# Web Genre Identification Using Open-class Models

No Author Given

No Institute Given

**Abstract.** Genre identification of web texts fits an open-set classification task. In this approach fundamental problems such the lack of consensus in the web genre pallet and the noise notion, which overwhelm the domain can be easily bypassed. In this paper web documents not belonging to any predefined genre or where multiple genres co-exist is considered as noise. However, the method is adaptable to the noise notion depending on the application. In this work we study the impact of noise on web genre identification within an open-set classification framework. We examine alternative classification models and document representation schemes based on the most common corpora used in the domain; two without noise and one with noise. It is shown that the recently proposed RFSE model is more robust with noise, compare to One Class SVM. Moreover, we show how noise is affecting performance of all the methods under consideration. Additionally, we show how the use of different similarity measures can affect performance of the RFSE depending on the objectives of the application.

## 1 Introduction

Web Genre Identification (WGI), a.k.a Automated Genre Identification of Web pages, is web documents taxonomy task associating web page with their form, communicative purpose and style rather than content. Although debatable in its notion, genre utility is recognized but several scientific communities, form pure linguistics to information retrieval. In particular it is recognized as the spontaneous (Web) texts taxonomy in order to accelerate the social communication procedure.

Despite the difficulties, there is a great amount of work on WGI because, the ability to automatically recognize genre of web documents can enhance modern information retrieval systems by providing genre-based grouping/filtering of search results or intuitive hierarchies of web page collections. The most notorious difficulties of the WGI domain are; (1) the definition of a common genre pallet which inherits on the idiosyncrasy of the genre notion itself (what exactly genre is and how many different genres and sub-genres exist); (2) the definition of genre unit, whether it is the web-page, the web-page section, or the neighbouring web-graph; (3) what is the negative samples, i.e. it is very ease to collect some samples of a genre class it is impossible to collect the complimentary sample of this class; (4) consequently of (2) and (3) it is very hard to define *the notion of*

*genre noise* which as we conclude is inextricably connected to the application [20, 19, 27, 12].

Traditionally and event most recently, WGI has been viewed as a closed-set classification problem, i.e. a machine learning (or similarity based) method has been used for predicting the proper class for an arbitrary web page (or other web unit as explained above) amongst a predefined genre pallet.

Only recently, it has been suggested that AGI better fits an open-set classification task since in any practical application it would not be easy to predefine the whole set of possible genres [21]. All web documents not belonging to one of the predefined classes or documents where multiple (known or unknown) genres co-exist can be viewed as noise in WGI [25]. However, it has not been studied yet *how such noise affects the effectiveness of WGI in an open-set scenario.*

In this paper we focus on measuring and analysing *the impact of noise in open-set WGI.* In particular, similar to [21], we are testing two open-set models *Random Feature Subspacing Ensembles* (RFSE) and *One-Class Support Vector Machines* (OC-SVM). We are applying these models to corpora without noise and another corpus with noise and examine differences in performance. The experiments indicate that both models are affected, RFSE still outperforms OC-SVM while the extracted results are more realistic. Other contributions of this paper are the examination of alternative text representation schemes for both WGI models. In addition we examine three different similarity measure cases and how they affecting the RFSE performance on certain genres.

## 2 Previous Work

In linguistic studies there is a great debate in defining *the notion of genre* as an *abstract categorization* of texts and the relation between them. Despite the methodological differences the linguistic community concluded that the idiosyncrasy of the *genre taxonomy* is mutable and diverse [5]. The idiosyncrasy of the *genre taxonomy* yields due to the spontaneous genesis of the genre classes through the socio-centric interaction which emerging for describing the texts in order to accelerate the social communication procedure.

In consideration of *web genres taxonomy* it has been also eloquently analysed the utilities and the difficulties for the web users. It has been pointed out that the genre taxonomy is summarising the type and the style of the text in a single term as a communicative act [This conclusion cited also in [7]]. Although people can do that efficiently, when one has to report about the attributes of the text that lead to the decision there is a great confusion. In fact there is a plethora of textual, stylistic and conceptual terms which are different per individual and/or per group (e.g. teachers, scientists, engineers) for the same or similar web pages. [6].

Overcoming the difficulties related to the genre taxonomy pointed out in linguistic and empirical studies, in text in *text categorisation* there is a great amount of work related to the automated categorisation of texts based on *genre taxonomy.* Although, starting from fundamentally different routes than the lin-

guistic studies, they ended up with the same notion of genre, which is eventually having two complimentary meanings. That is, *style* and *genus*, which in Greek means *type* or *class* [30, 5].

The aforementioned variations of the genre taxonomy's notion is related more to the methodology and the objective of the text categorisation task's specification rather than the philosophical difference. Particularly in author attribution domain there is a focus on identifying the *style of the author* [29, 15, 16]. On the other hand in the information retrieval (IR) domain, the interest is to classify the texts based on a predefined *genre pallet*, thus the interest is focused on the *style of the authors group*, such as scientists, journalists, bloggers, etc. In the domain of *web genre identification (WGI)* (a.k.a. Automated Genre Identification or AGI), *the web genre taxonomy pallet* which mostly used for research has been formed in a top-down approach [6]. That is a group of expert with a well defined scientific procedure are forming and refining a set of genre tags, for every experiment.

After a significant amount of work related to WGI for about two decades since the first effort of [] there is a consensus about the components which are mainly define the web genres. That is the *form, the function/purpose, the context* of the text and *the groups defining, forming, and using the genre tag/category.* As mostly reported *context* in this case it doesn't have the meaning of *topic.* On the contrary context is a *convolved notion* of *the author's style* and *the subject* the author is deals. This coincide to the linguistic outcome in respect of the components of the genre's identification [5, 6, 12, **?**][THERE ARE SEVERAL MORE].

In the very early studies it has been pointed out that the users are the most important component in the WGI research. There there was an effort of several user studies for eliciting the mechanics they in the process of *genre tagging.* As expected, due to the earlier linguistics studies on genres, the results on user agreement were very discouraging, in categorising the web pages upon their genres and there was also great luck of agreement on their criteria even when agreement in categorisation was occurring [22, **?**,**?**][THERE IS ONE MORE]. Knowing about the luck of consensus on the linguistics studies about the notion of genres and the lack of user agreement about the genre classed, WGI research has been focused on all the other also critical aspects.

Several aspect of the WGI has thoroughly been studied [20, 19, 13, 28, 27, 12]. The *document representation* , i.e. character n-grams, word n-grams and part-of-speech features. The *feature selection* i.e. frequency-based, chi-square, information gain, mutual information, random selection etc. The *term weighting schema*, i.e. TF, TF-IDF, Binary and most recently the TF-IGF. The later is a special weighting schema where the occurring terms (e.g. words), in a representative sample of a web genre, are *genre specific* therefore they are weighted differently for improving the performance of the classification method for the WGI task [30].

In respect to the *classification models* has been used in WGI research, are mostly based on SVM, decision trees, neural networks, naive Bayesian classifiers,

etc, machine learning models [27, 12, 8, 23, 17, 9, 19, 10]. Most recently 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 [1].

In the WGI task experiments there is a pre-procession part where the web pages are disassembled into their main components, *i.e text, the HTML structure and the Hyper-Links*. The main components of the web pages are individually utilised or in combination for farther analysis or for multi-class classification. It has been reported that the combination of all the parts of the pages yields the best performance for the classification methods used in these studies. However, the most informative part is as expected the text alone which is sufficient for the task while the rest just contributing in performance increment [12, **?**][SOME MORE MAYBE?]. In addition there are but a few is at least one study we are aware of where URL alone used for the WGI task where the results where poor but very impressive given the amount of information is used compare to the textual information [][STATHIS TOLD ME]. Moreover, the textual information of the URL itself has been used in combination with other textual information with very encouraging results in WGI applications [7].

A recent alternative approach the WGI task is exploiting the natural connection of the web pages, i.e. *Hyper-Links*. In the aforementioned study the information carried by the Hyper-links themselves (together with their anchor text) are exploited. However, the interconnection between the web pages, a.k.a *the web graph*, can be also exploited, where this interconnection is the implicit connection of the web pages of the same type, i.e. the genus or genre. An ranking algorithm named GenreSim has used this additional information in combination to the textual (and structural information) for improving performance [31]. On the contrary to the URL it self in this study the interconnection of the web pages is exploited similarly to the Google's PageRank and other's similar methods concept. There is a other study also is exploiting the the web-graph and the implicit genre relation of the web pages. In particular based on type assumption is using a semi-supervised methodology where for an arbitrary web page the textual information of its neighbour pages (e.i. ancestors, descendants, and siblings) is employed form improving the genre taxonomy compare to single-page textual information results [2]. In this study it is assumed that the web pages of the same genre are having the idiom of *homophily*, which assumes that neighbour pages having the same genre should have the same content as content defined above in the context of genre identification.

On the above approaches the web page is the main unit under evaluation, which will be classified in the proper genre amongst the genres of an expert based (and predefined) genre pallet. However, there is an other approach where *the page section* is the unit under consideration. Consequently, it is assumed that a web page might be multi-genre... [][STATHIS SENT ME THIS PAPER]

As we are aware of all the studies on WGI are based on a *closed-set* approach. That is, the classifiers had to predict that an arbitrary web pages is belonging to one of the predefined web genres, irrespectively of the models; either a multi-class

classification model or a similarity-based model [19]; either exploiting the single-page or graph-based information. To the best of our knowledge there is only one recent work where WGI is applied on an *open-set* classification framework [21]. The novelty of this work is that the WGI problem has been tackled based on the approach of NLP world, in particular *author identification* [15, 16]. Two *open-set classification models* were tested, the *Random Feature Space Ensemble (RFSE) and the One class SVM (OC-SVM) ensemble*. The former was by far better in experiments based on three corpora, where one of them was including a set of noise-full web pages in the context of genre taxonomy. However, the used corpora were designed for closed-set WGI, that is all web pages belong to one of the predefined genres. There is an other work based on the one-class classification approach in the context of WGI, however the framework was fundamentally different... [18] ??? [OCSVM Method Paper]

**The Indeterminate Negative and Noise Samples in WGI** Noise is an other problem added to the fundamental issues of the notion of genre, which in the first place motivated the WGI research. While the genre are constantly reassessed, reformed, abandoned and emerged by the users them shelves, noise is autonomously emerging and is mutated depending on the way we deal with the problem of *genre identification*.

In previous studies related to noise in WGI, noise has been defined as the set of web pages not belonging to any of the known genres of the corpus in [28, 13, 8]. In these works noise was used as negative examples for training binary classifiers. In a more resent work Santini in [20] (or [][INDIVIDUAL REF]) defines *noise-set* as a collection of web-pages having *no genre* or *multiple genres* same as the non-noise genres of the corpus. Thus, the notion of noise it has been used interchangeably with *the notion of negative samples* in the closed-set classification framework has been used in the history of WGI research, while the close-set approach is the natural approach when one thinks genre tags as genus tags or type tags.

Our conclusion is that this ambiguity can be clarified based on the genre unit's and pallet's properties which has to be very clear a priory of a set of experiments. In particular one has first has to answare in the question; "Which is my genre unit?", i.e. the web site, the web page, the web section, or any other division one might think of. The second question is the properties of the genre pallet, i.e. whether or not the genre pallet will be pre-defined and temporally stable and whether or not the pallet will be open or not.

In this study, we are following the properties are derives by the above Santini's definition of noise, i.e. no-genre or multi-genre. In particular for our experiments, we assure that the genre pallet is open, thus, there are genre that we are not aware of. I addition we assume that the genre pallet is not temporally changing thus our available genre pallet (of the corpora we use in this study) is the so called *"common" genre pallet*, i.e. the widely accepted genre pallet of the WGI community. Moreover, we use as our main genre unit, the web page it self.

As we will it will be explained in detail in this study, following the Santini's definition for noise (i.e. samples can be multi-genre on non-common genres) we evaluate the Ensembles behaviour is descried in sections 2 and 1. These ensembles are based on the one-class classification class of machine learning (ML) methods which has been used in several domains other than WGI.

**Novelty detection** *One-class classification* or novelty-detection handles data where only positive examples are available and has been applied to several domains [14]. One-class SVM (OC-SVM) is perhaps the most popular method. The key concept of OC-SVM is based on the $\nu$-SVM model proposed by Scholkopf et al.[26] *considers the origin as the only negative example.* OC-SVM is discussed in section 3.1 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* [18]. 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 [18]. On the contrary, the latest work in [21] the results were very encouraging in the context of WGI task.

**WGI applications** There are already several effort for showing the utility of WGI in other related applications, for example there is study on *genre aware crawling* where a similarity based method was intercepted into the crawling procedure in order to drive *a focused web crawling* procedure on specific genres. The results where very encouraging, additionally, their page selection procedure was in a broad sense similar to our *open-set classification method*, however, the topic and the genre were really convolved thus our results are not comparable at all[7]. An other very interesting application is the efforts of using the genre taxonomy for improving the query results ranking on a typical IR task. In particular genre taxonomy is employed for designating a priori the *formality level* of the web texts, then the *formality ranking* is linearly combined with the query results ranking for improving the results. The utility of this method is outlined in this study, together with the conclusion that the more formal an arbitrary web text the more relevant to the query it is [4].

