Some Issues in Automatic Genre Classification of Web Pages

Marina Santini

University of Brighton, Lewes Rd, Brighton, UK

Abstract

In this paper, two experiments in automatic genre classification of web pages are presented. These two experiments are designed to highlight three important issues related to genre classification : corpus composition and genre palettes, feature representativeness, and exportability of classification models. Results show the influence of corpus composition and genre palette on classification rates. They also show how well and to what extent feature sets represent genres in a palette, and give an idea of the limitations of the classification models when exported and used for predictive tasks.

Résumé

Dans cet article nous présentons deux expériences d'apprentissage pour le classement automatique des pages web en fonction de différents genres textuels. Ces deux expériences ont été conçues pour mettre en lumière trois aspects importants qui peuvent influer sur le résultat du classement : la composition du corpus et les genres utilisés, la représentativité des traits linguistiques et non-linguistiques utilisés dans les modèles et, enfin, l'exportation des modèles de classement. La première expérience montre que les résultats sont clairement influencés par la composition du corpus et par les genres utilisés. La seconde expérience montre les limites de la représentativité des traits et donne aussi une idées des limites des modèles de classement quand on les exporte sur un autre corpus pour des fonctions prédictives.

Keywords : genre classification, web pages, machine learning, genre prediction

1. Introduction

In this paper, we present two experiments that use machine learning for automatically classifying web pages according to genre. These two experiments are designed to highlight three important issues that should be taken into account when building genre classification models and that have not been addressed so far. The three issues are the following :

1. Corpus composition and genre palette
2. Feature representativeness
3. Exportability of classification models

The first issue, corpus composition and genre palette, concerns the influence that the prototypicality of a document and the genre palette have on the accuracy results of automatic genre classification experiments. Document prototypicality indicates how unambiguously a document represents a genre, while a genre palette is the list of genres included in a collection. Building a genre collection with a palette of disparate genres, and choosing exemplars, i.e. prototypical documents, to unambiguously represent these genres help the

classification algorithm a lot. We will see how different collections built with different criteria return different accuracy results.

The second issue, *feature representativeness*, is closely connected with the previous one. In general, when a genre is not well represented by the features (i.e. the features do not capture the core traits of a genre), the discrimination power of the features is low, and this affects the accuracy results of the automatic classification.

The third issue, *exportability of classification models*, is related to the degree of generalization of the classification models built on one or more collections of documents when applied or transferred to a different collection.

The results of the two experiments give some insight into these three issues. More specifically, Experiment 1 shows differences in accuracy results of classification models built with different document collections and genre palettes (Issue 1). It also shows the differentiated performance of three feature sets, which can be interpreted in terms of how well these features represent the genres in the palette (Issue 2). Experiment 2 is centered upon genre predictions made on an unclassified collection using classification models learned from other corpora. The results of this experiment show how effectively these models can be exported, and consequently the level of generalization they allow (Issue 3).

These two experiments use a single-label discrete, or hard, classification strategy (see Santini, 2005c), following the tradition of automatic genre classification studies.

The inadequacy of the single-label discrete strategy has already been acknowledged theoretically by several scholars (for example, Crowston and Kwasnik, 2004 ; Meyer zu Eissen and Stein, 2004), and seems inappropriate also for our view of genre. We see genres as cultural artifacts, linked to a society or a community, bearing standardized traits but leaving space for the creativity of the text producer. Genres induce predictable expectations in the receiver. They change or are introduced over time, especially under the impulse of a new communication medium (see Santini, 2006). For example, the personal home page (cf. also Roberts, 1998 ; Dillon and Gushrowski, 2000) has standard traits, such as self-narration, personal interests, contact details, and often pictures related to one's life. Nevertheless, these conventions do not hinder the creativity of the producer. When browsing a personal home page as receivers, we expect a blend of standardized information and personal touch. The personal home page has no evident antecedent in the paper world. It sprang up on the web, a new communication medium, to meet web users' need and can be considered a new genre, i.e. a cultural object servicing a community. How many new genres are on the web? At which stage of evolution? Showing what level of hybridism? We do not know. Intra-genre and inter-genre variations, genre transgression, genre colonization, multi-genre documents, genre hybridism, etc. are particularly acute when dealing with web pages, much more unpredictable and individualized than paper documents. However, these issues are hard to handle computationally and statistically. In fact, no statistical or computational model has been proposed so far to address them, apart from the pioneering attempt of a multi-faceted approach by Kessler et al. (1997) and the ongoing work by Santini (2006). Although the single-label discrete classification does not seem appropriate when dealing with genre, its application here allows us to make some comparisons with previous work and highlight some crucial points.

