Common Criteria for Genre Classification: Annotation and Granularity

Marina Santini[1]

ABSTRACT

In this paper, we present two experiments that use machine learning for automatically classifying web pages by genre. These experiments highlight the influence that genre annotation and genre granularity can have on the accuracy of the classification. From a practical point of view these experiments show that a collection annotated with the criteria of 'objective sources' and consistent genre granularity ensures a very good classification accuracy (Experiment 1). Additionally, the classification model built out of such a collection can be exported more profitably for predictive tasks on an unclassified web page collection (Experiment 2). These experiments represent a starting point for a discussion about the need of common criteria for building a genre collection in the absence of an official genre-annotated benchmark.

1 INTRODUCTION

In this paper, we present two experiments that use machine learning for automatically classifying web pages by genre.

Many definitions of genre have been proposed so far in literary studies (e.g. [20]), academic writing (e.g. [23]), professional settings (e.g. [2] and [24]), organizational environment (e.g. [26]), and so on. More specifically, in automatic genre classification studies, genres have often been seen as non-topical categories that could help reduce information overload (e.g. [16] or [15]). In this area, not only text categories such as 'article', 'FAQs', 'home page', etc. have been considered to be genres, but also polarities, such as subjective-objective and positive-negative ([7]), and style ([1], [9] and [5]). Regardless the different definitions and connotations, a classification by genre has been acknowledged to be useful in information retrieval (e.g. [9], [12], etc.), information filtering ([7]), digital libraries ([19]) and other practical applications.

In this paper we present two experiments of genre classification of web pages based on a simplified and intuitive definition of genre, which is suitable for all kind of genres – including genres on the web – and for an automatic approach. In our view, genres can be defined as named *socio-cultural* communication artefacts, linked to a society or a community, bearing standardized traits, leaving space for the creativity of the text producer, and raising expectations in the text receiver. For example, the personal home page (cf. also [6]) has standard traits, such as self-narration, personal interests, contact details, and often pictures related to one's life. However, these conventions do not hinder the creativity of the producer, and as receivers, we expect a blend of standardized information and personal touch. Though unsophisticated, this definition of genre allows us to suggest a practical solution to the main shortcoming in genre classification, i.e. the lack of a genre-annotated benchmark. Because of this lack, the main tendency has always been to build one's own collection according to subjective criteria as for genre annotation and genre granularity. This is especially true for genre studies based on collections of web pages. Although building a genre-annotated benchmark of web pages is difficult and maybe not feasible, because annotating a web page by genre is both hard and controversial (cf. [21]), a few criteria should be discussed and agreed upon. Without some kind of commonality, any comparison becomes unfeasible. For instance, can we state that the 92% accuracy achieved by [3] is better than the accuracy (about 70%) achieved by [17]? The solution we suggest for building more comparable genre collections is to exploit the *socio-cultural* aspect of the concept of genre. As pointed out earlier, genres have a function in a society, culture or community, i.e. they have a social or public role that implies a number of conventions and raises predictable expectations. This means that the role or the function of different genres is recognized and correctly used in the communication interaction. Leveraging on this public and collective acknowledgement it is possible to create a genre-annotated collection without involving human annotators. The key is to download documents from genre-specific archives or portals and use their membership in these containers as an automatic membership in a specific genre. For example, eshops can be randomly downloaded from the portal http://www.eshops.co.uk/ and considered to be eshops without any further manual annotation or inter-rater agreement assessment. We include in the public acknowledgement also genres used as title of documents (for example, "Insects Hotlist"). The idea behind selecting documents with a genre in the title or picking them up randomly from public resources, such as an archives or a portals, is the following: if there is an archive, a portal or a website specialized in, say, pointing to or collecting genres such as eshops, blogs or search engines, this means that the documents pointed to or collected there are considered to belong to these genres by the collectivity of web users. We call this criterion 'annotation by objective sources'. A genre collection annotated by objective sources tends to be more representative as for intra-genre variation than a collection annotated relying on the genre stereotypicality that two, three, or more annotators have in mind. We suggest that annotating a collection using objective sources is faster and closer to real-world conditions.