## 3 Ensembles Description

### 3.1 One-class SVM Ensemble

One-class SVM is actually an $\nu$-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 ($\nu$-SVM) is providing an alternative *trade-off control method of misclassification*, proposed from Scholkopf

et al. [26]. In $\nu$-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 analysed in [3], 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 $\nu$ in $\nu$-SVM has the flowing properties:

- $\nu$ is an upper bound on the fraction of *Outliers*.
- $\nu$ is a lower bound on the fraction of *Support Vectors*.
- $\nu$ 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. [26] 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 whether 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", details in 1. We used the *scikit-learn Python package* to implement this method. [1].

**Algorithm 1.1:** The *OC-SVM Ensemble* algorithm.

**Data**: a set of known web-pages $W_g$ for each $g$ of $G$ genres, $C_g$ class for genre $g$, $w$ an unknown web-page vector, $F$ the available feature set, $\boldsymbol{\nu}$ the nu hyper-parameter of OC-SVM,

**Result**: $w \in \{C_g, \emptyset\}$

```
for each genre g in G do
    Train OCSVM(W_g, ν), in vector space F with hyper-paramenter ν to
    build the Learner of C_g class;
end
for each genre g in G do
    OCSVM_score(g) ≡ the highest positive distance of the unknown page w
    from the hyperplane of the Learner. The higher the value the closer to
    the centre of class C_g the w page is located.;
    if OCSVM_prediction(g) > 0 then
        argmax( OCSVM_score(g) ) ≡ C_g, i.e. w ∈ C_g;
    else
        w ∈ ∅, i.e. "Don't Know" or OTHER;
    end
end
```

### 3.2 Random Feature Space Ensemble Algorithm

Our ensemble-based algorithm is a variation of the method presented by Koppel et al. [15] 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* 2.

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

**Algorithm 1.2:** The *Random Feature Sub-spacing Ensemble* algorithm.

**Data**: a set of known web-pages $W_g$ for each $g$ of $G$ genres, $c_g$ a vector representative for each genre class $C_g$, $w$ an unknown web-page vector, $V$ the available vocabulary, $F$ a feature set, $I$ a number of iterations, $\boldsymbol{\sigma}$ the decision threshold

**Result**: $w \in \{C_g, \emptyset\}$

```
for each genre g in G do
    Average known web-pages W_g of genre g to build one centroid vector c
    representing C_g class;
    repeat
        a.Randomly choose some fraction F of the full feature set V;
        b.Find top match of unknown page W in centroid vectors c using
        some Similarity(c, w) measure;
    until I times;
end
for each genre g in G do
    Score(g) ≡ proportion of times the w unknown page top matches with
    C_g;
    if max Score(g) > σ decision threshold then
        argmax( Score(g) ) ≡ C_g, i.e. w ∈ C_g;
    else
        w ∈ ∅, i.e. "Don't Know" or OTHER;
    end
end
```

## 4 The Experiments: Ensembles Performance

In this study we mainly focusing on the ensembles performance and behaviour on the WGI task. We use primarily F1 statistic for selecting the optimal parameter combinations and evaluate the behaviour of the ensembles. However, there are several cases where precision is more important than recall, especially in application level (see section 4.8). Thus, we are also examine precision maximisation, while keeping the highest possible recall.

The F1 statistic is the harmonic mean of the precision and recall. Due to the unbalanced nature of our available corpora, besides 7Genres, we have gotten the *macro-averaged precision and recall* values for evaluating the performance of the ensembles in the whole corpus. Since, we are primarily focusing on the behaviour of ensembles we use Precision Recall Curves (PRC) for depicting this in more detail. We have also measured the Area Under the PRCs (AUC) deriving for optimal values which are shown in table 2 together with the respective optimal F1. As it will be explained in section 4.8 it is possible to get an other optimal parameters set when parameters are selected upon F0.5 statistic or AUC maximisation, for application requiring ensembles with high precision.

### 4.1 Experimental set-up

In this paper, we use two corpora already used in previous work in AGI:

1. *7-GENRE* [23]: This is a collection of 1,400 English web pages evenly distributed into 7 genres (blogs, e-shops, FAQs, on-line front pages, listing, personal home pages, search pages). Details in table 12.
2. *KI-04* [9]: This is a collection of 1,205 English web pages (unevenly) categorized into 8 genres (link collection, help, shop, portrayal non-private, portrayal private, articles, download, discussion). Details in table 12.
3. *SANTINIS* [20]: This is a corpus comprising 1,400 English web pages evenly distributed into 7 genres (blogs, e-shops, FAQs, online front pages, listing, personal home pages, search pages), 80 documents evenly categorized to 4 additional genres taken from BBC web pages (DIY, editorial, bio, features) and a random selection of 1,000 English web pages taken from the SPIRIT corpus [11]. The latter can be viewed as noise in this corpus. Details in table 12.

The genre palettes of these corpora have some similarities (e.g. e-shops and personal home pages(P.H.P.) are included in all three) and differences (e.g. 7-genre and SANTINIS comprises blogs while KI-04 doesn't). They have been extensively used in many WGI studies but following the closed-set classification scenario *[24, 27]*, although, SANTINIS has only used for examine *the noise impact* to the WGI task, still more that half of it is consist of *7-genre corpus*. However, the results of these studies are not directly comparable to our results due to the essential difference of open-set and closed-set classification. Moreover, 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 web pages excluding any structural information, URLs, etc. Based on the good results reported in [27, 21] as well as some preliminary experiments, the following document representation schemes are examined: *Character 4-grams*, *Words uni-grams*, *Word 3-grams*.

In our experiments, we do not use the noisy pages at all in the training phase. We only use them in evaluation phase. To obtain results comparable with previous studies, we followed the practice of performing 10-fold cross-validation with these corpora. We use the Term-Frequency (TF) weighting scheme. The feature space is defined by a *Vocabulary which is comprised of the training set terms only*. The training set is selected with *stratified sampling* divided into 10-folds, for conducting cross-validation experiments.

As concerns OC-SVM ensemble, two parameters have to be tuned: the number of features $fs$ and $\nu$. For the former, we used $fs = \{$*1k, 5k, 10k, 50k, 90k*$\}$, of most frequent terms of the vocabulary. Following the reports of previous studies [26] and some preliminary experiments, we examined $\nu = \{$*0.05, 0.07, 0.1, 0.15, 0.17, 0.3, 0.5, 0.7, 0.9*$\}$. In comparison to [21], this set of parameter values is more extended.

With respect to RFSE, four parameters should be set: the vocabulary size $V$, the number of feature used in each iteration $f$, the number of iterations $I$,

and the threshold $\sigma$. We examined $V$=$\{$*5k, 10k, 50k, 100k*$\}$, $f$=$\{$*1k, 5k, 10k, 50k, 90k*$\}$, $I$=$\{$*10, 50, 100*$\}$ (following the suggestion in [15] that more than 100 iterations does not improve significantly the results) and $\sigma_s$=$\{$*0.5, 0.7, 0.9*$\}$ (based on some preliminary tests). Additionally, in this work we are testing two document similarity measures: cosine similarity (similar to [21]) and MinMax similarity (used also in a similar task by [16]).

Related to the Spirit 1000 which is the noise sample of the SANTINIS corpus it has used only for evaluating the noise tolerance on the ensembles. Therefore, there was no learner trained for recognising the noise pages neither for RFSE nor for OCSVM ensemble. Due to the 10-fold cross-validation applied in all of our experiments *the whole amount of 1000 noise pages were repeatedly tested for every fold. Thus in our evaluation measures the Spirit 1000 pages are counted as being 10,000.*

Finally, a grid-search ran, for RFSE, of *3240 experiments* for every combination of parameters, i.e. document representations, distance measures and the three (3) available corpora, described in section 4.1. In addition, for OCSVM Ensemble, a grid-search for *405 experiments* also ran.

### 4.2 OCSVM Ensemble: Baseline Selection

As said in section 2 OCSVM has been widely used for one-class classification tasks, consequently, its ensemble (OC-SVM ensemble) form, described in section 3.1, is used in this study as the baseline for examining the behaviour of the ensembles in WGI task.

In table 1 the are summarized the best results OC-SVM ensemble could achieve in respect to F1 statistic maximisation, on the three available corpora described in 4.1. The parameter combination for these results are included together with the document representation selected each time and the AUC for this combination.

| Corpus | Doc. Rep. | Feat. | $\nu$ | Pre | Rec | F1 | AUC |
|---|---|---|---|---|---|---|---|
| 7Genres | 3W | 5,000 | 0.1 | 0.711 | 0.521 | 0.601 | 0.525 |
| | 1W | 5,000 | 0.07 | 0.658 | 0.521 | 0.582 | 0.432 |
| | 4C | 50,000 | 0.1 | 0.701 | 0.565 | 0.626 | 0.53 |
| KI04 | 3W | 5,000 | 0.3 | 0.583 | 0.242 | 0.342 | 0.177 |
| | 1W | 1,000 | 0.1 | 0.551 | 0.308 | 0.395 | 0.176 |
| | 4C | 90,000 | 0.07 | 0.502 | 0.316 | 0.388 | 0.185 |
| SANTINIS | 3W | 10,000 | 0.1 | 0.631 | 0.654 | 0.643 | 0.3 |
| | 1W | 50,000 | 0.3 | 0.259 | 0.283 | 0.27 | 0.094 |
| | 4C | 5,000 | 0.5 | 0.662 | 0.367 | 0.472 | 0.334 |

**Table 1.** Optimal combinations of parameters for OC-SVM ensemble and the respective F1 statistic, after Precision/Recall macro-averaging, for each combination. AUC has been calculated under the full PR curves, i.e. PR curve has one point for every web page of the corpus.

Considering the performance of the OC-SVM ensemble, it seems WGI to be a hard task for this ensemble because, on every corpus it is clear form the figures 1, 2 and 3, that both precision and recall is lower than one would expect in a real worlds application.

Starting with the performance of the OC-SVM ensemble on 7Genre corpus (see figure 1) it seems that word 3-grams is a better choice considering that the precision remains high up to 50% of the corpus, however, then it drops to 0. On the other hand the highest F1 is for character 4-grams (4C) as document representation, where the precision is as good as for word uni-grams (1W) up to 40% of the corpus. Then the curve for 1W drops significantly while for 4C the precision remains the same (about 0.78) up to 70% of the corpus web pages and then drops to 0, again.

Considering the parameters selection, one might has to choose between the third and the first row, of table 1, for selecting the parameters for tuning the OC-SVM ensemble upon the specification of a particular WGI task. On the one hand, selecting the parameters of the third row the PRC with the stars, in figure 1, is occurring where F1 and AUC is maximised as shown in the table. On the other hand selecting the first row precision is maximised for the 50% of the documents and then drops to 0. As we will see in section 4.8 for an application such as *a genre aware web crawler* this would be the case, where we would be interested in achieving no less than 0.88 precision and we wouldn't mind for omitting the 50% of the web pages belonging to the genre of interest.

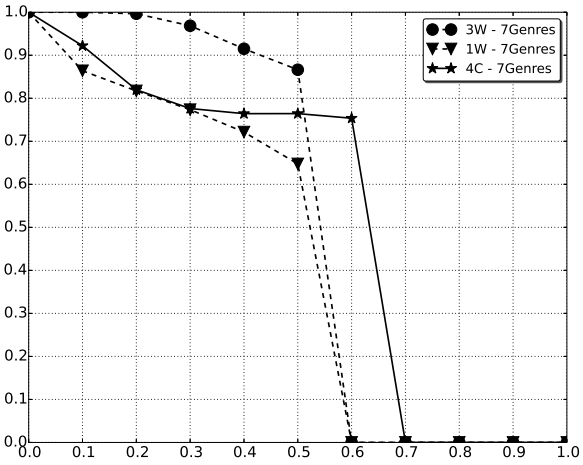


**Fig. 1.** Optimal Precision-Recall Curves of OC-SVM ensemble based on parameters found in **table 1** for 7Genres corpus. The annotation boxes shows the F1 and the AUC of each PRC, for every document representation respectively. The grey said annotation box indicates that the pointed PRC corresponds to parameter combination for which the OC-SVM ensemble achieved the best F1.