The paper is organized as follows : Section 2 provides an overview of recent work in genre classification ; Section 3 describes some additional issues that should be taken into account when setting up experiments for genre classification ; after a short description of the web page

collections and three feature sets employed in the experiments, Section 4 presents results and discussion. Conclusions are drawn in Section 5.

2. Recent Work in Automatic Genre Classification of Web Pages

Several experiments have been recently carried out with genres and web pages. Here we list the latest work and refer to Santini (2004) for a more comprehensive review.

What becomes evident when looking at them is not only the lack of an agreed definition of genre or web genre. Equally conspicuous is the absence of standardized criteria for building a genre collection. The tendency is to build one's own web page collection following subjective criteria as for the number of genres, genre palette and number of web pages in the collection. Although we think that building a benchmark for genre classification with a single label is difficult and maybe not feasible, because labelling a web page is both hard and controversial (cf. Santini, 2005c), some criteria about corpus composition should be discussed and agreed upon. Without some kind of commonality, any comparison becomes unfeasible. For instance, can we state that the 91% accuracy achieved with 78 features across 10 genres (see Boese, 2005) is better than the accuracy (about 70%) achieved with 35 features across eight genres (see Meyer zu Eissen and Stein, 2004)? These two experiments are based on collections differing in size, web page selection criteria, and genre palette. Although all the experiments reported below are valuable pieces of experience, the overall picture is fragmentary, and the interaction among corpus composition, genre palette and feature representativeness remains obscure.

For all the studies listed here, we report the number of web pages included in the collection, how many people were involved in the annotation, and the categories used for the classification.

Finn and Kushmerick (2006). *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 Finn and Kushmerick, 2006 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 indicate specific genre (such as the '*review*'), these two polarities can hardly be considered as "genres" in themselves (cf. the definition of genre above). Nonetheless, Finn and Kushmerick (2006) did a valuable job because shed some light on the performance of different feature sets across domains.

Bravslavski and Tselischev (2005). *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 functional styles based on the Russian theoretical approach) to a single continuous parameter. Regardless the promising preliminary results, they could see little improvement in relevance ranking when stylistic parameters were included.

Boese (2005). *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 prototypicality on discrimination tasks. Her best accuracy (92.1%) was achieved by one of the feature combinations when applying an automatic feature selection technique.

Kennedy and Shepherd (2005). *Number of web pages : 321 ; Genre annotation : do not say ; Genres : home pages subgenres (personal, corporate, organizational) and some non-home pages, as noise*. They tried the hard task of subgenre discrimination. The best accuracy (71.4%) seems to be achieved on personal home pages with a single classifier, manual feature selection, and without noisy pages.

Lim et al. (2005). *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.

Meyer zu Eissen and Stein (2004). *Number of web pages : 800 ; Genre annotation : do not say ; Genres : help, article, discussion, shop, portrayal (non-private), portrayal (private), link collection, download*. They worked out a genre palette of eight genre following the outcome of their user 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 computational effort. They achieved around 70% accuracy with discriminant analysis on the full set of eight genres. Other results relate to groups of genres tailored for web user profiles.

Lee and Myaeng (2002) and the follow up Lee and Myaeng (2004). *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% for six genre classes listed above.

