Implementing a Characterization of Genre for Automatic Genre Identification of Web Pages

Marina Santini
NLTG
University of Brighton
UK
M.Santini@brighton.ac.uk

Richard Power
Computing Department
Open University
UK
r.power@open.ac.uk

Roger Evans
NLTG
University of Brighton
UK
R.P.Evans@brighton.ac.uk

Abstract

In this paper, we propose an implementable characterization of genre suitable for automatic genre identification of web pages. This characterization is implemented as an inferential model based on a modified version of Bayes' theorem. Such a model can deal with genre hybridism and individualization, two important forces behind genre evolution. Results show that this approach is effective and is worth further research.

1 Introduction

The term 'genre' is employed in virtually all cultural fields: literature, music, art, architecture, dance, pedagogy, hypermedia studies, computer-mediated communication, and so forth. As has often been pointed out, it is hard to pin down the concept of genre from a unified perspective (cf. Kwasnik and Crowston, 2004). This lack is also experienced in the more restricted world of non-literary or non-fictional document genres, such as professional or instrumental genres, where variation due to personal style is less pronounced than in literary genres. In particular, scholars working with practical genres focus upon a specific environment. For instance Swales (1990) develops his notion of genre in academic and research settings, Bathia (1993) in professional settings, and so on. In automatic genre classification studies, genres have often been seen as non-topical categories that could help reduce information overload (e.g. Mayer zu Eissen and Stein, 2004; Lim et al., 2005).

Despite the lack of an agreed theoretical notion, genre is a well-established term, intuitively understood in its vagueness. What humans intuitively perceive is that there are categories created within a culture, a society or a community which are used to group documents that share some conventions. Each of these groups is a genre, i.e. a cultural object or artefact, purposely made to meet and streamline communicative needs. Genres show sets of standardized or conventional characteristics that make them recognizable, and this identity raises specific expectations.

Together with conventions and expectations, genres have many other traits. We would like to focus on three traits, namely hybridism, individualization and evolution. Genres are not mutually exclusive and different genres can be merged into a single document, generating hybrid forms. Also, genres allow a certain freedom of variation and consequently can be individualized. Finally, genre repertoires are dynamic, i.e. they change over time, thus triggering genre change and evolution. It is also important to notice that before genre conventions become fully standardized, a genre does not have an official name. A genre name becomes acknowledged when the genre itself has an active role and a communicative function in a community or society (Swales, 1990). Before this acknowledgement, a genre shows hybrid or individualized forms, and indistinct functions.

Putting all these traits together, we suggest the following broad theoretical characterization of genre of written texts: genres are named communication artefacts characterized by conventions, raising expectations, showing hybridism and individualization, and undergoing evolution.

This characterization is flexible enough to encompass not only paper genres (both literary and practical genres), but also digital genres, and more specifically web genres. Web genres or cybergenres (Shepherd and Watters 1998) are those genres created by the combination of the use of the computer and the Internet.

Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions, pages 699–706,
Sydney, July 2006.

Genre hybridism and individualization are very evident on the web. In fact, web pages are often very hybrid because of the wider intra-genre variation and the smaller inter-genre differentiation. They can also be highly individualized because of the creative freedom provided by HTML tags (the building blocks of web pages) or programming languages such as Javascript. We suggest that genre hybridism and individualization can be seen as forces acting behind genre evolution. They allow the upgrade of existing genres and the creation of novel genres.

The change of genre repertoire and the creation of new genres were well illustrated by Crowston and Williams (2000) and Shepherd and Watters (1998). Both these studies describe a similar process. Web genres can start either as reproductions or as unprecedented types of documents. In the first case, existing genres are gradually upgraded and modified to adapt to potentials offered by the web. These variants might become very different from the original genres with time passing by. In the second case, novel genres can be generated from specific needs and requirements of the web. Crowston and Williams (2000) have traced this evolution through a manual qualitative survey of 1000 web pages. Shepherd and Watters (1998) have proposed a fuzzy taxonomy for web genres.

We would like to add a new force in this scenario, namely emerging genres. Emerging genre are those genres still in formation, not fully standardized and without any name or fixed function. For example, before 1998 web logs (or blogs) were already present on the web, but they were not yet identified as a genre. They were just "web pages", with similar characteristics and functions. In 1999, suddenly a community sprang up using this new genre (Blood, 2000). Only at this point, the genre "web log" or "blog" started being spread and being recognized.

Emerging genres may account for all those web pages, which remain unclassified or unclassifiable (cf. Crowston and Williams, 2000) because they show genre mixture or no genre at all. Authors often point out that assigning a genre to a web page might be difficult and controversial (e.g. Roussinov et al., 2001; Meyer zu Eissen and Stein, 2004; Shepherd et al., 2004) because web pages can appear hybrid or peculiar. Genre-mixed web pages or web pages without any evident genre can represent the antecedent of a future genre, but currently they might be considered as belonging to a genre still in formation. It is also important to highlight, however, that since the acknowledgement of genre relies on social acceptance, it is impossible to define the exact point at which a new genre emerges (Crowston and Williams 2000). The multi-facetted model capable of hosting new genres wished for by Kwasnik and Crowston (2004), and the adaptive learning system that can identify genre as they emerge announced by Shepherd et al. (2004) are hard to implement. For this reason, the focus of the method proposed below is not to detect emerging genres, but to show a flexible approach capable of giving account of genre hybridism and individualization.