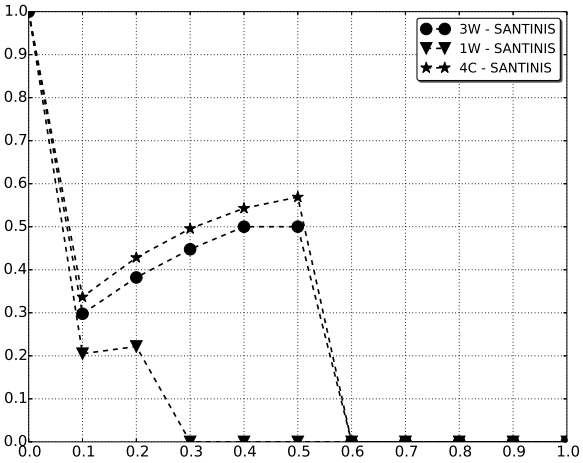


**Fig. 2.** Optimal Precision-Recall Curves of OC-SVM ensemble based on parameters found in **table 1** for SANTINIS corpus. The annotation boxes shows the F1 and the AUC of each PRC, for every document representation respectively. The grey said annotation box indicates that the pointed PRC corresponds to parameter combination for which the OC-SVM ensemble achieved the best F1.

In figure 3 the results on KI04 corpus are depicted. As one can tell also by the table 1, neither F1 nor the precision could be maximised, while for most of the corpus the performance of the OC-SVM ensemble is worst than random selection. The best precision achieved by OC-SVM ensemble on KI04 corpus is 0.65 for only 20% of the corpus while for the rest it drops to 0.

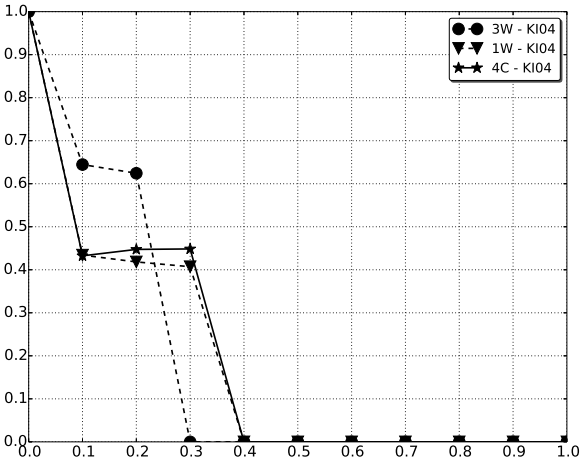


**Fig. 3.** Optimal Precision-Recall Curves of OC-SVM ensemble based on parameters found in **table 1** for KI04 corpus. The annotation boxes shows the F1 and the AUC of each PRC, for every document representation respectively. The grey said annotation box indicates that the pointed PRC corresponds to parameter combination for which the OC-SVM ensemble achieved the best F1.

Before we proceed in evaluating the performance behaviour of OC-SVM ensemble in the noise-full case, we have to reason the sharp drop of PRC curves to 0 precision. As explained above in section 3 the ensembles are following an open-set approach where they can select either to classify the web pages on one of the genres in the pallet their are trained on, or they can let the web pages to the pile of "Don't Know" pages, a.k.a. OTHERS. In all the PRCs presented in this study, the braking point where the curves are dropping to 0 precision is the point after which the ensembles have decided that the web pages are remaining as unknown for the rest of the corpus. To put it in an other way, The PRC are showing how strictly the ensembles has to be tuned-up for keeping the precision as high as possible, as a result lowering the recall level as low as it is required.

As for the behaviour of the OC-SVM ensemble on SANTINIS corpus. In figure 2 the PRCs show a significant drop in performance compare to the performance on 7Genres corpus which in fact is the noise-free version of SANTINIS. In best case, when 4C is used as document representation OC-SVM ensemble can achieve up to about 0.59 precision and for only 50% of the corpus. As for the rest of document representations precision drops to random or worst. It has to be noted though that F1 for 3W is much higher than the F1 for 4C as one can see in table 1 which is not actually captured by the PRCs.

More details on reasoning the above will be found in 4.4, a brief explanation is given when one is observing the confusion matrices in tables 3 and 3. If one will sum up the diagonals in these table will find that about 50% of the documents is retrieved for 3W document representation while about 57% is retrieved for 4C

for the respective precision depicted in the curves. This is happening because 4C document representation is forcing OC-SVM ensemble to be much more strict in presence of noise and more efficient, since is letting about 1000 more web pages into the OTHERS pile. However, the recall for ever genre individually is getting really low thus it drops significantly macro-F1, i.e. F1 calculated by macro-precision an macro-recall.

Note also in figure 2, that the PRCs are dropping to a very small value at the begging of the curve and then start to raise until the braking point of zero, as reasoned above. This is a main side effect we observing is cased due to noise presence. In particular, what is happing here is the confusion of a great amount of web pages from the noise part of the SANTINIS corpus, predicted as one of the non-noise genre classes (see tables 3 and 4 first raw and first column).

In the following chapters where the ensembles performance behaviour is farther analysed, OC-SVM ensemble performance is used as the baseline for analysing the RFSE performance. In cases where the results are not directly comparable the OC-SVM ensemble RPC are selected with the highest F1 where recall is the highest possible for the respective precision curve.

### 4.3 Random Feature Space Ensemble in Practice

In table 2 they are summarised the best results of the RFSE algorithm's optimal behaviour in WGI task, on the the three corpora used for the evaluation of the ensembles. In this table the parameters combinations paired with the document representation and the distance measure required, are depicted. The optimal combinations have been selected in respect the maximisation of F1, which they are shown in the table together with their respective AUC.

Considering the F1 statistic values in table 1 F1 statistic is most cases is above 0.72 with the great exception of KI04 where the maximum F1 is 0.612 and the minimum close to 0.55, As for precision exceeds 0.6 in worst case scenario, however it can reach to 0.933 even at noise-full case of SANTINIS corpus. In addition, as we will see in section 4.8 it is possible to achieve even higher precision in application based requirements.

| Dist. Meas. | Corpus | Doc. Rep. | Vocab. | Feat. | $\sigma$ | Iter. | Pre | Rec | F1 | AUC |
|---|---|---|---|---|---|---|---|---|---|---|
| Cosine | 7Genres | 3W | 100,000 | 50,000 | 0.5 | 50 | 0.809 | 0.756 | 0.782 | 0.835 |
| | | 1W | 100,000 | 50,000 | 0.5 | 100 | 0.776 | 0.696 | 0.733 | 0.733 |
| | | 4C | 50,000 | 5,000 | 0.5 | 50 | 0.79 | 0.723 | 0.755 | 0.829 |
| | KI04 | 3W | 100,000 | 90,000 | 0.5 | 50 | 0.664 | 0.518 | 0.582 | 0.497 |
| | | 1W | 100,000 | 50,000 | 0.5 | 50 | 0.704 | 0.536 | 0.609 | 0.5 |
| | | 4C | 10,000 | 1,000 | 0.5 | 10 | 0.69 | 0.468 | 0.558 | 0.473 |
| | SANTINIS | 3W | 50,000 | 10,000 | 0.7 | 100 | 0.797 | 0.722 | 0.758 | 0.768 |
| | | 1W | 50,000 | 5,000 | 0.5 | 100 | 0.839 | 0.702 | 0.765 | 0.77 |
| | | 4C | 50,000 | 1,000 | 0.5 | 100 | 0.739 | 0.78 | 0.759 | 0.606 |
| MinMax | 7Genres | 3W | 100,000 | 90,000 | 0.5 | 10 | 0.777 | 0.764 | 0.77 | 0.819 |
| | | 1W | 5,000 | 1,000 | 0.5 | 10 | 0.761 | 0.664 | 0.709 | 0.711 |
| | | 4C | 10,000 | 5,000 | 0.5 | 100 | 0.768 | 0.74 | 0.754 | 0.807 |
| | KI04 | 3W | 100,000 | 90,000 | 0.5 | 50 | 0.663 | 0.529 | 0.589 | 0.487 |
| | | 1W | 10,000 | 5,000 | 0.5 | 100 | 0.659 | 0.572 | 0.612 | 0.572 |
| | | 4C | 10,000 | 1,000 | 0.5 | 100 | 0.711 | 0.512 | 0.595 | 0.48 |
| | SANTINIS | 3W | 50,000 | 5,000 | 0.7 | 100 | 0.933 | 0.68 | 0.787 | 0.893 |
| | | 1W | 100,000 | 10,000 | 0.7 | 100 | 0.826 | 0.69 | 0.752 | 0.866 |
| | | 4C | 100,000 | 5,000 | 0.9 | 100 | 0.864 | 0.682 | 0.762 | 0.862 |
| Cos & MinMax | 7Genres | 3W | 100,000 | 50,000 | 0.5 | 100 | 0.793 | 0.757 | 0.775 | 0.835 |
| | | 1W | 10,000 | 5,000 | 0.5 | 100 | 0.769 | 0.68 | 0.722 | 0.73 |
| | | 4C | 10,000 | 1,000 | 0.5 | 50 | 0.81 | 0.725 | 0.765 | 0.831 |
| | KI04 | 3W | 100,000 | 90,000 | 0.5 | 50 | 0.643 | 0.528 | 0.58 | 0.489 |
| | | 1W | 10,000 | 5,000 | 0.5 | 100 | 0.675 | 0.549 | 0.605 | 0.486 |
| | | 4C | 10,000 | 5,000 | 0.5 | 100 | 0.602 | 0.478 | 0.533 | 0.438 |
| | SANTINIS | 3W | 100,000 | 1,000 | 0.5 | 100 | 0.878 | 0.683 | 0.768 | 0.903 |
| | | 1W | 50,000 | 5,000 | 0.5 | 100 | 0.854 | 0.689 | 0.763 | 0.862 |
| | | 4C | 50,000 | 1,000 | 0.5 | 100 | 0.752 | 0.773 | 0.762 | 0.725 |

**Table 2.** Optimal combinations of parameters for RFSE and the respective F1 statistic, after Precision/Recall macro-averaging, for each combination. AUC has been calculated under the full PR curves, i.e. PR curve has one point for every web page of the corpus. SANTINIS corpus curves has been calculated using strata sampling over the RFSE results of the noise part of the corpus.

Comparing the PRCs of RFSE to the best PRCs of OC-SVM ensemble, we see clearly (in figures 4, 6 and 5) that RFSE outperforms over OC-SVM ensemble on all corpora. Moreover, even in the most difficulty case, i.e. on KI04 corpus, RFSE manages to sustain more than 0.7 precision for the 60% of the corpus as depicted in figure 6.

In response to the performance behaviour of the RFSE on noise, i.e. on SANTINIS corpus, as shown in figure 5, and it will explained later in detain (section 4.4), RFSE seems to be really tolerant to noise on the contrary to OC-SVM ensemble, where its performance drops really low. Moreover, it seems that he distance measure selected for the algorithm is very important factor for the performance of the RFSE.

### 4.4 The Noise Impact

Noise in the web is a very important factor one has to be considered in a potential application of a WGI filter. That is, an application where web pages with the genre of interest would only be presented in a query's ranked results. As explained in section 2 above, there is an ambiguity of that is noise in the context of genres, since in lots of studies it has an interchangeable meaning with the negative samples, where it is also ambiguous and by nature constantly incomplete set of web pages. However, there is already defined a definition of noise by Santini in [25] which we are using it as the research framework for analysing the performance of RFSE and OC-SVM ensemble in the presence of noise. Thus, noise pages are defined the ones where *noise-pages* either they include multiple genres in their context or are not belonging into the *"common" genre pallets.*

In SANTINIS corpus where noise has been introduces we see clearly the advantage of RFSE over OC-SVM ensemble, as shown in figure 5. As explained above 7Genres is a subset of SANTINIS corpus where noise is omitted. There, we seen in figure 1 that OC-SVM ensemble manages to perform quite well with precision over 0.88 at least for 50% for the corpus. In figure 4 we compare the best OC-SVM ensemble's PRC with the PRCs of RFSE. The RFSE PRCs have also been selected based on the F1 maximisation. We clearly see that the RFSE PRCs outperforms OC-SVM's because the preserve their highest precision, i.e over 0.95, for 80% of the corpus.

In the extend with noise corpus of 7Genres, i.e. SANTINIS, in figure 5, OC-SVM ensembles performance drops to random (when F1 is maximised). In the contrary and unexpectedly RFSE performance remains almost as high as in 7Gernes corpus (the noise free one) even on the presence of noise. In particular the drop in precision is less than 0.05 on average, while the recall is even better for MinMax and Comb cases of similarity measures where it reaches 90% of the corpus.

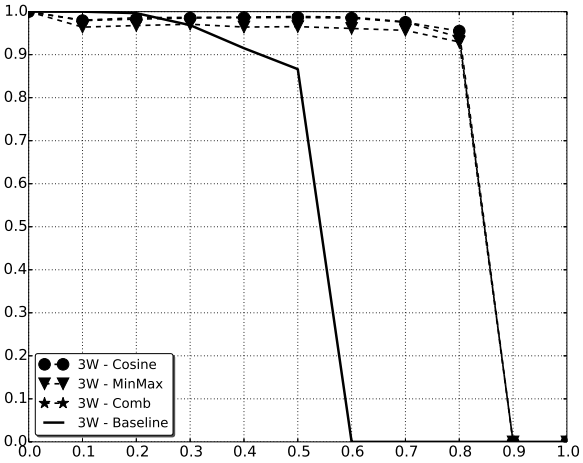


**Fig. 4.** Optimal Precision-Recall Curves of RFSE based on parameters found in **table 1** for 7Genres corpus, using word 3-grams as document representation. The annotation boxes shows the F1 and the AUC of each PRC, for every document representation respectively. The grey said annotation box indicates that the pointed PRC corresponds to parameter combination for which the RFSE achieved the best F1.