Genre granularity is also important when building a collection for genre classification. In fact, genre palettes often show different levels of granularity. For instance, [9] includes in his genre palette both FAQs (genre) and journalistic materials (super-genre). We suggest the use of the prototype theory (cf. [18] and [13]) to achieve a consistent level of genre granularity. A prototype is the most typical instance of a more encompassing or fuzzy category. Categories that can be dealt with the prototype theory can be ordered into a three-tiered hierarchy: superordinate level, basic level and subordinate level. For example, the genre 'advertisement' represents the basic level (genre) of the superordinate level 'advertising' (super-genre), while a 'web ad' represents the subordinate level (subgenre) of the basic level. The

[1] University of Brighton (UK); M.Santini@brighton.ac.uk

basic level embodies the information level at which concepts are most easily recognized, remembered and learned with respect to their function. The basic level included in the prototype theory should not be mixed up with document stereotypicality or exemplarity. Building a genre collection choosing exemplars, i.e. only stereotypical documents, to unambiguously represent a genre can return biased results. According to the prototype theory, instead, instances of a genre may vary in their prototypicality, thus allowing intra-genre variation.

The two experiments presented in this paper highlight the influence that genre annotation and genre granularity can have on the accuracy of genre classification of web pages. They were designed to point out several issues (some already covered in [22]). In this paper, these two experiments allow us to emphasize two general aspects of genre classification, one practical and one theoretical. From a practical point of view these experiments show that a collection annotated with the criteria of objective sources and consistent genre granularity ensures a very good classification accuracy (Experiment 1). Additionally, the classification model built out of such a collection can be exported more profitably for predictive tasks on an unclassified web page collection (Experiment 2). From a theoretical point of view, they represent a starting point for a discussion about the need of common criteria in the absence of an official genre-annotated benchmark

In order to ensure replicability, all the materials used for these experiments, including web page collections, feature sets and the manual evaluation of Experiment 2, are available at http://www.nltg.brighton.ac.uk/home/Marina.Santini/, bottom of the page.

The paper is organized as follows: Section 2 provides an overview of recent work in genre classification of web pages; Section 3 presents the web page collections and the two experiments; conclusions are drawn in Section 4.

2 PREVIOUS WORK

Several experiments have been recently carried out with genres and web pages. Here we list the latest studies in order to show how difficult is to compare their results in the absence of common criteria as for corpus building and genre palettes.

[7]: *Number of web pages: 2150; Annotation: single rater; Categories: subjectivity, positive-ness*. They tried to discriminate among texts coming from different domains in terms of two polarities: subjective vs. objective and positive vs. negative. Their aim was to see how a classification model tuned on one domain performed in another domain. According to their results, in single domain classification the best accuracy is achieved with Multi-View-Ensemble (MVE) (see [7] for details) for subjectivity, and with bag-of-words (BOW) features for positive-ness. In domain transfer classification, the best accuracy is achieved with Parts-of-Speech (POS) tags for subjectivity and MVE for positive-ness. Although it is true that genres can be divided into more subjective genres (e.g. editorials), or more objective genres (e.g. surveys), and that the opposition positive-negative can suggest specific genres (such as reviews), these two polarities can hardly be considered as “genres” in themselves. Nonetheless, [7]’s contribution is extremely valuable because they shed some light on the performance of different feature sets across several domains, providing insight into the extent of feature exportability.

[5]: *Number of web pages: 2700; Annotation: one or more raters; Categories: functional styles*. They carried out an experiment on style-dependent document ranking. Their research explored the possibility of incorporating style-dependent ranking into ranking schemata for searching the web and digital libraries. Their basic idea was to reduce styles (more specifically, the five functional styles theorized by the School of Prague) to a single continuous parameter. Regardless the promising preliminary results, they could see little improvement in relevance ranking when stylistic parameters were included.