As already stressed at the beginning of this section, the absence of common criteria or evaluation ground makes most of the experiments on automatic genre classification difficult to compare, however fruitful each study can be in itself. The interaction of the three issues mentioned in the introduction on the results remains opaque and unexplored.

3. Food for Thought : Some Additional Issues

Apart from the difficulties in comparing different studies with each other, there are other problems to take into consideration in genre classification of web pages : noise, overfitting, word features, feature exportability.

Noisy Input. Raw web pages, i.e. web pages downloaded from the web, are very noisy documents, especially if in HTML format. Irregularity of punctuation, spelling mistakes, extra-linguistic elements such as HTML tags, code snippets, etc. can make feature extraction hard. It is difficult to regularize HTML coding, first because its syntax is permissive and second because HTML code is written by humans and software packages (such as Microsoft Frontpage, Dreameaver, and Microsoft Word.) that can have different coding conventions. Cleaning or standardizing utilities, such as the freeware TidyHTML, have low power in this tangle of different coding styles.

But noise is not only physical. There is also noise at textual level. While the linear organization of most of paper documents is still reflected in traditional electronic corpora, such as the British National Corpus (BNC), web pages have a visual organization that allows the inclusion of several functions or different texts with different aims in a single document. The effect of hyperlinking (cf. Haas and Grams, 1998 ; Crowston and Williams, 1999), interactivity and multi-functionality (cf. Shepherd and Watters, 1999) can deeply affect the textuality of web pages, which tend to be more mixed than traditional paper documents.

Number of Features, Corpus Size, and Overfitting. While one of the curses of traditional topical text categorization is the high dimensionality of the search space, the reduction of this space (dimensionality reduction) is not an issue in genre classification. At least it not an issue when content/topic words are not used, because non-topical feature sets tend to be limited. A low number of features prevents overfitting, which occurs when a classifier "is tuned to the contingent characteristics of training data, rather than the constitutive characteristic of the category" (Sebastiani, 2002). Cross-validation is a technique that helps overcome overfitting, but it does not seem very effective, because when a corpus is small and the number of features and categories is high, the accuracy rate tends to be high too. What is a reliable proportion between corpus size and number of features when doing genre classification? How to spot a classification model that overfits regardless of cross-validated results? More findings in relation to these questions are welcomed.

Word features. In automatic genre classification, word features are traditionally topic-neutral words. Usually content/topic words – commonly employed for topical text categorization (cf. Sebastiani, 2002) – are not included.

Karlgren and Cutting 1994, one of the first experiment in genre classification, applied discriminant analysis across the categories of the Brown corpus without using any content/topic words. The authors shrunk Biber's features[1] to easily extractable cues. Content/topic words were not used by Kessler et al. (1997) either. Stamatatos et al. (2000), borrowing from stylometrics, tried the discriminating power of "the 50 most common words in the BNC", mostly function words, across the press genres of part of the Wall Street Journal corpus with encouraging results.

[1] Biber (1988) was not involved in automatic genre classification. His main interest was the variation across speech and writing using a corpus-based approach. He made a clear-cut distinction between genres and text types, and his research focuses on the latter (cf. Biber, 1988 : 68-70).

In general, genre is mostly topic-independent, apart from special cases. In fact, it is true that some topics tend to be dealt with the same genre, for example obituaries are always about somebody's death. Or some genres bear their specialization in their name, such as biography or weather report. But generally speaking, most genres, such as report, editorial, and FAQs, are not linked to any topic. Therefore, it is rather intuitive that, when not dealing with specialized genres, content/topic words cannot capture genre-related differences. Nonetheless, some experiments in genre detection include content/topic words in their feature sets. For instance, Dewdney et al. (2001) compared the efficiency of content/topic words (called "word features"), presentation features (POSs, etc.), and a combined set of the two. Interestingly, although they declared that the combined set performed better, they also acknowledged that the use of presentation features yields a significant advantage over the use of word frequencies in most cases. That some words help genre discrimination is self-evident, for example pronouns and genre-specific terms, such as "FAQs", or "home page". That all content/topic words contribute to topic-independent genre classification is more doubtful.