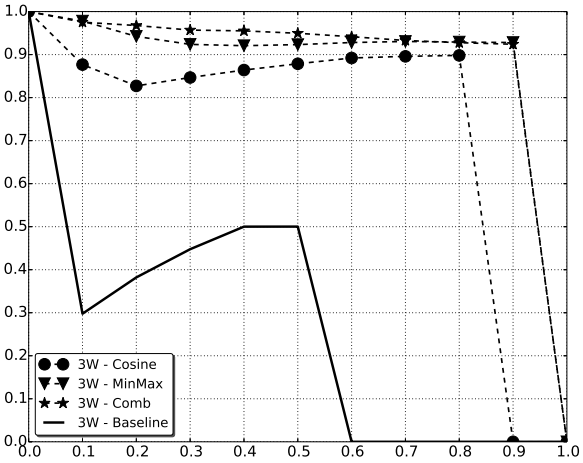


**Fig. 5.** Optimal Precision-Recall Curves of RFSE based on parameters found in **table 1** for SANTINIS corpus, using word 3-grams as document representation. The annotation boxes shows the F1 and the AUC of each PRC, for every document representation respectively. The grey said annotation box indicates that the pointed PRC corresponds to parameter combination for which the RFSE achieved the best F1.

Notice also that, for OC-SVM ensemble as explained in section 4.2 there is a great drop at the beginning of the respective PRC. However, for RFSE this drop is insignificant and only cosine similarity measure while for the rest of the similarity measures there is no effect of noise at the beginning of the curve.

In order to have a better insight we compare the confusion matrices for the optimal parameters of OC-SVM ensemble and RFSE. In tables 3 and 4 for OC-SVM and tables 5, 6 and 7 for RFSE. The first columns of both tables are depicting the prediction for the ensembles for the noise-pages of SANTINIS corpus, when for both the parameters set-up offers the maximum performance in respect to F1 maximisation.

Here we have to note again that, both ensembles have not any learner in order to identify the noise-pages, the learners exists in the ensemble only or the non-noise genres. Therefore, it is implied that the amount of noise pages which the ensembles classified correctly as OTHER, is the proportion of the noise pages that they *correctly left them unclassified*. In addition, we have to note that the first row of the confusion matrices showing the amount of pages which, the ensembles, left unclassified. Moreover the first column is showing the proportion of noise-pages which they have predicted as correctly as OTHERS (i.e. correctly classified) or misclassified as some of the genres from the "common" genre pallet.

Comparing OC-SVM ensemble and RFSE in tables 3 and 5, the later is showing great performance compare to the former. In particular the amount of web pages from the noise-full genre that the RFSE letting correctly as unclassified to the pile of OTHERs pages is almost the whole amount of this pages. On the contrary, about the half of the noise pages are misclassified by the OC-SVM ensemble.

| | Real Classes | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | OTR | G1 | G2 | G3 | G4 | G5 | G6 | G7 | B1 | B2 | B3 | B4 |
| OTR | 4,881 | 77 | 89 | 29 | 16 | 100 | 95 | 73 | 1 | 3 | 0 | 2 |
| G1 | 261 | 117 | 0 | 0 | 0 | 1 | 2 | 0 | 0 | 0 | 0 | 0 |
| G2 | 1,419 | 2 | 99 | 2 | 10 | 14 | 15 | 13 | 1 | 1 | 2 | 0 |
| G3 | 0 | 0 | 0 | 162 | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 |
| G4 | 0 | 0 | 0 | 0 | 158 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| G5 | 2,215 | 1 | 9 | 7 | 11 | 76 | 6 | 11 | 0 | 1 | 0 | 0 |
| G6 | 782 | 2 | 0 | 0 | 0 | 1 | 80 | 2 | 0 | 0 | 0 | 0 |
| G7 | 442 | 1 | 3 | 0 | 5 | 6 | 2 | 101 | 0 | 0 | 0 | 1 |
| B1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 18 | 0 | 0 | 0 |
| B2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 15 | 0 | 0 |
| B3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 18 | 0 |
| B4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 17 |
| SUMs | 10,000 | 200 | 200 | 200 | 200 | 200 | 200 | 200 | 20 | 20 | 20 | 20 |

**Table 3.** Confusion Matrix of OC-SVM ensemble predictions on *SANTINIS* for an optimal parameter set, i.e. 10,000 features, $\nu = 0.1$. Document representation: Word 3-Grams.

| | Real Classes | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | OTR | G1 | G2 | G3 | G4 | G5 | G6 | G7 | B1 | B2 | B3 | B4 |
| OTR | 5,986 | 100 | 104 | 67 | 51 | 101 | 96 | 101 | 10 | 10 | 10 | 11 |
| G1 | 236 | 99 | 0 | 0 | 0 | 2 | 9 | 1 | 0 | 1 | 1 | 0 |
| G2 | 189 | 0 | 86 | 0 | 1 | 1 | 1 | 4 | 0 | 0 | 0 | 0 |
| G3 | 0 | 0 | 0 | 71 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| G4 | 0 | 0 | 0 | 0 | 20 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| G5 | 3,396 | 1 | 8 | 62 | 128 | 95 | 27 | 11 | 0 | 4 | 4 | 5 |
| G6 | 154 | 0 | 0 | 0 | 0 | 0 | 67 | 0 | 0 | 0 | 0 | 0 |
| G7 | 10 | 0 | 2 | 0 | 0 | 1 | 0 | 83 | 0 | 0 | 0 | 0 |
| B1 | 29 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 10 | 0 | 0 | 0 |
| B2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5 | 0 | 0 |
| B3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 5 | 0 |
| B4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 4 |
| SUMs | 10,000 | 200 | 200 | 200 | 200 | 200 | 200 | 200 | 20 | 20 | 20 | 20 |

**Table 4.** Confusion Matrix of OC-SVM ensemble predictions on *SANTINIS* for an optimal parameter set, i.e. 5,000 features, $\nu = 0.5$. Document representation: Character 4-Grams.

| | Real Classes | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | OTR | G1 | G2 | G3 | G4 | G5 | G6 | G7 | B1 | B2 | B3 | B4 |
| OTR | 9,887 | 48 | 178 | 2 | 6 | 194 | 192 | 141 | 0 | 0 | 0 | 0 |
| G1 | 63 | 152 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| G2 | 1 | 0 | 22 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| G3 | 0 | 0 | 0 | 198 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| G4 | 29 | 0 | 0 | 0 | 194 | 2 | 0 | 0 | 0 | 0 | 0 | 0 |
| G5 | 0 | 0 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 0 | 0 |
| G6 | 0 | 0 | 0 | 0 | 0 | 0 | 7 | 0 | 0 | 0 | 0 | 0 |
| G7 | 20 | 0 | 0 | 0 | 0 | 0 | 0 | 59 | 0 | 0 | 0 | 0 |
| B1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 20 | 0 | 0 | 0 |
| B2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 20 | 0 | 0 |
| B3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 20 | 0 |
| B4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 20 |
| SUMs | 10,000 | 200 | 200 | 200 | 200 | 200 | 200 | 200 | 20 | 20 | 20 | 20 |

**Table 5.** Confusion Matrix of RFSE Ensemble predictions on *SANTINIS* for an optimal parameter set, i.e. 50,000 vocabulary size, 5,000 features, $\sigma = 0.7$ and 100 random feature selection iterations. Document representation: Word 3-Grams. Distance measure: MinMax.

Related to the rest of the genres, the non-noise ones, again RFSE it has a more clear off-diagonal confusion matrix than the OC-SVM ensemble. The weakness of RFSE is on G5-Listing, G6-Personal Home Pages and G2-Eshops genres where

| | Real Classes | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | OTR | G1 | G2 | G3 | G4 | G5 | G6 | G7 | B1 | B2 | B3 | B4 |
| OTR | 9,867 | 25 | 177 | 3 | 3 | 192 | 196 | 155 | 0 | 0 | 0 | 0 |
| G1 | 90 | 175 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 |
| G2 | 6 | 0 | 21 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| G3 | 0 | 0 | 0 | 197 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| G4 | 25 | 0 | 0 | 0 | 197 | 2 | 0 | 0 | 0 | 0 | 0 | 0 |
| G5 | 4 | 0 | 0 | 0 | 0 | 4 | 0 | 0 | 0 | 0 | 0 | 0 |
| G6 | 0 | 0 | 0 | 0 | 0 | 0 | 3 | 0 | 0 | 0 | 0 | 0 |
| G7 | 8 | 0 | 1 | 0 | 0 | 0 | 0 | 44 | 0 | 0 | 0 | 0 |
| B1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 20 | 0 | 0 | 0 |
| B2 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 20 | 0 | 0 |
| B3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 20 | 0 |
| B4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 20 |
| SUMs | 10,000 | 200 | 200 | 200 | 200 | 200 | 200 | 200 | 20 | 20 | 20 | 20 |

**Table 6.** Confusion Matrix of RFSE Ensemble predictions on *SANTINIS* for an optimal parameter set, i.e. 50,000 vocabulary size, 5,000 features, $\sigma = 0.7$ and 100 random feature selection iterations. Document representation: Word 3-Grams. Distance measure: Combination of Cosine and MinMax.

| | Real Classes | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | OTR | G1 | G2 | G3 | G4 | G5 | G6 | G7 | B1 | B2 | B3 | B4 |
| OTR | 9,506 | 5 | 129 | 71 | 12 | 186 | 166 | 94 | 0 | 3 | 0 | 0 |
| G1 | 425 | 195 | 0 | 0 | 0 | 1 | 9 | 0 | 0 | 0 | 0 | 0 |
| G2 | 3 | 0 | 71 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| G3 | 0 | 0 | 0 | 129 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| G4 | 9 | 0 | 0 | 0 | 188 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| G5 | 2 | 0 | 0 | 0 | 0 | 11 | 0 | 0 | 0 | 0 | 0 | 0 |
| G6 | 20 | 0 | 0 | 0 | 0 | 0 | 25 | 0 | 0 | 0 | 0 | 0 |
| G7 | 29 | 0 | 0 | 0 | 0 | 1 | 0 | 106 | 0 | 0 | 0 | 0 |
| B1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 20 | 0 | 0 | 0 |
| B2 | 6 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 17 | 0 | 0 |
| B3 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 20 | 0 |
| B4 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 20 |
| SUMs | 10,000 | 200 | 200 | 200 | 200 | 200 | 200 | 200 | 20 | 20 | 20 | 20 |

**Table 7.** Confusion Matrix of RFSE Ensemble predictions on *SANTINIS* for an optimal parameter set, i.e. 50,000 vocabulary size, 5,000 features, $\sigma = 0.7$ and 100 random feature selection iterations. Document representation: Word Uni-grams. Distance measure: Cosine.

the recall is dropping dramatically in order to preserve almost excellent precision. However, in a potential real world application one could improve G2-Eshop and G5-Listing by changing the similarity measure from MinMax to Cosine, in order to improve recall. Unfortunately G6-PHP it cannot be improves because Cosine similarity decrease precision in half compare to MinMax similarity.

As for OC-SVM ensemble, it is clear that half more than half of the noise-pages OC-SVM ensemble has misclassified in several genres either we use 3W (see table 3 or 4), where the worst confusion is for *Eshops* and *Listing* and for *Blogs*, Personal Home Pages and *Search Pages*. The main difference when selecting 4C over 3W is the improvement in precision of non-noise genres, because about 1000 more noise pages are correctly predicted as OTHERs. However, at the same time recall is dropping a lot. The greatest improvement for OC-SVM ensemble with 4C, is for G2-Eshops.

In respect to RFSE performance in the presence of noise, as we will see in the following section, similarity measure (or distance measure) is affecting the results although in over all performance are very similar and equally high. Comparing the PRCs of MinMax similarity and the Comb, i.e. the combination of Cosine and MinMax similarity, in figure 5, the later curve is slightly higher than the MinMax's one. However, as we have seen in table 2 for MinMax, F1 statistic is optimised in higher value than for the Comb. It seems that the combination of both similarity measures increases precision, particularly by improving the noise tolerance as one can tell by observing the first column of confusion matrices table 5 and 6.

More details on similarity measure and document representations impact is discussed in the following chapters.

### 4.5 RFSE: Similarity Measure Impact

In our experiments we have tested the two most common choices of similarity (or distance) measures both previously used in author identification domain [15, 16]. Cosine similarity has also used with success in previous work [21] on open-set WGI, while the MinMax similarity has for the first time tested on WGI task in this study as we are aware of. As we have seen previously, we have conducted some additional experiments combining both similarity measures, which we called it *the Comb*.