[3]: *Number of web pages: 343; Genre annotation: the author plus at least one or more raters; Genres: abstract, call for papers, FAQs, hub/sitemap, job description, resume/C.V., statistics, syllabus, technical paper*. She tried out the efficiency of several feature sets and automatic feature selection techniques on a small corpus of 10 genres, using a number of classification algorithms. Although her results can be considered only indicative given the reduced number of pages per genre (an average of 20 web pages per genre class), she made interesting remarks about discrimination across similar genres, and the influence of the genre palette and document exemplarity on discrimination tasks. Her best accuracy (92.1%) was achieved by one of the feature combinations resulting from an automatic feature selection technique.

[10]: *Number of web pages: 321; Genre annotation: do not say; Genres: personal, corporate, organizational home pages, including also non-home pages, as noise*. They tried the hard task of home page genre discrimination. The best accuracy (71.4%) is achieved on personal home pages with a single classifier, manual feature selection, and without noisy pages.

[16]: *Number of web pages: 1224; Genre annotation: two graduate students; Genres: personal home page, public home page, commercial home page, bulletin collection, link collection, image collection, simple table/lists, input pages, journalistic material, research report, official materials, FAQs, discussions, product specification, informal texts (poem, fiction, etc.)*. They investigated the efficiency of several feature sets to discriminate across these 16 genres. They also tested the classification efficiency on different parts of the web page space (title and meta-content, body, and anchors). The best accuracy (75.7%) was achieved with one of their features sets when applied only to the body and anchors.

[17]: *Number of web pages: 800; Genre annotation: three raters; Genres: help, article, discussion, shop, portrayal (non-private), portrayal (private), link collection, download*. They worked out a genre palette of eight genres following the outcome of a study on genre usefulness. As they aimed at a classification performed on the fly, they assessed features according to the computational effort they required, giving preference to those requiring low or medium effort. They achieved around 70% accuracy with discriminant analysis on the palette of eight genres. Other results relate to groups of genres tailored for web user profiles.

[14] and the follow up [15]: *Number of web pages: 321; Genre annotation: at least two raters; Genres: reportage-editorial, research article, review, home page, Q&A, specification*. They aimed at selecting genre-revealing terms from the training document set using collection of web pages annotated both at topic level and at genre level. Their formula (the deviation formula) makes use of both genre-classified documents and subject-classified documents and eliminate terms that are more subject-related than genre-related. They report a micro-average of precision and recall of about 90%.

As already stressed, the absence of common criteria or evaluation ground makes most of these experiments (see Table 1 for a summary) difficult to compare, however fruitful each study can be in itself. A cross-evaluation of these experiments remains virtually unfeasible because genre palettes are mostly disparate. Also in the case of 'home page', which is probably one of the few genres in common in several experiments, any comparison appear to be difficult, because selection criteria and level of exemplarity are not declared. The two criteria of annotation by objective sources and consistent level of granularity are suggested to overcome this un-comparability.

Table 1. Summary Table

Studies	**No. of web pages**	**Annotation**	**Labels**
[7]	2,150	single rater	Subjectivity vs. objectivity, positive vs. negative
[5]	2,700	One or more raters	public affairs style, everyday communication style, scientific style, journalistic style, literary style
[3]	343	Two or more raters	abstract, call for papers, FAQs, hub/sitemap, job description, resume/C.V., statistics, syllabus, technical paper
[10]	321	do not say	home pages (personal, corporate, organizational)
[16]	1,224	two graduate students	personal home page, public home page, commercial home page, bulletin collection, link collection, image collection, simple table/lists, input pages, journalistic material, research report, official materials, FAQs, discussions, product specification, informal texts
[17]	800	3 raters	article, discussion, shop, portrayal (non-private), portrayal (private), link collection, download
[14] and [15]	321	at least two raters	reportage-editorial, research article, review, home page, Q&A, specification

3 EXPERIMENTS

3.1 7-Web-Genre Collection

The *7-web-genre collection* includes 200 English web pages per genre, amounting to a total of 1,400 web pages (available online at the URL reported in the Introduction). These web pages were collected by the author of this paper in early spring 2005. This collection was built with genres belonging to a consistent level of granularity and applying the annotation by objective source. The seven web genres included in the collection are the following:

1. blog
2. eshop
3. FAQs
4. online newspaper front page
5. list
6. personal home page[2]
7. search page

The web pages included in the 7-web-genre collection were randomly downloaded from the following public archives or portals (download date: Feb-March 2005):

- Blogs:
  - http://www.britblog.com/
  - http://www.nataliedarbeloff.com/augustinearchive.html.
- Eshops:
  - http://www.shops.co.uk/
  - http://www.eshops.co.uk/
- FAQs:
  - http://www.cybernothing.org/faqs/net-abuse-faq.html
  - http://www.irs.gov/faqs/
  - http://www.copyright.gov/help/faq/
  - http://www.aoml.noaa.gov/hrd/tcfaq/tcfaqHED.html
- Newspaper front pages belong to a number of different online newspaper and are available at Internet Archive:
  - www.archive.org
- Personal home pages are heterogeneous, and include academic and administrative personal home pages, as well as more informal personal home pages. They were downloaded from:
  - http://dmoz.org/Society/People/Personal_Homepages/
  - http://www.math.unl.edu/~mbritten/ldt/homepage.html
  - http://www.bradley.edu/people/fac-staff.html
  - http://www.daimi.au.dk/local/map/PeopleandLocationsPeopleFrame.html
  - http://www.mit.edu/Home-byUser.html
  - http://dir.yahoo.com/Society_and_Culture/People/Personal_Home_Pages
  - http://hpsearch.uni-trier.de/hp/a-tree/
  - Search pages comes from:
  - http://www.searchenginecolossus.com/

The web pages included in the genre 'list', were selected searching keywords in Google and selecting relevant web pages from the results. All the lists include one of the following keywords (and orthographic variants) in the heading: *checklist, hot list, table of content,* and *sitemap* (see, for example, Insect Hotlist at http://www.fi.edu/tfi/hotlists/insects.html).

3.2 KI-04 corpus

KI-04 corpus was built following a palette of eight genres suggested by a user study on genre usefulness ([17]). It includes 1,295 English web pages (HTML documents), but only 800 web pages (100 per genre) were used in the experiment described in [17]. In Experiment 1, we used 1,205 web pages because some web pages were empty (both original version, 1,295 web pages, and working version, 1,205 web pages, are available online at the URL reported in the Introduction). KI-04 corpus includes:

1. article (127 web pages)
2. download (151 w. p)
3. link collection (205 w. p)
4. portrayal (priv.) (126 w. p)
5. discussion (127 w. p)
6. help (139 w. p)
7. portrayal (non-priv) (163 w. p.)
8. shop (167 w. p)

The KI-04 corpus was collected using bookmarks from about five people. Some genres were extended to get a better balance. The corpus was sorted by three people, one of them wrote a bachelor thesis (in German) on the corpus building process. One of the author of [17] checked many of the pages, and most of the sorting complied with his understanding of the genre categories. The download date was January 26th, 2004.

[2] 'Personal home page' is the basic level of the superordinate level 'home page' and has 'academic personal home page', 'administrative personal home page', etc. as subordinate level.

3.3 SPIRIT collection

The *SPIRIT collection* is a random crawl carried out in 2001 (see [8]). It contains single web pages and not full websites. The size of the whole collection is about one terabyte, and the number of web pages (mostly HTML files) is about 95 millions. It is multilingual and without any meta-information, apart from a short header including the original URL, the date and time when the pages were crawled from the web, and few other details. It represents a genuine slice of the real web. In Experiment 2, we used only **1,000** English web pages (available online at the URL reported in the Introduction) from this random, multilingual and unclassified collection.

3.4 Experiment 1

The practical aim of Experiment 1 was to build two single-label discrete classification models, one out of the 7-web-genre collection, the other from KI-04 corpus, and compare their accuracy results. Both collections were submitted to the same pre-processing. The unit of analysis was a single static web page in HTML format.