Flexible genre classification systems are uncommon in automatic genre classification studies. Apart from two notable exceptions, namely Kessler et al. (1997) and Rehm (2006) whose implementations require extensive manual annotation (Kessler et al., 1997) or analysis (Rehm, 2006), genres are usually classified as single-label discrete entities, relying on the simplified assumption that a document can be assigned to only one genre.

In this paper, we propose a tuple representation that maps onto the theoretical characterization of genre suggested above and that can be implemented without much overhead. The implementable tuple includes the following attributes:

(genre(s)) of web pages=<***linguistic features, HTML, text types, [...]***>

This tuple means that web pages can have zero, one or more genres (`(genre(s)) of web pages`) and that this situation can be captured by a number of attributes. For the time being these attributes are limited to `linguistic features, HTML tags, text types`, but in future other attributes can be added (`[...]`). The attributes of the tuple can capture the presence of textual conventions or their absence. The presence of conventions brings about expectations, and can be used to identify acknowledged genres. The absence of conventions brings about hybridism and individualisation and can be interpreted in terms of emerging genres and genre evolution.

In this paper we present a simple model that implement the tuple and can deal with this complex situation. This model is based on statistical inference, performs automatic text analysis and has a classification scheme that includes zero labels, one label or multiple labels. More specifically, in addition to the traditional single-label classification, a zero-label

classification is useful when, for example, a web page is so peculiar from a textual point of view that it does not show any similarity with the genres included in the model. Conversely, a multi-label classification is useful when web pages show several genres at the same time. As there is no standard evaluation metrics for a comprehensive evaluation of such a model, we defer to further research the assessment of the model as a whole. In this paper, we report a partial evaluation based on single-label classification accuracy and predictions.

From a theoretical point of view, the inferential model makes a clear-cut separation between the concepts of 'text types' and 'genres'. Text types are rhetorical/discourse patterns dictated by the purposes of a text. For example, when the purpose of a text producer is to narrate, the narration text type is used. On the contrary, genres are cultural objects created by a society or a community, characterized by a set of linguistic and non-linguistic conventions, which can be fulfilled, personalized, transgressed, colonized, etc., but that are nonetheless recognized by the members of the society and community that have created them, raising predictable expectations. For example, what we expect from a personal blog is diary-form narration of the self, where opinions and comments are freely expressed.

The model presented here is capable of inferring text types from web pages using a modified form of Bayes' theorem, and derive genres through *if-then* rules.

With this model, emerging genres can be hypothesized through the analysis of unexpected combinations of text types and/or other traits in a large number of web pages. However, this potential will be investigated in future work. The results presented here are just a first step towards a more dynamic view of a genre classification system.

Automatic identification of text types and genres represents a great advantage in many fields because manual annotation is expensive and time-consuming. Apart from the benefits that it could bring to information retrieval, information extraction, digital libraries and so forth, automatic identification of text types and genres could be particularly useful for problems that natural language processing (NLP) is concerned with. For example, parsing accuracy could be increased if parsers were tested on different text types or genres, as certain constructions may occur only in certain types of texts. The same is true for Part-of-Speech (POS) tagging and word sense disambiguation. More accurate NLP tools could in turn be beneficial for automatic genre identification, because many features used for this task are extracted from the output of taggers and parsers, such as POS frequencies and syntactic constructions.

The paper is organized as follows: Section 2 reports previous characterization that have been implemented as statistical or computational models; Section 3 illustrates the attributes of the tuple; Section 4 describes the inferential model and reports evaluation; finally in Section 5 we draw some conclusions and outline points for future work.

2 Background

Although both Crowston and Williams (2000) and Shepherd and Watters (1998) have well described the evolution of genres on the web, when it comes to the actual genre identification of web pages (Roussinov et al., 2001; and Shepherd et al., 2004, respectively), they set aside the evolutionary aspect and consider genre from a static point of view. For Crowston and Williams (2000) and the follow-up Roussinov et al. (2001) most genres imply a combination of <purpose/function, form, content>, and, as they are complex entities, a multi-facetted classification seems appropriate (Kwasnik and Crowston, 2004). For Shepherd and Watters (1998) and the practical implementation Shepherd et al. (2004), cybergenres or web genres are characterized by the triple <content, form, functionality>, where functionality is a key evolutionary aspect afforded by the web. Crowston and co-workers have not yet implemented the combination of <purpose/function, form, content> together with the facetted classification in any automatic classification model, but the tuple <content, form, function> has been employed by Rehm (2006) for an original approach to single-web genre analysis, the personal home pages in the domain of academia. Rehm (2006) describes the relationship between HTML and web genres and depicts the evolutionary processes that shape and form web genres. In the practical implementation, however, he focuses only on a single web genre, the academic's personal home page, that is seen from a static point of view. As far as we know, Boese and Howe (2005) is the only study that tries to implement a diachronic view on genre of web pages using the triple

<style, form, content>. This study has the practical aim of finding out whether feature sets for genre identification need to be changed or updated because of genre evolution. They tried to detect the change through the use of a classifier on two parallel corpora separated by a six-year gap. Although this study does not focus on how to detect newly created web genres or how to deal with difficult web pages, it is an interesting starting point for traditional diachronic analysis applied to automatic genre classification.