In the Comb experiments, the combination of both similarity measures was done in the following way: the RFSE was comparing the normalized values Cosine and MinMax similarities and the decision was taken upon the high *normalized similarity value*. In this way the decision was based on the similarity measure yielding the highest certainty for the decision.

In figure 4 the PRCs of RFSE are depicted for all different similarity measures and their combination for 3W as document representation, where for 3W RFSE achieved the maximum F1. As mentioned above, in section 4.4, the Comb similarity measure is yielding a PRCs better than MinMax in respect to precision maximisation and in particular into *certainty maximisation*.

In the noise-free case, in 7Genres corpus, in figure 4, MinMax similarity and the Comb seems to have almost the same curve, while cosine is significantly lower. In particular it seems that with cosine similarity RFSE is mostly affected by noise. On the contrary, it seems that the Comb is the most noise tolerant set-up for the RFSE and it outperforms the expectations in our experiments by preserving precision as high as 0.93 for 90% of the SANTINIS corpus.

As mentioned above KI04 corpus is the hardest cases for both ensembles, however, RFSE preserves precision over 0.7 compare to worst than random performance of the OC-SVM ensemble. As for the best similarity measure MinMax is the best choice because even thought the performance is almost the same to the cosine similarity, as shown in figure 6, the former manages to recall 70% of the corpus compare to 50% or less for the rest of the cases.

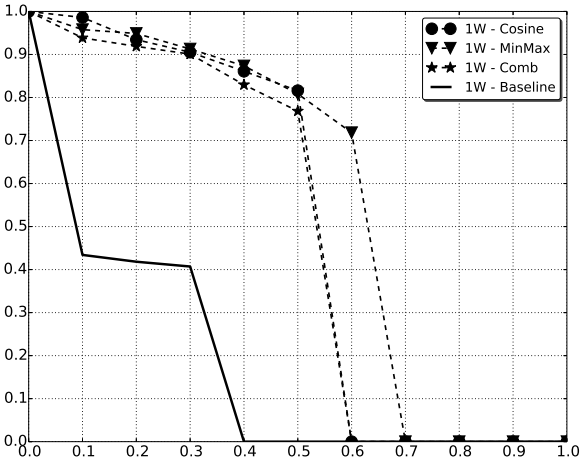


**Fig. 6.** Optimal Precision-Recall Curves of RFSE based on parameters found in **table 1** for KI04 corpus, using word uni-grams as document representation. The annotation boxes shows the F1 and the AUC of each PRC, for every document representation respectively. The grey said annotation box indicates that the pointed PRC corresponds to parameter combination for which the RFSE achieved the best F1.

Finally, note that in figure 6 for the KI04 we have selected to show the performance for different similarity measure using 1W as document representation for which RFSE achieved the maximum performance in respect the F1 statistic, on the contrary to the 7Genres and SANTINIS corpus where 3W is yielding the optimal performance of the RFSE.

### 4.6 Document Representation Impact

In section 4.2 we have seen that there is a difference in performance of the OC-SVM ensemble depending on the document representation. In particular we see that the behaviour of this ensemble is chaining significantly due to the document representation choice, which it varies for the three different corpora.

This is the case for RFSE. In particular observing the figures 7, 8 and 9, we see that document representation is affecting the RFSE performance more in non-noise cases than in noise-full one.

In SANTINIS case for MinMax similarity for which RFSE maximises performance 3W seems to boost performance even farther compare to single words and character 4-grams (see figure 9). However, the performance is preserves almost as high, for all document representations for 90% or the corpus. Again, it seems that 3W is the document representation for which RFSE is less affected due to the fact that the PRC is remaining very high at the beginning of the curves, while for the other two document representations the performance drops at the beginning of the curves.

On the contrary, for 7Genres corpus (see figure 7) recall is affected by 11% increment when 3W or 4C is selected as document representations compare to 1W. In addition, 4C requires about the half of the size of the vocabulary and one tenth (1/10) of the features are required when 3W or 1W are used, while the performance is very close as one case see in table 2. In a real worlds applications one might had to choose between the maximisation of performance or the efficiency of the algorithm in therms or resources required and in this case the weight most likely will favour 4C.

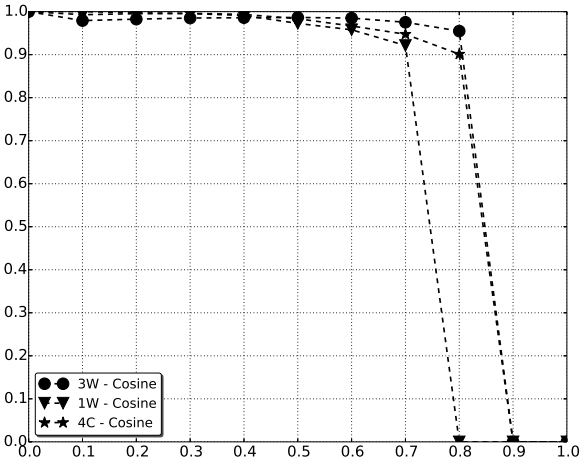


**Fig. 7.** Optimal Precision-Recall Curves of RFSE based on parameters found in **table 1** for 7Genres corpus, using Cosine similarity. The annotation boxes shows the F1 and the AUC of each PRC, for every document representation respectively. The grey said annotation box indicates that the pointed PRC corresponds to parameter combination for which the RFSE achieved the best F1.

As the RFSE performance on KI04, it seems that document representation and similarity measure selection is equally affecting it as it is shown in figures 8 and 6. Together with the results shown in table 2 the combination of 1W and MinMax yields the best performance since is increasing recall by 14% and preserving performance over 0.7. However, in a real worlds application one might had to choose 3W in order to preserve precision over 0.8 when recall wouldn't be so important.

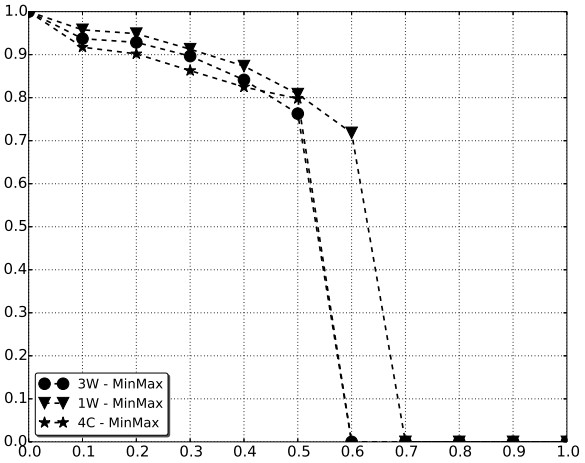


**Fig. 8.** Optimal Precision-Recall Curves of RFSE based on parameters found in **table 1** for KI04 corpus, using MinMax similarity. The annotation boxes shows the F1 and the AUC of each PRC, for every document representation respectively. The grey said annotation box indicates that the pointed PRC corresponds to parameter combination for which the RFSE achieved the best F1.

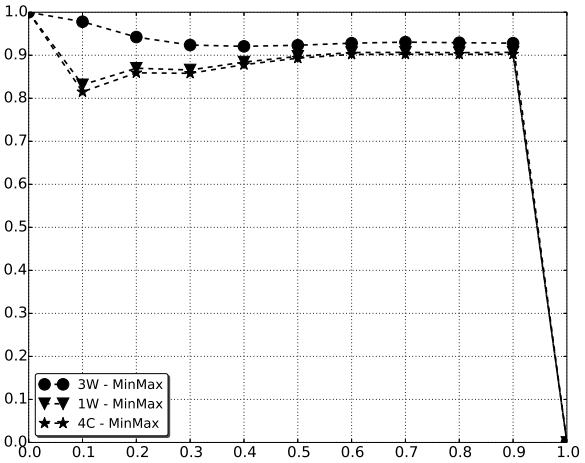


**Fig. 9.** Optimal Precision-Recall Curves of RFSE based on parameters found in **table 1** for SANTINIS corpus, using MinMax similarity. The annotation boxes shows the F1 and the AUC of each PRC, for every document representation respectively. The grey said annotation box indicates that the pointed PRC corresponds to parameter combination for which the RFSE achieved the best F1.

In respect to noise, in SANTINIS corpus, document representation choice seem to be critical in noise tolerance for the first 20% of the corpus. To be specific, as PRCs are formed the predictions with the highest certainty, i.e highest similarity score, are ordered at the beginning of the curve, by Precision-Recall-Curve definition. There, at the first 20%, we see a great difference in precision, (see figure 9), for about 0.07 on average, when 3W document representation is used compare to the 4C and 1W.

Interestingly the same amount of features are required for 3W and 4C document representation, i.e 5,000, as depicted in table 1. On the contrary, for non-noise corpora 4C document representation has a great advantage over the other two in small sized feature set is required for maximising the RFSE's performance.

To conclude, other than the case of OC-SVM ensemble in figure 2 in all other cases and in particular for RFSE precision is maximised using 3W document representation as one can tell in figures 1, 3, 5, 7, 8 and 9.

In the next section we analyse the performance behaviour of the ensembles on genre level, particularly, for the corpora we used in this study.

### 4.7 Genre Level Corpora Analysis

In this section we are focusing on the genre level behaviour of the ensembles. In particular we are interested in the performance of our open-set approach, for the WGI task, *on the "common" genre pallet*.

Starting with 7Genre, the most popular corpus of the WGI domain, in table 8 they are depicted the precision and the recall for the optimal parameter combination in respect to F1 maximization, with the highest F1 from left to right.

Blogs, Eshop and Listing are the most difficult genre case for OC-SVM ensemble, i.e the baseline, in respect of precision, while in respect of recall Personal Home Pages is also a difficult for this ensemble to perform hight. As for RFSE again for the same genres the task is slightly more difficult compare to the rest of the genres. However, for Listing when RFSE is maximising precision up to 0.951, for 3W representation and *the Comb*, precision drops to 0.58. Moreover, as we have seen in section 4.5 for cosine similarity RFSE manages to increase precision to the max with the main exception of Search Pages, where the Comb together with 4C yield even better performance.

| | 3W Cosine | | 3W Comb | | 4C Comb | | BASELINE | |
|---|---|---|---|---|---|---|---|---|
| | P | R | P | R | P | R | P | R |
| Blog | 0.862 | 1 | 0.862 | 1 | 0.773 | 0.99 | 0.615 | 0.805 |
| Eshop | 0.888 | 0.835 | 0.88 | 0.805 | 0.97 | 0.81 | 0.769 | 0.6 |
| FAQs | 1 | 0.975 | 0.99 | 0.99 | 0.994 | 0.87 | 0.974 | 0.74 |
| Front Page | 0.99 | 0.975 | 0.98 | 0.99 | 0.901 | 1 | 0.972 | 0.52 |
| Listing | 0.919 | 0.685 | 0.951 | 0.58 | 0.868 | 0.625 | 0.436 | 0.68 |
| Per. Home P. | 0.91 | 0.76 | 0.885 | 0.805 | 0.976 | 0.615 | 0.964 | 0.405 |
| Search Page | 0.906 | 0.82 | 0.797 | 0.885 | 0.994 | 0.89 | 0.875 | 0.77 |
| | F1 = .782 | | F1 = .775 | | F1 = .765 | | F1 = .626 | |

**Table 8.** Precision-Recall table of *7Genres corpus*. The scores derived using the optimal parameter combinations for getting F1 maximisation. The baseline is for character 4-grams with parameters $\nu = 0.1$ and 50,000 feature size.

| | 3W MinMax | | 3W Comb | | 1W Cosine | | BASELINE | |
|---|---|---|---|---|---|---|---|---|
| | P | R | P | R | P | R | P | R |
| OTHER | 0.929 | 0.989 | 0.929 | 0.987 | 0.935 | 0.951 | 0.91 | 0.488 |
| Blog | 0.7 | 0.76 | 0.653 | 0.875 | 0.31 | 0.975 | 0.307 | 0.585 |
| Eshop | 0.957 | 0.11 | 0.75 | 0.105 | 0.959 | 0.355 | 0.063 | 0.495 |
| FAQs | 1 | 0.99 | 0.995 | 0.985 | 1 | 0.645 | 0.988 | 0.81 |
| Front Page | 0.862 | 0.97 | 0.879 | 0.985 | 0.949 | 0.94 | 1 | 0.79 |
| Listing | 1 | 0.015 | 0.5 | 0.02 | 0.846 | 0.055 | 0.033 | 0.38 |
| Per. Home P. | 1 | 0.035 | 1 | 0.015 | 0.556 | 0.125 | 0.092 | 0.4 |
| Search Page | 0.747 | 0.295 | 0.83 | 0.22 | 0.779 | 0.53 | 0.18 | 0.505 |
| DIY Guides | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.9 |
| Editorial | 1 | 1 | 1 | 1 | 0.739 | 0.85 | 1 | 0.75 |
| Features | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.9 |
| Short Bio | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0.85 |
| | F1 = .787 | | F1 = .768 | | F1 = .765 | | F1 = .643 | |

**Table 9.** Precision-Recall table of *SATNINIS corpus*. The scores derived using the optimal parameter combinations for getting the F1 maximisation. The baseline is for word 3-grams with parameters $\nu = 0.1$ and 10,000 feature size.