Feature Exportability. One of the advantages of content/topic-neutral features is that they can be easily exported to other corpora. Once the set has proved successful on a corpus, it can be directly transposed to another collection without any adaptation, because only frequency counts need to be updated according to the new texts. On the contrary, as content/topic words are corpus-dependent, they must be reworked for each corpus. However, not all topic-neutral features can be smoothly exported. For example, POS trigrams (Argamon et al., 1998) must be reworked on each collection. We suggest that the computational effort required by a feature set be assessed not only in terms of easy extractability (cf. Meyer zu Eissen and Stein, 2004), but also in terms of exportability, which can be seen as a contribution to generalization.

4. Experiments

Web Page Collections. The web page collections described below were built by different people, and with different purposes in mind. These differences are reflected in their composition criteria, such as genre palette, annotation of web pages, number of pages representing a genre, and intra-genre variation (prototypicality). As results will show, these factors affect accuracy rates of genre classification models.

The *seven web genre collection* includes 200 web pages per genre, amounting to 1400 web pages. They were collected by the author of this paper in early spring 2005 and are available online (http ://www.nltg.brighton.ac.uk/home/Marina.Santini/, bottom of the page). The seven web genres included in the collection are the following :

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

Meyer zu Eissen web page collection[2] was built following a palette of eight genres suggested by their user study on genre usefulness (see Meyer zu Eissen and Stein, 2004). This collection includes 1,209 web pages (HTML documents), but only 800 web pages (100 per genre) were

[2] Many thanks to S. Meyer zu Eissen for making this collection available for our research.

used in the experiment described in Meyer zu Eissen and Stein (2004). In Experiment 1, we used 1,205 web pages from this collection. The genre palette of Meyer zu Eissen web page collection includes :

1. article
2. download
3. link collection
4. portrayal (private)
5. discussion
6. help
7. portrayal (non-private)
8. shop

The *SPIRIT collection*[3] is a random crawl carried out in 2001 (see Joho and Sanderson, 2004). 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 web pages in English from this random and unclassified collection (this subset is available online at http ://www.nltg.brighton.ac.uk/home/Marina.Santini/, bottom of the page).

Feature Sets. Three feature sets were used for Experiments 1 and 2. Some of the features come from previous genre classification studies, others, such as linguistic facets (Santini, 2005a), genre-specific facets and HTML facets are new (Santini, 2006).

The first feature set (abbreviated as *1_set*) contains:

- the 50 most common words in English ;
- 24 POS tags ;
- 8 punctuation symbols : colon (:), semi-colon (;), comma (,), esclamation mark (!), question mark (?), apostrophe ('), double quotes (") ;
- 7 genre-specific facets for the seven web genre collection and 8 genre-specific facets for Meyer zu Eissen collection ;
- 28 HTML tags ;
- 1 nominal attribute representing the length of the web page (SHORT, MEDIUM and LONG).

The second set (abbreviated as *2_set*) contains:

- 100 POS trigrams for the seven web genre collection and 76 POS trigrams for Meyer zu Eissen collection ;
- 8 punctuation symbols (as above) ;
- genre-specific facets (as above) ;
- HTML tags (as above) ;
- 1 nominal attribute (as above).

[3] Many thanks to M. Sanderson and H. Joho for making this collection available for our research.

The third set (abbreviated as *3_set*) contains:

- 86 linguistic facets ;
- genre-specific facets (as above) ;
- 6 HTML facets ;
- 1 nominal attribute (as above).

4.1. Experiment 1. Building Classification Models

The practical aim of Experiment 1 is to build two sets of single-label discrete classification models. Each of the two sets of classification models is learned from two different collections containing web pages belonging to two different genre palettes, the seven web genre collection and Meyer zu Eissen collection. Each of the two sets of classification models includes three models, one model per feature set. Each feature set represents a different view on the data. Figure 1 shows a diagram of Experiment 1, with three models per set at the bottom level.