The feature set, called *1_set*, used in Experiment 1 includes:

- the 50 most common words in English;
- 24 Part-of-Speech (POS) tags;
- 8 punctuation marks: full stop (.), colon (:), semi-colon (;), comma (,), exclamation mark (!), question mark (?), apostrophe ('), and quotes (");
- genre-specific words[3];
- 28 HTML tags;
- 1 nominal attribute representing the length of the web page (SHORT, MEDIUM and LONG).

(This feature set, together with a description, is available online at the URL reported in the Introduction). The classification algorithm used both in Experiments 1 and 2 is SMO (which implements the Sequential Minimal Optimisation (SMO) for training support vectors) with default parameters and logistic regression model, from Weka machine learning workbench ([25]). Accuracy results, shown in Table 2, are averaged over stratified 10-fold crossvalidations repeated 10 times.

Table 2. Averaged Accuracies with SMO

Averaged Accuracy on the 7-web-genre collection	Averaged Accuracy on KI-04 corpus
90.6%	**68.9%**

As you can see in Table 2, the accuracy of the model built with the 7-web-genre collection is much higher than the model built with KI-04 corpus, namely +21.7%.

In order to see whether the feature set was too tailored or biased towards the 7-web-genre collection, we compared the accuracy of this feature set on KI-04 corpus with the accuracy rates reported in [17]. To make this comparison possible, we ran discriminant analysis using our feature set on KI-04 corpus. As [17] ran their discriminant analysis only on 800 web pages while we used 1,205 web pages, we converted all the results into percentages. A breakdown of the different accuracy rates achieved with discriminant analysis and two different feature set is shown in Table 3.

Table 3. Accuracy rates with discriminant analysis

KI-04 corpus	**Our feature set**	**[17]'s feature set**
Article	80.3%	81.3%
Discussion	76.4%	68.5%
Download	74.2%	79.6%
Help	59.7%	55.1%
Link Collection	69.3%	67.6%
Portrayal (non-priv)	59.5%	57.9%
Portrayal (priv)	73.8%	67.7%
Shop	68.3%	66.9%
Accuracy	**70.2%**	**68.1%**

Our feature set performs better than [17]'s feature set. Although the difference is rather small (+2.1%), it is statistically significant (chi-square test). This means that our feature set is not biased toward the 7-web-genre collection, but it performs significantly better than [17]'s feature set on KI-04 corpus with discriminant analysis, i.e. the same algorithm used in [17].

3.4.1 Discussion

Experiment 1 compares the accuracies of two models built with the same classification algorithm, the same feature set but different web page collections, the 7-web-genre collection and KI-04 corpus. The accuracy on the 7-web-genre collection (1,400 web pages) is above 90% while the accuracy on KI-04 corpus is definitely lower. A first thought was that our feature set did not represent the genre palette of KI-04 corpus adequately. However, after having compared the performance of our feature set with [17]'s feature set using the same algorithm (discriminant analysis) on the same collection, we saw that the accuracy achieved by our feature set was slightly higher than the accuracy stated in [17]. Although KI-04 corpus contains eight genres, i.e. one genre more than the 7-web-genre collection (error rate usually increases with the number of categories), this does not justify such a wide the gap in the classification accuracy. Also, it is important to stress that genre-specific words are tailored to the genre palette. This means, the genre-specific words used for the 7-web-genre collection account for blogs, search, front page, etc., while those employed for KI-04 corpus include words relate to articles, discussion, download, etc. Since these two genre palettes have two web genres in common, i.e. home page/portrayal (priv) and eshop/shop, in these two cases the same set of genre-specific words was used for both web genre collections. That the feature set used in the KI-04 corpus is not biased towards the 7-web genre collection is confirmed by the results shown in Table 3, where the performance of our features set is higher than [17]'s feature set.