In the case of noise, in table 9 of SATNINIS corpus, as explained in detail previously (in section 4.4) RFSE is way more noise tolerant compare to OC-SVM ensemble. However, it the difficulty is not the same for all the genre of the corpus as expected. In particular, OC-SVM ensemble showing a dramatic difference in performance when we come on genre level, as seen in this table. That is, Front Pages, FAQs and the whole BBC genres it seems a very easy prediction task for the ensemble while for the rest of the genres precision drops as low as 0.063 (for Eshop).

Although, RFSE as we have seen in previous section is slightly affected by the noise there is great difference compare to the non-noise case for some genres. Particularly, Listing and Personal Home Pages suffering on recall mainly but also in precision depending on the combination of document representation and similarity measure. Moreover, for Blogs we have significantly lower performance in precision compare to the other genres. This is most likely ought to the great amount of web pages into the noise part of the SANTINIS corpus, i.e. Spirit1000, which they were multi-genre and one of their genres found to be Blog. This conclusion is after the qualitative analysis of Santini for the SANTINIS corpus in [20]. This is the reason why in confusion matrix of table 5 in section 4.4 we see so many pages of the OTHERs pile to be classified as blogs.

In respect to KI04 corpus, the genres of Article, Help and Portrait seems to be the most difficult cases for both RFSE and OC-SVM ensemble. However, there is at least one parameter combination for RFSE where a good precision can be achieve for these genre, while for OC-SVM ensemble performance is worst than random. Considering Link List and Discussion genres RFSE is suffering in recall maximisation where it achieves to identify only 38% to 63% of the samples of these genres.

Due to the small size of this corpus conclusion quite specific on the corpus. However, it seems that with a very small amount of samples and in a high imbalanced case like open-set approach cannot perform as good as on SANTINIS corpus, which is a noise-full corpus. Of course as we have seen in section 4.3 RFSE manages to yield precision more than 0.7 for at least 50% of the KI04 corpus, while OC-SVM ensemble shows one of it worst performances in respect of precision and recall (see figure 6).

| | 1W MinMax | | 1W Cosine | | 1W Comb | | BASELINE | |
|---|---|---|---|---|---|---|---|---|
| | P | R | P | R | P | R | P | R |
| Article | 0.576 | 0.866 | 0.553 | 0.866 | 0.489 | 0.874 | 0.278 | 0.354 |
| Discusion | 0.988 | 0.63 | 0.94 | 0.622 | 0.975 | 0.614 | 0.433 | 0.228 |
| Download | 0.847 | 0.808 | 0.963 | 0.695 | 0.924 | 0.722 | 0.393 | 0.57 |
| help | 0.441 | 0.669 | 0.688 | 0.554 | 0.576 | 0.576 | 0.5 | 0.101 |
| Link List | 0.951 | 0.38 | 0.904 | 0.502 | 0.946 | 0.429 | 0.343 | 0.473 |
| Portrait | 0.563 | 0.356 | 0.733 | 0.27 | 0.639 | 0.325 | 0.22 | 0.27 |
| Port. Priv. | 0.797 | 0.778 | 0.853 | 0.69 | 0.835 | 0.762 | 0.69 | 0.159 |
| Shop | 0.764 | 0.659 | 0.703 | 0.623 | 0.688 | 0.635 | 0.528 | 0.281 |
| | F1 = .612 | | F1 = .609 | | F1 = .605 | | F1 = .395 | |

**Table 10.** Precision-Recall table of *KI04 corpus*. The scores derived using the optimal parameter combinations for getting F1 maximisation. The baseline is for word unigrams with parameters $\nu = 0.1$ and 1,000 feature size.

### 4.8 Maximising Precision: WGI Filter Application

In to the previous section the performance behaviour of RFSE and OC-SVM ensemble has been analysed in tree different corpora. Firstly, the 7Genre, which is including the "common" genre pallet. Secondly, the KI04 which is the most widely used imbalanced corpus. Finally, the SANTINIS corpus which is introducing a more solid approach of what noise in web genres must be. In this section we are focusing on the case where our suggested open-set classification ensembles could be exploited in real world applications by tuning the parameters properly upon the application, based on the above analysis of the ensembles behaviour.

Several domains could exploit *an open-set web genre identification ensemble* such as *faceted search* [REF?], *genre aware crawling* [REF], *task retrieval* [REF?] and *automated translation*[REF?] to name a few. In all these cases precision maximisation is most likely the main requirement in a noise-full environment. Thus, in the next few lines we are showing how the precision can be maximised in a noise-full environment and the side effects of this tuning.

In order analyse the different behaviours of the ensembles upon precision maximisation we have tried F05, macro-precision and AUC maximisation. Although, it seem a straightforward problem it is not due to the variation in performance per genre as explained in section 4.7. Thus, in order to maximise precision for the greatest possible value and for the greatest amount of the corpus we had to select a different measure for maximisation. In the Appendix the results of our experiments are summarised for all measure we maximised, i.e. F1 (tables 13 and 17), F0.5 (tables 14 and 18), macro-Precision (tables 16 and 20) and AUC (tables 15 and 19).

In figure 10 and table 11 are depicted a set of parameters for both the ensembles in order to maximise the performance according to precision maximisation. Amongst all the maximisation methods we used as mentioned above RFSE seem to maximised its precision for the whole corpus based on PRC's AUC maximisation. OC-SVM ensemble is maximises its performance in respect to precision when macro-precision is maximised.

However, in both ensembles precision maximisation has some significant side effects. RFSE seem to recognise non of the Personal Home Pages while for Listing and Short Bio recall is dropping very low for maximising precision for the whole corpus. As for OC-SVM ensemble it seems that only about 5% of the pages has been identified and classified in the correct class while the rest has been rejected as OTHERs, moreover, Short Bio is not recognisable. However, even the weak OC-SVM ensemble manages to achieve precision close to 0.9 with the proper tuning, although it is not very useful in practise.

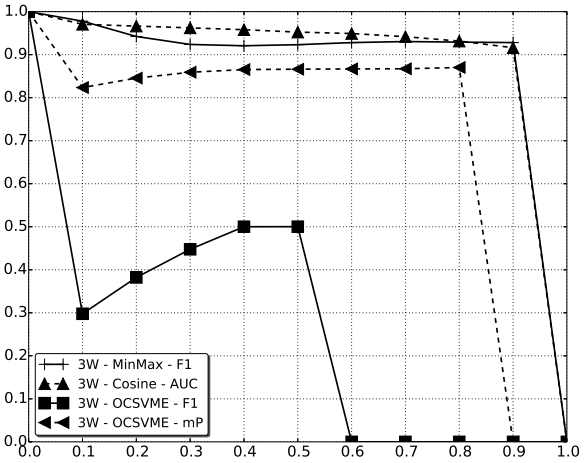


**Fig. 10.** Precision-Recall curves of RFSE and OC-SVM ensemble, for different maximisation criteria, i.e. F1, F0.5 and AUC, for locating the optimal parameters combination in respect to precision maximisation in the whole SANTINIS corpus.

| | 3W Cosine | | 3W MinMax | | 3W OCSVME mP | | 3W OCSVME F1 | |
|---|---|---|---|---|---|---|---|---|
| | P | R | P | R | P | R | P | R |
| OTHER | 0.917 | 0.989 | 0.929 | 0.989 | 0.876 | 0.991 | 0.91 | 0.488 |
| Blog | 0.607 | 0.81 | 0.7 | 0.76 | 1 | 0.06 | 0.307 | 0.585 |
| Eshop | 0.714 | 0.075 | 0.957 | 0.11 | 0.094 | 0.04 | 0.063 | 0.495 |
| FAQs | 1 | 0.655 | 1 | 0.99 | 1 | 0.075 | 0.988 | 0.81 |
| Front Page | 0.984 | 0.93 | 0.862 | 0.97 | 1 | 0.065 | 1 | 0.79 |
| Listing | 0.375 | 0.015 | 1 | 0.015 | 0.421 | 0.04 | 0.033 | 0.38 |
| Per. Home P. | NaN | 0 | 1 | 0.035 | 0.727 | 0.04 | 0.092 | 0.4 |
| Search Page | 0.905 | 0.095 | 0.747 | 0.295 | 0.8 | 0.06 | 0.18 | 0.505 |
| DIY Guides | 1 | 1 | 1 | 1 | 1 | 0.05 | 1 | 0.9 |
| Editorial | 1 | 1 | 1 | 1 | 1 | 0.05 | 1 | 0.75 |
| Features | 1 | 1 | 1 | 1 | 1 | 0.05 | 1 | 0.9 |
| Short Bio | 1 | 0.2 | 1 | 1 | NaN | 0 | 1 | 0.85 |
| | AUC = .905 | | F1=.787 | | mP = .746 | | F1 = .643 | |

**Table 11.** Precision-Recall table of *SATNINIS corpus*. The scores derived using the optimal parameter combinations for getting the AUC, F1 and macro-precision maximisation respectively.

An other aspect we have to note is the *certainty maximisation* which is implied when AUC is maximised, in particular case of the RFSE. As we have

seen up to know RFSE is seem to be very useful for application because not only is significantly more tolerant than OC-SVM ensemble but when AUC is maximised it is implied that the certainty of the classification decisions of the RFSE are maximised. In order to reason this, one should think of the procedure while calculating PRCs, where the web pages are ranked in a descending score order. In the open-set calcification approach the score returned by the ensemble is indicating the certainty of the decision, i.e. the farther for the threshold the more certain is the prediction genre tag is returned.

To conclude, in this section the behaviour of the ensemble has been analysed when precision maximisation is critical and it has been shown that the RFSE outperforms OC-SVM ensemble in a noise-full corpus such as SANTINIS. In addition, we have shown that our open-set classification best fits the web genre identification task (WGI a.k.a AGI) in real world applications.

## 5 Conclusion and Future Work

In this paper we focused on the impact of noise in AGI. It is the first time where a number of instances not belonging to any of the known genres were used in the evaluation phase to test the robustness of AGI models. Moreover, we examine appropriate classification models in an open-set scenario which is more realistic taking into account the lack of a complete genre palette and the constantly evolving web genres. Experimental results show that both RFSE and OC-SVM models are affected by noise but in general they remain robust. Moreover, only a few genres are heavily affected by noise. Finally, the choice of text similarity function and text representation features for RFSE models affect the effectiveness on certain genres. Thus, one can build a more reliable model if the appropriate combination of features and similarity measure is used for each genre separately.

| 7-Genres (and SANTINIS) | | | KI04 | | | BBC (and SANTINIS) | | |
|---|---|---|---|---|---|---|---|---|
| Label | Genre | Pages | Label | Genre | Pages | Label | Genre | Pages |
| G1 | Blogs | 200 | K1 | Articles | 127 | B1 | DIY Mini Guides | 20 |
| G2 | Eshops | 200 | K2 | Discussion | 127 | B2 | Editorial | 20 |
| G3 | FAQs | 200 | K3 | Download | 152 | B3 | Features | 20 |
| G4 | Front pages | 200 | K4 | Helps | 140 | B4 | Short Bio | 20 |
| G5 | Listings | 200 | K5 | Link List | 208 | | | |
| G6 | Personal Home Pages | 200 | K6 | Portrait | 179 | | | |
| G7 | Search Pages | 200 | K7 | Portrait Private | 131 | | | |
| | | | K8 | Shop | 175 | | | |
| OTR: | Other pages, a.k.a "Don't Know" or Spirit1000 Noise, i.e. Expected "Don't Know" | | | | | | | 1000 |

