In: Biundo, S., Frühwirth, T., Palm, G. (eds.) KI 2004. LNCS (LNAI), vol. 3238, pp. 256–269. Springer, Heidelberg (2004)
10. Santini, M.: Automatic identification of genre in web pages. PhD thesis, University of Brighton (2007)
11. Lim, C.S., Lee, K.J., Kim, G.C.: Multiple sets of features for automatic genre classification of web documents. Information Processing and Management 41(5), 1263–1276 (2005)
12. Mason, J., Shepherd, M., Duffy, J.: An n-gram based approach to automatically identifying web page genre. In: HICSS, pp. 1–10. IEEE Computer Society (2009)
13. Khan, S.S., Madden, M.G.: A Survey of Recent Trends in One Class Classification. In: Coyle, L., Freyne, J. (eds.) AICS 2009. LNCS, vol. 6206, pp. 188–197. Springer, Heidelberg (2010)
14. Scholkopf, B., Platt, J., Shawe-Taylor, J., Smola, A., Williamson, R.: Estimating the support of a high-dimensional distribution. Technical Report MSR-TR-99-87 (1999)
15. Manevitz, L., Yousef, M.: One-class svms for document classification. The Journal of Machine Learning Research 2, 139–154 (2002)
16. Anderka, M., Stein, B., Lipka, N.: Detection of text quality as as a one-class classification problem. In: 20th ACM International Conference on Information and Knowledge Management (CIKM 2011), pp. 2313–2316 (2011)
17. Ferretti, E., Fusilier, D., Cabrera, R., y Gómez, M., Errecalde, M., Rosso, P.: On the use of pu learning for quality flaw prediction in wikipedia. In: Working Notes, CLEF 2012 Evaluation Labs and Workshop, Rome, Italy, 17-20 (2012)
18. Bishop, C.: Pattern Recognition and Machine Learning, 331–336 (2006)