In conclusion, if neither the feature set nor the classification algorithm is the cause of this large discrepancy in accuracy, then the suspicion is that the selection of the web pages representing genres in KI-04 corpus might be responsible for the lower performance. Although the issue of subjectivity of the assignment of genre to web pages needs further investigation (cf. also [4]), for the time being we interpret the higher performance on the 7-web-genre collection as a result of the application of the two criteria of

[3] Genre-specific words were selected through a cursory manual analysis. A total of 13 sets of genre-specific words were built. 13 and not 15 because two sets were shared across the two collections, namely those related to home-page/portrayal (priv) and eshop/shop. It is worth saying that genre-specific words (available online at the URL reported in the Introduction) are not numerous. For example, genre-specific words for the search web genre are only: *search, crawl, directories, engine, find,* and *see*.

annotation by objective sources annotation and consistent genre granularity.

3.5 Experiment 2

The goal of Experiment 2 was to see whether the classification model built with the collection complying to the criteria of annotation by objective source and consistent genre granularity is more effective also for predictive tasks. In other words, predictions are used here as a kind of evaluation metrics of the efficiency of classification models.

In this experiment we used the two classification models built in the previous experiment together with additional models. The practical aim was to make predictions on unclassified and non-annotated web pages, i.e. 1,000 random English web pages from the SPIRIT collection. The relevance of the agreed upon web pages (see Tables 5 and 6) to a genre was manually assessed by the author of this paper (the breakdown of this manual evaluation is available online at the URL reported in the Introduction).

When making a prediction, the classifier returns a probability score to be interpreted in terms of classification confidence. This confidence score can be exploited when assessing the value of a prediction and for setting a threshold for reliable guesses. In order to get predictions on genre labels which were as reliable as possible, we devised an approach inspired by co-training. The basic idea was to exploit three different views (i.e. three different feature sets) on the same data. When the three models built with the three feature sets agreed on the same genre label (3-out-of-3 agreement) at very high confidence score, namely >=0.9, this was for us an indication of a good prediction. Additionally, as we have two web page collections with two different genre palettes, we can have multi-label predictions. Ideally, a web page might get a prediction of "personal home page", following the palette adopted in the 7-web-genre collection, and "portrayal (private)", following the genre palette adopted in KI-04 corpus. Also, as the two palettes are mostly not overlapping, it is interesting to see which palette is more suitable for the classification of this SPIRIT random sample. From the previous experiment we had two models built with a single feature set (*1_set*). To these models, we add four additional models (two per collection) in order to get the three simultaneous views on each collection. The additional two models were built using the feature sets called *2_set* and *3_se*t (these feature sets, together with a description, are available online at the URL reported in the Introduction).

2_set contains the following features:

- POS trigrams;
- 8 punctuation symbols (as above);
- genre-specific words (as above);
- 28 HTML tags (as above);
- 1 nominal attribute representing the length of the web page (as above).

3_set contains the following features:

- 86 linguistic facets[4];
- genre-specific words;
- 6 HTML facets;
- 1 nominal attribute representing the length of the web page (as above).

Table 4 shows the performance of the three feature sets on the two web genre collections.

Table 4. Accuracies of three feature sets on two collections

Classification algorithm: Weka SMO	Averaged accuracy on the 7-web-genre collection	Averaged accuracy on KI-04 corpus
1_set	90.6%	68.9%
2_set	89.4%	64.1%
3_set	88.8%	65.9%

From the summary shown in Table 5, we can see that a very low number of pages were agreed upon by the three classification models (second column) built on the 7-web-page collection. This is not necessarily bad when aiming at high precision (future work will explore the possibility of increasing precision).

Table 5. Correct predictions with the 7-web-genre palette

7 WEB GENRE PALETTE	# OF AGREED UPON WEB PAGES (OUT OF 1,000)	CORRECT GUESSES	INCORRECT GUESSES AND UNCERTAIN	ERROR RATE
BLOG	17	1	16	0.94
ESHOP	11	3	8	0.73
FAQs	8	1	7	0.88
FRONTPAGE	7	0	7	1.00
LISTING	18	7	11	0.61
PHP	44	10	34	0.77
SPAGE	12	6	6	0.50
TOTAL	117	28	89	
PERCENTAGE	**11.7%**	**2.8%**	**8.9%**	

However, predictions are even sparer with the models built using KI-04 corpus (Table 6). As there was no 3-out-of-3 agreement for discussion, download, help, and portrayal (non-private), these genres were evaluated with 2-out-of-3 agreement. No correct guesses were returned for article, discussion, download, and help.