**Table 12.** Genre tags, descriptions, and amount of pages for all corpora.

# A Appendix

| Dist. Meas. | Corpus | Doc. Rep. | Vocab. | Feat. | $\sigma$ | Iter. | Pre | Rec | F1 | AUC |
|---|---|---|---|---|---|---|---|---|---|---|
| Cosine | 7Genres | 3W | 100,000 | 50,000 | 0.5 | 50 | 0.809 | 0.756 | 0.782 | 0.835 |
| | | 1W | 100,000 | 50,000 | 0.5 | 100 | 0.776 | 0.696 | 0.733 | 0.733 |
| | | 4C | 50,000 | 5,000 | 0.5 | 50 | 0.79 | 0.723 | 0.755 | 0.829 |
| | KI04 | 3W | 100,000 | 90,000 | 0.5 | 50 | 0.664 | 0.518 | 0.582 | 0.497 |
| | | 1W | 100,000 | 50,000 | 0.5 | 50 | 0.704 | 0.536 | 0.609 | 0.5 |
| | | 4C | 10,000 | 1,000 | 0.5 | 10 | 0.69 | 0.468 | 0.558 | 0.473 |
| | SANTINIS | 3W | 50,000 | 10,000 | 0.7 | 100 | 0.797 | 0.722 | 0.758 | 0.768 |
| | | 1W | 50,000 | 5,000 | 0.5 | 100 | 0.839 | 0.702 | 0.765 | 0.77 |
| | | 4C | 50,000 | 1,000 | 0.5 | 100 | 0.739 | 0.78 | 0.759 | 0.606 |
| MinMax | 7Genres | 3W | 100,000 | 90,000 | 0.5 | 10 | 0.777 | 0.764 | 0.77 | 0.819 |
| | | 1W | 5,000 | 1,000 | 0.5 | 10 | 0.761 | 0.664 | 0.709 | 0.711 |
| | | 4C | 10,000 | 5,000 | 0.5 | 100 | 0.768 | 0.74 | 0.754 | 0.807 |
| | KI04 | 3W | 100,000 | 90,000 | 0.5 | 50 | 0.663 | 0.529 | 0.589 | 0.487 |
| | | 1W | 10,000 | 5,000 | 0.5 | 100 | 0.659 | 0.572 | 0.612 | 0.572 |
| | | 4C | 10,000 | 1,000 | 0.5 | 100 | 0.711 | 0.512 | 0.595 | 0.48 |
| | SANTINIS | 3W | 50,000 | 5,000 | 0.7 | 100 | 0.933 | 0.68 | 0.787 | 0.893 |
| | | 1W | 100,000 | 10,000 | 0.7 | 100 | 0.826 | 0.69 | 0.752 | 0.866 |
| | | 4C | 100,000 | 5,000 | 0.9 | 100 | 0.864 | 0.682 | 0.762 | 0.862 |
| Cos & MinMax | 7Genres | 3W | 100,000 | 50,000 | 0.5 | 100 | 0.793 | 0.757 | 0.775 | 0.835 |
| | | 1W | 10,000 | 5,000 | 0.5 | 100 | 0.769 | 0.68 | 0.722 | 0.73 |
| | | 4C | 10,000 | 1,000 | 0.5 | 50 | 0.81 | 0.725 | 0.765 | 0.831 |
| | KI04 | 3W | 100,000 | 90,000 | 0.5 | 50 | 0.643 | 0.528 | 0.58 | 0.489 |
| | | 1W | 10,000 | 5,000 | 0.5 | 100 | 0.675 | 0.549 | 0.605 | 0.486 |
| | | 4C | 10,000 | 5,000 | 0.5 | 100 | 0.602 | 0.478 | 0.533 | 0.438 |
| | SANTINIS | 3W | 100,000 | 1,000 | 0.5 | 100 | 0.878 | 0.683 | 0.768 | 0.903 |
| | | 1W | 50,000 | 5,000 | 0.5 | 100 | 0.854 | 0.689 | 0.763 | 0.862 |
| | | 4C | 50,000 | 1,000 | 0.5 | 100 | 0.752 | 0.773 | 0.762 | 0.725 |

**Table 13.** Optimal combinations of parameters for RFSE and the respective F1 statistic, after Precision/Recall macro-averaging, for each combination. AUC has been calculated under the full PR curves, i.e. PR curve has one point for every web page of the corpus. SANTINIS copus curves has been calculated using strata sampling over the RFSE results of the noise part of the corpus.

| Dist. Meas. | Corpus | Doc. Rep. | Vocab. | Feat. | $\sigma$ | Iter. | Pre | Rec | F0.5 | AUC |
|---|---|---|---|---|---|---|---|---|---|---|
| Cosine | 7Genres | 3W | 100,000 | 50,000 | 0.7 | 50 | 0.843 | 0.691 | 0.808 | 0.741 |
| | | 1W | 10,000 | 1,000 | 0.5 | 100 | 0.849 | 0.599 | 0.784 | 0.643 |
| | | 4C | 50,000 | 1,000 | 0.5 | 50 | 0.844 | 0.623 | 0.788 | 0.739 |
| | KI04 | 3W | 100,000 | 50,000 | 0.5 | 100 | 0.707 | 0.464 | 0.64 | 0.49 |
| | | 1W | 50,000 | 10,000 | 0.5 | 100 | 0.802 | 0.458 | 0.697 | 0.51 |
| | | 4C | 100,000 | 10,000 | 0.5 | 50 | 0.777 | 0.432 | 0.67 | 0.423 |
| | SANTINIS | 3W | 100,000 | 5,000 | 0.9 | 50 | 0.885 | 0.603 | 0.809 | 0.887 |
| | | 1W | 100,000 | 10,000 | 0.7 | 100 | 0.897 | 0.606 | 0.819 | 0.869 |
| | | 4C | 50,000 | 1,000 | 0.7 | 100 | 0.871 | 0.658 | 0.818 | 0.861 |
| MinMax | 7Genres | 3W | 100,000 | 90,000 | 0.9 | 100 | 0.84 | 0.678 | 0.802 | 0.735 |
| | | 1W | 10,000 | 1,000 | 0.5 | 100 | 0.809 | 0.595 | 0.755 | 0.636 |
| | | 4C | 10,000 | 1,000 | 0.7 | 50 | 0.827 | 0.644 | 0.782 | 0.731 |
| | KI04 | 3W | 100,000 | 50,000 | 0.5 | 100 | 0.702 | 0.466 | 0.638 | 0.481 |
| | | 1W | 5,000 | 1,000 | 0.5 | 100 | 0.73 | 0.517 | 0.674 | 0.507 |
| | | 4C | 10,000 | 1,000 | 0.5 | 100 | 0.711 | 0.512 | 0.66 | 0.48 |
| | SANTINIS | 3W | 50,000 | 5,000 | 0.7 | 100 | 0.933 | 0.68 | 0.868 | 0.893 |
| | | 1W | 50,000 | 5,000 | 0.7 | 100 | 0.929 | 0.601 | 0.838 | 0.872 |
| | | 4C | 100,000 | 5,000 | 0.9 | 100 | 0.864 | 0.682 | 0.82 | 0.862 |
| Cos & MinMax | 7Genres | 3W | 100,000 | 50,000 | 0.7 | 50 | 0.839 | 0.693 | 0.805 | 0.741 |
| | | 1W | 5,000 | 1,000 | 0.5 | 100 | 0.821 | 0.628 | 0.774 | 0.733 |
| | | 4C | 100,000 | 5,000 | 0.5 | 100 | 0.824 | 0.691 | 0.793 | 0.737 |
| | KI04 | 3W | 100,000 | 50,000 | 0.5 | 100 | 0.68 | 0.481 | 0.628 | 0.488 |
| | | 1W | 5,000 | 1,000 | 0.5 | 50 | 0.75 | 0.47 | 0.67 | 0.489 |
| | | 4C | 100,000 | 10,000 | 0.5 | 50 | 0.75 | 0.408 | 0.642 | 0.409 |
| | SANTINIS | 3W | 50,000 | 10,000 | 0.9 | 50 | 0.915 | 0.635 | 0.841 | 0.882 |
| | | 1W | 50,000 | 5,000 | 0.5 | 100 | 0.854 | 0.689 | 0.815 | 0.862 |
| | | 4C | 50,000 | 1,000 | 0.7 | 100 | 0.877 | 0.659 | 0.822 | 0.866 |

**Table 14.** Optimal combinations of parameters for RFSE and the respective F0.5 statistic, after Precision/Recall macro-averaging, for each combination. AUC has been calculated under the full PR curves, i.e. PR curve has one point for every web page of the corpus. SANTINIS copus curves has been calculated using strata sampling over the RFSE results of the noise part of the corpus.

| Dist. Meas. | Corpus | Doc. Rep. | Vocab. | Feat. | $\sigma$ | Iter. | Pre | Rec | AUC | F1 |
|---|---|---|---|---|---|---|---|---|---|---|
| Cosine | 7Genres | 3W | 100,000 | 50,000 | 0.5 | 100 | 0.811 | 0.751 | 0.836 | 0.78 |
| | | 1W | 10,000 | 5,000 | 0.5 | 10 | 0.739 | 0.701 | 0.809 | 0.72 |
| | | 4C | 10,000 | 1,000 | 0.5 | 100 | 0.793 | 0.708 | 0.829 | 0.748 |
| | KI04 | 3W | 100,000 | 90,000 | 0.5 | 50 | 0.664 | 0.518 | 0.497 | 0.582 |
| | | 1W | 100,000 | 50,000 | 0.5 | 10 | 0.666 | 0.554 | 0.557 | 0.605 |
| | | 4C | 50,000 | 5,000 | 0.5 | 10 | 0.672 | 0.463 | 0.477 | 0.549 |
| | SANTINIS | 3W | 100,000 | 1,000 | 0.5 | 50 | 0.792 | 0.564 | 0.905 | 0.659 |
| | | 1W | 50,000 | 5,000 | 0.7 | 100 | 0.934 | 0.536 | 0.873 | 0.681 |
| | | 4C | 50,000 | 1,000 | 0.9 | 100 | 0.799 | 0.47 | 0.874 | 0.592 |
| MinMax | 7Genres | 3W | 100,000 | 90,000 | 0.7 | 100 | 0.803 | 0.732 | 0.831 | 0.766 |
| | | 1W | 5,000 | 1,000 | 0.5 | 100 | 0.784 | 0.636 | 0.723 | 0.703 |
| | | 4C | 10,000 | 1,000 | 0.5 | 100 | 0.791 | 0.714 | 0.823 | 0.75 |
| | KI04 | 3W | 100,000 | 90,000 | 0.5 | 100 | 0.659 | 0.51 | 0.487 | 0.575 |
| | | 1W | 10,000 | 5,000 | 0.5 | 100 | 0.659 | 0.572 | 0.572 | 0.612 |
| | | 4C | 10,000 | 1,000 | 0.5 | 100 | 0.711 | 0.512 | 0.48 | 0.595 |
| | SANTINIS | 3W | 50,000 | 1,000 | 0.5 | 100 | 0.862 | 0.541 | 0.899 | 0.665 |
| | | 1W | 50,000 | 5,000 | 0.7 | 100 | 0.929 | 0.601 | 0.872 | 0.73 |
| | | 4C | 100,000 | 1,000 | 0.7 | 100 | 0.818 | 0.598 | 0.873 | 0.691 |
| Cos & MinMax | 7Genres | 3W | 100,000 | 50,000 | 0.5 | 100 | 0.793 | 0.757 | 0.835 | 0.775 |
| | | 1W | 5,000 | 1,000 | 0.5 | 100 | 0.821 | 0.628 | 0.733 | 0.712 |
| | | 4C | 10,000 | 1,000 | 0.5 | 100 | 0.802 | 0.723 | 0.832 | 0.76 |
| | KI04 | 3W | 100,000 | 50,000 | 0.5 | 50 | 0.674 | 0.475 | 0.49 | 0.557 |
| | | 1W | 100,000 | 90,000 | 0.5 | 100 | 0.658 | 0.554 | 0.546 | 0.601 |
| | | 4C | 100,000 | 50,000 | 0.5 | 100 | 0.616 | 0.468 | 0.442 | 0.532 |
| | SANTINIS | 3W | 100,000 | 1,000 | 0.5 | 100 | 0.878 | 0.683 | 0.903 | 0.768 |
| | | 1W | 50,000 | 5,000 | 0.7 | 100 | 0.878 | 0.544 | 0.875 | 0.672 |
| | | 4C | 50,000 | 1,000 | 0.9 | 100 | 0.882 | 0.495 | 0.875 | 0.634 |

**Table 15.** Optimal combinations of parameters for RFSE and the respective Precision-Recall AUC for each combination. AUC has been calculated under the full PR curves, i.e. PR curve has one point for every web page of the corpus. SANTINIS copus curves has been calculated using strata sampling over the RFSE results of the noise part of the corpus.