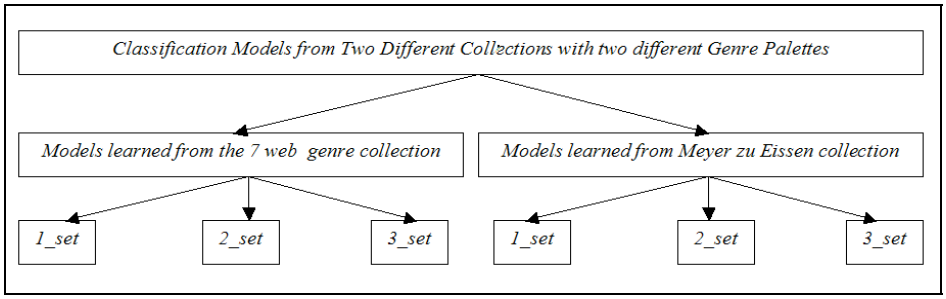

Fig. 1. *Diagram of Experiment 1*

The unit of analysis is a single static web page in HTML format. The classification algorithm used both in Experiment 1 and 2 is SMO (which implements the Sequential Minimal Optimization (SMO) for training support vectors) with default parameters and logistic regression model, from Weka machine learning workbench (Witten and Frank, 2005). Accuracy results, shown in Table 1, are averaged over stratified 10-fold crossvalidations repeated 10 times.

Classification algorithm : Weka SMO	Avg. Accuracy on the 7 web genre collection	Avg. Accuracy on Meyer zu Eissen collection
1_set	90.6%	68.9%
2_set	89.4%	64.1%
3_set	88.8%	65.9%

Table 1. *Accuracy results of three feature sets on two web page collections*

Chi-square tests were used to assess statistically significant differences in the accuracy of the three feature sets on each of the two collections. According to these tests, there are not statistical significant differences among the accuracy of the three feature sets in the seven web genre collection. As for Meyer zu Eissen collection, however, there is a significant difference between the accuracy of *1_set* and *2_set*, but not between *1_set* and *3_set*, neither between *2_set* and *3_set*.

In order to compare these results and the results reported in Meyer zu Eissen and Stein (2004), we ran discriminant analysis using our feature sets on Meyer zu Eissen collection. As Meyer zu Eissen and Stein (2004) ran their discriminant analysis only on 800 web pages, while we used 1,205 web pages, we converted all the results into percentage. A breakdown of the different accuracy rates is shown in Table 2.

Meyer zu Eissen collection	1_set	2_set	3_set	MzE's feature set
Article	80.3%	80.3%	66.9%	81.3%
Discussion	76.4%	71.7%	73.2%	68.5%
Download	74.2%	64.2%	68.9%	79.6%
Help	59.7%	55.4%	54.7%	55.1%
Link Collection	69.3%	70.7%	71.7%	67.6%
Portrayal (non-priv)	59.5%	52.8%	59.5%	57.9%
Portrayal (priv)	73.8%	65.1%	66.7%	67.7%
Shop	68.3%	71.3%	71.3%	66.9%
Accuracy	**70.2%**	**66.4%**	**66.6%**	**68.1%**

Table 2. *Comparison of the accuracy of the three feature sets and Meyer zu Eissen feature set on Meyer zu Eissen collection*

According chi-square tests, *1_set* performs significantly better than Meyer zu Eissen feature set, while Meyer zu Eissen feature set performs significantly better than *2_set* and *3_set*.