Table 6. Correct predictions with KI-04 corpus

KI-04 CORPUS	# OF AGREED UPON WEB PAGES (OUT OF 1,000)	CORRECT GUESSES	INCORRECT GUESSES AND UNCERTAIN	ERROR RATE
ARTICLE	4	0	4	1.00
DISCUSSION	8	0	8	1.00
DOWNLOAD	4	0	4	1.00
HELP	3	0	3	1.00
LINK	3	3	0	0.00
PORTRAYAL (NON-PRIVATE)	5	1	4	0.80
PORTRAYAL (PRIVATE)	7	3	4	0.57
SHOP	6	3	3	0.50
TOTAL	36	10	26	
PERCENTAGE	**3.6%**	**1%**	**2.6%**	

3.5.1 Discussion

Experiment 2 shows that the classification models built with the 7-web-genre collection return a higher number of predictions. This seems to confirm the interpretation that using the two criteria of objective source annotation and consistent level of granularity ensures better classification models and consequently a higher number of correct predictions. Also, this experiment shows a useful methodology to follow for multi-genre classification of web pages, which can be refined and further investigated in future.

4 CONCLUSIONS

In this paper we pointed out how classification models learned from a web collection annotated by genre using the two criteria of annotation by objective source and consistent level of granularity can return higher accuracy and a higher number of correct predictions.

[4] Linguistic facets and HTML facets are groups of features highlighting an aspect in the communicative context that is reflected in the use of language. They are listed in the URL reported in the Introduction.

The annotation by objective source is not only less subjective and closer to real-world conditions, but also much faster than annotation by human raters, which is usually time-consuming, controversial, and expensive. Further, a collection built with a consistent level of genre granularity seems to be learned more profitably by the classifier. Together, these two criteria enhance the performance of classification algorithms.

However, a full comparison between the results achieved with the two web page collections built with different criteria is not entirely feasible because the two genre palettes are mostly different. Nonetheless, these findings are indicative of a tendency that can be further investigated in future. It is also worth pointing out that objective sources may still contain biases. Biases in web collections relate to the well-known issue of 'corpus representativeness', dating back to Chomsky's aversion to the use of corpora. However, in the present days and with the web available, biases can be alleviated by randomly picking up web pages from several genre-specific web archives or portals.

Although the two criteria of annotation by objective source and consistent level of granularity represent a practical solution that can help genre classification, the concept of genre remains hard to capture computationally and statistically in its entirety.

First, it would be interesting to investigate more about the ideal proportion among corpus size, number of features and number of classes and its influence on classification results. Also, up to now only single-label discrete classification has been tried out in genre classification studies. Experiment 2 implicitly shows an easy method that can be exploited for multi-label classification: the use of concurrent genre palettes over the same unclassified collection. Ideally, the use of several classification models built with different collections annotated by external sources and a consistent granularity, and including different genre palettes can suggest several genre labels for the same web page. Multi-genre documents and genre hybridism are particularly acute when dealing with web pages, which appear much more unpredictable and individualized than paper documents. Using concurrent genre palettes might represent an alternative to the multi-faceted approach by [11]. What is less reassuring is the absence of a proper evaluation metrics for multi-label problems. We leave these problems open to further investigations and invite the genre classification community to make use of the three collections employed in these experiments and now available online.

5 REFERENCES

[1] Argamon, S., Koppel, M., Avneri, G. Routing documents according to style, *Proc. First International Workshop on Innovative Internet Information Systems*, 1998.

[2] Bathia, V. *Analysing Genre. Language Use in Professional Settings*, Longman, London and New York, 1993.