| Dist. Meas. | Corpus | Doc. Rep. | Vocab. | Feat. | $\sigma$ | Iter. | Pre | Rec | AUC |
|---|---|---|---|---|---|---|---|---|---|
| Cosine | 7Genres | 3W | 100,000 | 10,000 | 0.9 | 100 | 0.874 | 0.341 | 0.346 |
| | | 1W | 50,000 | 5,000 | 0.9 | 10 | 0.875 | 0.174 | 0.15 |
| | | 4C | 100,000 | 5,000 | 0.9 | 50 | 0.873 | 0.374 | 0.449 |
| | KI04 | 3W | 100,000 | 10,000 | 0.9 | 10 | 0.849 | 0.126 | 0.129 |
| | | 1W | 100,000 | 50,000 | 0.9 | 100 | 0.872 | 0.179 | 0.148 |
| | | 4C | 100,000 | 1,000 | 0.5 | 100 | 0.885 | 0.061 | 0.05 |
| | SANTINIS | 3W | 50,000 | 5,000 | 0.9 | 50 | 0.889 | 0.593 | 0.886 |
| | | 1W | 50,000 | 1,000 | 0.5 | 50 | 0.954 | 0.415 | 0.771 |
| | | 4C | 100,000 | 5,000 | 0.9 | 100 | 0.892 | 0.585 | 0.871 |
| MinMax | 7Genres | 3W | 10,000 | 1,000 | 0.9 | 10 | 0.873 | 0.353 | 0.416 |
| | | 1W | 100,000 | 10,000 | 0.9 | 50 | 0.872 | 0.329 | 0.348 |
| | | 4C | 100,000 | 1,000 | 0.9 | 10 | 0.872 | 0.306 | 0.348 |
| | KI04 | 3W | 50,000 | 5,000 | 0.9 | 100 | 0.855 | 0.112 | 0.136 |
| | | 1W | 10,000 | 1,000 | 0.9 | 10 | 0.859 | 0.132 | 0.148 |
| | | 4C | 10,000 | 1,000 | 0.9 | 100 | 0.87 | 0.137 | 0.143 |
| | SANTINIS | 3W | 50,000 | 5,000 | 0.9 | 10 | 0.952 | 0.611 | 0.889 |
| | | 1W | 50,000 | 1,000 | 0.5 | 50 | 0.945 | 0.4 | 0.771 |
| | | 4C | 100,000 | 5,000 | 0.9 | 100 | 0.864 | 0.682 | 0.862 |
| Cos & MinMax | 7Genres | 3W | 50,000 | 5,000 | 0.9 | 10 | 0.874 | 0.286 | 0.347 |
| | | 1W | 5,000 | 1,000 | 0.9 | 50 | 0.873 | 0.241 | 0.249 |
| | | 4C | 100,000 | 1,000 | 0.7 | 100 | 0.872 | 0.344 | 0.349 |
| | KI04 | 3W | 100,000 | 10,000 | 0.9 | 50 | 0.866 | 0.087 | 0.05 |
| | | 1W | 100,000 | 5,000 | 0.7 | 10 | 0.872 | 0.066 | 0.05 |
| | | 4C | 50,000 | 5,000 | 0.9 | 100 | 0.884 | 0.054 | 0.05 |
| | SANTINIS | 3W | 100,000 | 1,000 | 0.7 | 10 | 0.951 | 0.492 | 0.892 |
| | | 1W | 5,000 | 1,000 | 0.9 | 100 | 0.898 | 0.481 | 0.777 |
| | | 4C | 50,000 | 1,000 | 0.9 | 10 | 0.903 | 0.412 | 0.778 |

**Table 16.** Optimal combinations of parameters for RFSE, after Precision/Recall macro-averaging, for each combination. AUC has been calculated under the full PR curves, i.e. PR curve has one point for every web page of the corpus. SANTINIS copus curves has been calculated using strata sampling over the RFSE results of the noise part of the corpus.

| Corpus | Doc. Rep. | Vocab. | Feat. | $\nu$ | Pre | Rec | F1 | AUC |
|---|---|---|---|---|---|---|---|---|
| 7Genres | 3W | 10,000 | 5,000 | 0.1 | 0.711 | 0.521 | 0.601 | 0.525 |
| | 1W | 50,000 | 5,000 | 0.07 | 0.658 | 0.521 | 0.582 | 0.432 |
| | 4C | 100,000 | 50,000 | 0.1 | 0.701 | 0.565 | 0.626 | 0.53 |
| KI04 | 3W | 100,000 | 5,000 | 0.3 | 0.583 | 0.242 | 0.342 | 0.177 |
| | 1W | 10,000 | 1,000 | 0.1 | 0.551 | 0.308 | 0.395 | 0.176 |
| | 4C | 100,000 | 90,000 | 0.07 | 0.502 | 0.316 | 0.388 | 0.185 |
| SANTINIS | 3W | 50,000 | 10,000 | 0.1 | 0.631 | 0.654 | 0.643 | 0.3 |
| | 1W | 100,000 | 50,000 | 0.3 | 0.259 | 0.283 | 0.27 | 0.094 |
| | 4C | 10,000 | 5,000 | 0.5 | 0.662 | 0.367 | 0.472 | 0.334 |

**Table 17.** Optimal combinations of parameters for OC-SVM ensemble and the respective F1 statistic, after Precision/Recall macro-averaging, for each combination. AUC has been calculated under the full PR curves, i.e. PR curve has one point for every web page of the corpus.

| Corpus | Doc. Rep. | Vocab. | Feat. | $\nu$ | Pre | Rec | F05 | AUC |
|---|---|---|---|---|---|---|---|---|
| 7Genres | 3W | 50,000 | 10,000 | 0.1 | 0.75 | 0.496 | 0.68 | 0.531 |
| | 1W | 100,000 | 90,000 | 0.1 | 0.664 | 0.513 | 0.627 | 0.444 |
| | 4C | 100,000 | 90,000 | 0.1 | 0.703 | 0.563 | 0.67 | 0.529 |
| KI04 | 3W | 100,000 | 10,000 | 0.3 | 0.617 | 0.228 | 0.46 | 0.187 |
| | 1W | 10,000 | 5,000 | 0.17 | 0.58 | 0.295 | 0.486 | 0.179 |
| | 4C | 100,000 | 50,000 | 0.07 | 0.507 | 0.313 | 0.451 | 0.184 |
| SANTINIS | 3W | 100,000 | 50,000 | 0.07 | 0.647 | 0.603 | 0.638 | 0.584 |
| | 1W | 50,000 | 10,000 | 0.9 | 0.592 | 0.137 | 0.356 | 0.609 |
| | 4C | 100,000 | 50,000 | 0.5 | 0.687 | 0.348 | 0.575 | 0.333 |

**Table 18.** Optimal combinations of parameters for OC-SVM ensemble and the respective F0.5 statistic, after Precision/Recall macro-averaging, for each combination. AUC has been calculated under the full PR curves, i.e. PR curve has one point for every web page of the corpus.

| Corpus | Doc. Rep. | Vocab. | Feat. | $\nu$ | Pre | Rec | AUC | F1 |
|---|---|---|---|---|---|---|---|---|
| 7Genres | 3W | 50,000 | 10,000 | 0.15 | 0.752 | 0.483 | 0.532 | 0.588 |
| | 1W | 5,000 | 1,000 | 0.07 | 0.644 | 0.527 | 0.494 | 0.58 |
| | 4C | 50,000 | 10,000 | 0.1 | 0.687 | 0.571 | 0.53 | 0.623 |
| KI04 | 3W | 100,000 | 10,000 | 0.5 | 0.638 | 0.191 | 0.188 | 0.294 |
| | 1W | 10,000 | 1,000 | 0.07 | 0.496 | 0.32 | 0.181 | 0.389 |
| | 4C | 50,000 | 10,000 | 0.3 | 0.521 | 0.287 | 0.191 | 0.37 |
| SANTINIS | 3W | 100,000 | 90,000 | 0.9 | 0.743 | 0.127 | 0.746 | 0.217 |
| | 1W | 10,000 | 5,000 | 0.9 | 0.586 | 0.136 | 0.609 | 0.221 |
| | 4C | 10,000 | 5,000 | 0.9 | 0.615 | 0.153 | 0.717 | 0.246 |

**Table 19.** Optimal combinations of parameters for OC-SVM ensemble and the respective AUC, for each combination. AUC has been calculated under the full PR curves, i.e. PR curve has one point for every web page of the corpus.

| Corpus | Doc. Rep. | 0 | Feat. | $\nu$ | Pre | Rec | AUC |
|---|---|---|---|---|---|---|---|
| 7Genres | 3W | 100,000 | 90,000 | 0.17 | 0.829 | 0.334 | 0.348 |
| | 1W | 50,000 | 5,000 | 0.17 | 0.67 | 0.481 | 0.443 |
| | 4C | 10,000 | 5,000 | 0.3 | 0.763 | 0.426 | 0.323 |
| KI04 | 3W | 100,000 | 90,000 | 0.9 | 0.721 | 0.032 | 0.05 |
| | 1W | 50,000 | 5,000 | 0.9 | 0.625 | 0.061 | 0.05 |
| | 4C | 50,000 | 5,000 | 0.7 | 0.526 | 0.143 | 0.096 |
| SANTINIS | 3W | 100,000 | 90,000 | 0.9 | 0.743 | 0.127 | 0.746 |
| | 1W | 100,000 | 50,000 | 0.9 | 0.595 | 0.136 | 0.609 |
| | 4C | 100,000 | 50,000 | 0.7 | 0.732 | 0.238 | 0.468 |

**Table 20.** Optimal combinations of parameters for OC-SVM ensemble, after Precision/Recall macro-averaging, for each combination. AUC has been calculated under the full PR curves, i.e. PR curve has one point for every web page of the corpus.

## References


1. Anderka, M., S.B.L.N.: Detection of text quality as as a one-class classification problem. 20th ACM International Conference on Information and Knowledge Management (CIKM'11) pp. 2313–2316 (2011)
2. Asheghi, N.R., Markert, K., Sharoff, S.: Semi-supervised graph-based genre classification for web pages. TextGraphs-9 p. 39 (2014)
3. Bishop, C.: Pattern Recognition and Machine Learning pp. 331–336 (2006)
4. 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)
5. Coutinho, M.A., Miranda, F.: 3 to describe genres: Problems and strategies. PERSPECTIVES ON WRITING p. 35

6. Crowston, K., Kwaśnik, B., Rubleske, J.: Problems in the use-centered development of a taxonomy of web genres. In: Genres on the Web, pp. 69–84. Springer (2011)
7. De Assis, G.T., Laender, A.H., Gonçalves, M.A., Da Silva, A.S.: A genre-aware approach to focused crawling. World Wide Web 12(3), 285–319 (2009)
8. Dong, L., Watters, C., Duffy, J., Shepherd, M.: Binary cybergenre classification using theoretic feature measures (2006)
9. Meyer zu Eissen, S., Stein, B.: Genre classification of web pages. KI 2004: Advances in Artificial Intelligence pp. 256–269 (2004)
10. 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 NACACL, Companion Volume: Short Papers. pp. 173–176. Association for Computational Linguistics (2009)
11. Joho, H., Sanderson, M.: The spirit collection: an overview of a large web collection. In: ACM SIGIR Forum. vol. 38, pp. 57–61. ACM (2004)
12. Kanaris, I., Stamatatos, E.: Learning to recognize webpage genres. Information Processing & Management 45(5), 499–512 (2009)
13. Kennedy, A., Shepherd, M.: Automatic identification of home pages on the web. In: System Sciences, 2005. HICSS'05. Proceedings of the 38th Annual Hawaii International Conference on. pp. 99c–99c. IEEE (2005)
14. Khan, S., Madden, M.: A survey of recent trends in one class classification. Artificial Intelligence and Cognitive Science pp. 188–197 (2010)
15. Koppel, M., Schler, J., Argamon, S.: Authorship attribution in the wild. Language Resources and Evaluation 45(1), 83–94 (2011)
16. Koppel, M., Winter, Y.: Determining if two documents are written by the same author. Journal of the Association for Information Science and Technology 65(1), 178–187 (2014)
17. 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)
18. Manevitz, L., Yousef, M.: One-class svms for document classification. The Journal of Machine Learning Research 2, 139–154 (2002)
19. 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)
20. Mehler, A., Sharoff, S., Santini, M.: Genres on the Web: Computational Models and Empirical Studies. Text, Speech and Language Technology, Springer (2010)
21. Pritsos, D.A., Stamatatos, E.: Open-set classification for automated genre identification. In: Advances in Information Retrieval, pp. 207–217. Springer (2013)
22. Roussinov, D., Crowston, K., Nilan, M., Kwasnik, B., Cai, J., Liu, X.: Genre based navigation on the web. In: System Sciences, 2001. Proceedings of the 34th Annual Hawaii International Conference on. pp. 10–pp. IEEE (2001)
23. Santini, M.: Automatic identification of genre in web pages. Ph.D. thesis, University of Brighton (2007)
24. Santini, M., Sharoff, S.: Web genre benchmark under construction. Journal for Language Technology and Computational Linguistics 24(1), 129–145 (2009)
25. Santini, M.: Cross-testing a genre classification model for the web. In: Genres on the Web, pp. 87–128. Springer (2011)
26. 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)

27. 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)
28. Shepherd, M.A., Watters, C.R., Kennedy, A.: Cybergenre: Automatic identification of home pages on the web. J. Web Eng. 3(3-4), 236–251 (2004)
29. Stamatatos, E.: A survey of modern authorship attribution methods. Journal of the American Society for information Science and Technology 60(3), 538–556 (2009)
30. Sugiyanto, S., Rozi, N.F., Putri, T.E., Arifin, A.Z.: Term weighting based on index of genre for web page genre classification. JUTI: Jurnal Ilmiah Teknologi Informasi 12(1), 27–34 (2014)
31. Zhu, J., Zhou, X., Fung, G.: Enhance web pages genre identification using neighboring pages. In: Web Information System Engineering–WISE 2011, pp. 282–289. Springer (2011)