Discussion. Experiment 1 compares the accuracy results of several models, built with the same classification algorithm, but different document collections and different features sets. The three feature sets performs very well on the seven web genre collection with an accuracy of about 90%, with small variation due to sampling effect, but no significant differences. Given this good accuracy, we can deduct that they represent the genre palette of the seven web collection appropriately. Accuracy rates returned by the three feature sets on Meyer zu Eissen collection, however, are definitely lower. The first thought is that their representativeness of Meyer zu Eissen genre palette is not ideal. However, if we compare these accuracy rates with the accuracy results achieved by Meyer zu Eissen and Stein (2004) (see Table 2), we can notice that accuracy values are very similar and rather close to each other, even if *1_set* performs significantly better than Meyer zu Eissen feature set, and the latter performs significantly better than *2_set* and *3_set*. Chi-square does not say how large this difference in performance is. Discrepancies can be statistically significant, but very small, therefore almost "insignificant" in practical terms.

4.2. Experiment 2. Exporting Classification Models

The practical aim of Experiment 2 is to use the two sets of classification models built in the previous experiment to make predictions on unclassified web pages, the 1,000 English web pages from the SPIRIT collection. 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 predictions.

In order to get predictions on genre labels which are as reliable as possible, we devised an approach inspired by co-training. The basic idea is to exploit the three different views on the data represented by the three feature sets. When the three models built with the three feature sets agree on the same genre label at very high confidence score, namely >=0.9, this is for us

an indication of a good prediction. Additionally, as we built two sets of models, one per each web page collection, we can have predictions with two different genre palettes. Ideally, a web page might get a prediction of "personal home page", following the palette adopted in the seven web genre collection, and "portrayal (private)", following the genre palette adopted in Meyer zu Eissen collection. 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. The relevance of a web page to a genre was assessed by the author. From the summary shown in Table 3, we can see that a very low number of pages were agreed upon by the three classification models (second column) built on the seven web genre collection. This is not necessarily bad when aiming at high precision. What is less reassuring is the low number of correct guesses (third column) and, consequently, the high error rate (last column).

7 WEB GENRE PALETTE	N. OF AGREED UPON WEB PAGES	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%**	

Table 3. *Correct predictions agreed upon using models built with the seven web genre palette*

Results are even less encouraging with models built using Meyer zu Eissen collection (Table 4). 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.

8 GENRE PALETTE	N. OF AGREED UPON WEB PAGES	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%**	

Table 4. *Correct predictions agreed upon using models built with Meyer zu Eissen palette*

Discussion. Although the classification models built with Experiment 1 looked promising, when applied for predictions on an unclassified random sample of 1,000 web pages, results are spare and error rate high. Classification models built with the seven web genre palette seem more suitable for this random sample than models built with Meyer zu Eissen genre palette.

5. Conclusions

Experiment 1 showed that corpus composition, genre palette and feature representativeness influence and affect the accuracy results of genre classification models. The three feature sets used in this experiment seem more representative of the prototypicality and palette used to build the seven web genre collection (accuracy is around 90%) than of the prototypicality and palette employed for Meyer zu Eissen collection (accuracy is around 70%). On the other hand, the accuracy results achieved by our three feature sets on Meyer zu Eissen collection are very close (sometime better, sometime worse) to the accuracy results achieved by the collection creators.

Experiment 2 showed that it is not straightforward to export classification models learned from specific collections (even when the accuracy of those models is high, as in the case of the seven web genre collection) to a random unclassified web page collection. Do the exported models overfit the data on which they were built upon? Or is the problem represented by the distribution and the proportion of genres in the unclassified set? These questions remain unanswered. Exporting classification models to make predictions seems to be a challenging issue if we think of the unpredictability of web pages on the live web.

In conclusion, the results of these two experiments provide insight into the interaction of corpus composition and genre palettes on classification results, show how well and to what extent the feature sets represent the genres in the palettes, and give an idea of the limitations of the classification models when exported and used for predictive tasks. Automatically classifying web pages by genre using machine learning is hard when approximating a real-world situation. Particularly, with a single-label discrete approach.

Références

Argamon S., Koppel M. and Avneri G. (1998). Routing documents according to style. *Proceedings of the 1st International Workshop on Innovative Internet Information Systems*.