[3] Boese, E. *Stereotyping the Web: Genre Classification of Web Documents*, M.S. Thesis, Colorado State Univ., 2005.

[4] Boese, E and Howe A. Effects of Web Document Evolution on Genre Classification, *CIKM'05*, 2005.

[5] Bravslavski, P. and Tselischev, A. Experiment on Style-Dependent Document Ranking, *Proc. of the 7th Russian Conference on Digital Libraries*, 2005.

[6] Dillon, A. and Gushrowski, B. Genres and the Web: is the personal home page the first uniquely digital genre?, *JASIS*, 51(2), 2000.

[7] Finn, A. and Kushmerick, N. Learning to classify documents according to genre. *JASIST*, Special Issue, 7(5), 2006.

[8] Joho, H. and Sanderson, M. The SPIRIT collection: an overview of a large web collection, *SIGIR Forum*, 38(2) 2004.

[9] Karlgren, J. *Stylistic Experiments for Information Retrieval*, Thesis submitted for the degree of Doctor of Philosophy, Stockholm University, Sweden, 2000.

[10] Kennedy, A. and Shepherd, M. Automatic Identification of Home Pages on the Web, *Proc. 38 HICSS*, 2005.

[11] Kessler, B., Numberg, G. and Shütze, H. Automatic Detection of Text Genre, *Proc. 35 Annual Meeting of the ACL and 8th Conference of the EACL*, 1997.

[12] Kwasnik, B., Crowston, K., Nilan, M. and Roussinov, D. Identifying document genre to improve web search effectiveness. *The Bulletin of the American Society for Information Science and Technology*, 27(2), 23–26, 2000.

[13] Lee, D. Genres, Registers, Text types, Domains, and Styles: Clarifying the concepts and navigating a path through the BNC Jungle, *Language Learning and Technology*, 5(3), 37-72, 2001.

[14] Lee, Y. and Myaeng, S. Automatic Identification of Text Genres and Their Roles in Subject-Based Categorization, *Proc. 37 HICSS*, 2004.

[15] Lee, Y. and Myaeng, S. Text Genre Classification with Genre-Revealing and Subject-Revealing Features, *Proc. 25 Annual International ACM SIGIR*, 145-150, 2002.

[16] Lim, C., Lee, K. and Kim G., Automatic Genre Detection of Web Documents, in Su K., Tsujii J., Lee J., Kwong O. Y. (eds.) *Natural Language Processing*, Springer, Berlin, 2005.

[17] Meyer zu Eissen S. and Stein B. Genre Classification of Web Pages: User Study and Feasibility Analysis, in Biundo S., Fruhwirth T., Palm G. (eds.), *Advances in Artificial Intelligence*, Springer, Berlin, 256-269, 2004.

[18] Paltridge, B. Working with genre: A pragmatic perspective, *Journal of Pragmatics*, 24, 393-406, 1995.

[19] Rauber, A. and Müller-Kögler, A. Integrating Automatic Genre Analysis into Digital Libraries, *ACM/IEEE joint Conference on Digital Libraries*, Roanoke, USA, 2001.

[20] Rosmarin, A. *The Power of Genre*, University of Minnesota Press, Minneapolis, 1985.

[21] Santini, M. Genres In Formation? An Exploratory Study of Web Pages using Cluster Analysis, *Proc. CLUK 05*, 2005.

[22] Santini, M. Some Issues in Automatic Genre Classification of Web Pages. *Proc. of the JADT 2006* Besançon 2006.

[23] Swales, J. *Genre Analysis*, Cambridge University Press, Cambridge, 1990.

[24] Trosborg, A. (ed.), *Analysing Professional Genres*, J. Benjamins Publishing Company, Amsterdam, 2000.

[25] Witten, I. and Frank, E. *Data Mining: Practical Machine Learning Tools and Techniques*, Morgan Kaufmann Publishers, Amsterdam, second edition, 2005.

[26] Yates, J., and Orlikowski, W. Genres of organizational communication: A structural approach to studying communications and media, *Academy of Management Review*, 17(2), 229-326, 1992.