Biber D. (1988). *Variation across speech and writing*. Cambridge University Press, Cambridge.

Boese E. (2005). *Stereotyping the Web : Genre Classification of Web Documents, M.S.* Thesis, Computer Science Department, Colorado State University.

Bravslavski P. and Tselischev A. (2005). Experiment on Style-Dependent Document Ranking. *Proceedings of the 7th Russian Conference on Digital Libraries*, RCDL.

Crowston K. and Kwasnik B. (2004). A Framework for Creating a Facetted Classification for Genres : Addressing Issues of Multidimensionality. *Proceedings of the 37th Hawaii International Conference on System Sciences*.

Crowston K. and Williams M. (1999). The Effects of Linking on Genres of Web Documents. *Proceedings of the 32nd Hawaii International Conference on System Sciences*.

Dewdney N., Vaness-Dikema C. and Macmillan R. (2001). The form is the Substance : Classification of Genres in Text. *ACL '2001 Conference*, Toulouse, France.

Dillon A. and Gushrowski B. (2000). Genres and the Web : is the personal home page the first uniquely digital genre ? *JASIS*, Vol. 51, No. 2.

Finn A. and Kushmerick N. (2006). Learning to classify documents according to genre. To appear *JASIST*, Special Issue on Computational Analysis of Style, Vol. 7, N. 5, March 2006.

Haas, S. and Grams, E. (1998). Page and Link Classifications : Connecting Diverse Resources. *Proceedings of Digital Libraries '98* : 99-107.

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

Karlgren J. and Cutting D. (1994). Recognizing Text Genre with Simple Metrics Using Discriminant Analysis. *Proceedings of COLING 1994*, Kyoto.

Kennedy A. and Shepherd M. (2005). Automatic Identification of Home Pages on the Web. *Proceedings of the 38th Hawaii International Conference on System Sciences*.

Kessler B., Numberg G. and Shütze H. (1997). Automatic Detection of Text Genre. *Proceedings of the 35th Annual Meeting of the ACL and 8th Conference of the EACL*.

Lee Y. and Myaeng S. (2002). Text Genre Classification with Genre-Revealing and Subject-Revealing Features. *Proceedings of the 25th Annual International ACM SIGIR* : 145-150.

Lee Y. and Myaeng S. (2004). Automatic Identification of Text Genres and Their Roles in Subject-Based Categorization. *Proceedings of the 37th Hawaii International Conference on System Sciences*.

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

Meyer zu Eissen S., Stein B. (2004). 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.

Roberts G. (1998). The Home Page as Genre : A Narrative Approach. *Proceedings of the 31st Hawaii International Conference on System Sciences*.

Santini M. (2004). State-of-the-art on Automatic Genre Identification, Tech. Rep. ITRI-04-03.

Santini M. (2005a). Linguistic Facets for Genre and Text Type Identification : A Description of Linguistically-Motivated Features, Tech. Rep. ITRI-05-02.

Santini M. (2005b). Automatic Text Analysis : Gradations of Text Types in Web Pages, *Proceedings of the 10th ESSLLI Student Session, Edinburgh*, UK : 276-285.

Santini M. (2005c). Genres In Formation? An Exploratory Study of Web Pages using Cluster Analysis, *Proceedings of the CLUK 05*.

Santini M. (2006), forthcoming.

Sebastiani F. (2002). Machine learning in automated text categorization. *ACM Computing Surveys*, Vol. 34, N. 1 : 1-47.

Shepherd M. and Watters C. (1999). The Functionality Attribute of Cybergenres. *Proceedings of the 32nd Hawaii International Conference on System Sciences*.

Stamatatos E., Fakotakis N. and Kokkinakis G. (2000). Text Genre Detection Using Common Word Frequencies. *Proceedings of COLING 2000*, Saarbrücken, Germany.

Witten I. and Frank E. (2005). *Data Mining : Practical Machine Learning Tools and Techniques*. Morgan Kaufmann Publishers, Amsterdam.