In contrast, the model described in this paper aims at pointing out genre hybridism and individualisation in web pages. These two phenomena can be interpreted in terms of genre evolution in future investigations.

3 Attributes of the Tuple

The attributes *<linguistic features, HTML tags, text types>* of the tuple represent the computationally tractable version of the combination <purpose, form> often used to define the concept of genre (e.g. cf. Roussinov et al. 2001).

In our view, the purpose corresponds to text types, i.e. the rhetorical patterns that indicate what a text has been written for. For example, a text can be produced to narrate, instruct, argue, etc. Narration, instruction, and argumentation are examples of text types. As stressed earlier, text types are usually considered separate entities from genres (cf. Biber, 1988; Lee, 2001).

Form is a more heterogeneous attribute. Form can refer to linguistic form and to the shape (layout etc.). From an automatic point of view, linguistic form is represented by linguistic features, while shape is represented by HTML tags. Also the functionality attribute introduced by Shepherd and Watters (1998) can be seen in terms of HTML tags (e.g. tags for links and scripts). While content words or terms show some drawbacks for automatic genre identification (cf. Boese and Howe, 2005), there are several types of linguistic features that return good results, for instance, Biberian features (Biber, 1988). In the model presented here we use a mixture of Biberian features and additional syntactic traits. The total number of features used in this implementation of the model is 100. These features are available online at: http://www.nltg.brighton.ac.uk/home/Marina.Santini/

4 Inferential Model

The inferential model presented here (partially discussed in Santini (2006a) combines the advantages of deductive and inductive approaches. It is deductive because the co-occurrence and the combination of features in text types is decided a priori by the linguist on the basis on previous studies, and not derived by a statistical procedure, which is too biased towards high frequencies (some linguistic phenomena can be rare, but they are nonetheless discriminating). It is also inductive because the inference process is corpus-based, which means that it is based on a pool of data used to predict some text types. A few handcrafted *if-then* rules combine the inferred text types with other traits (mainly layout and functionality tags) in order to suggest genres. These rules are worked out either on the basis of previous genre studies or of a cursory qualitative analysis. For example, rules for personal home pages are based on the observations by Roberts (1998), Dillon and Gushrowski (2000). When previous studies were not available, as in the cases of eshops or search pages, the author of this paper has briefly analysed these genres to extract generalizations useful to write few rules.

It is important to stress that there is no hand-coding in the model. Web pages were randomly downloaded from genre-specific portals or archives without any further annotation. Web pages were parsed, linguistic features were automatically extracted and counted from the parsed outputs, while frequencies of HTML tags were automatically counted from the raw web pages. All feature frequencies were normalized by the length of web pages (in tokens) and then submitted to the model.

As stated earlier, the inferential model makes a clear-cut separation between text types and genres. The four text types included in this implementation are: *descriptive_narrative, expository_informational, argumentative_persuasive,* and *instructional*. The linguistic features for these text types come from previous (corpus-)linguistic studies (Werlich 1976; Biber, 1988; etc.), and are not extracted from the corpus using statistical methods. For each web page the model returns the probability of belonging to the four text types. For example, a web page can have 0.9 probabilities of being argumentative_persuasive, 0.7 of being instructional and so on. Probabilities are interpreted in terms of degree or gradation. For example, a web page with 0.9 probabilities

of being argumentative_persuasive shows a high gradation of argumentation. Gradations/ probabilities are ranked for each web page.
The computation of text types as intermediate step between linguistic and non-linguistic features and genres is useful if we see genres as conventionalised and standardized cultural objects raising expectations. For example, what we expect from an editorial is an 'opinion' or a 'comment' by the editor, which represents, broadly speaking, the view of the newspaper or magazine. Opinions are a form of 'argumentation'. Argumentation is a rhetorical pattern, or text type, expressed by a combination of linguistic features. If a document shows a high probability of being argumentative, i.e. it has a high gradation of argumentation, this document has a good chance of belonging to argumentative genres, such as editorials, sermons, pleadings, academic papers, etc. It has less chances of being a story, a biography, etc. We suggest that the exploitation of this knowledge about the textuality of a web page can add flexibility to the model and this flexibility can capture hybridism and individualization, the key forces behind genre evolution.

4.1 The Web Corpus

The inferential model is based on a corpus representative of the web. In this implementation of the model we approximated one of the possible compositions of a random slice of the web, statistically supported by reliable standard error measures. We built a web corpus with four BBC web genres (editorial, Do-It-Yourself (DIY) mini-guide, short biography, and feature), seven novel web genres (blog, eshop, FAQs, front page, listing, personal home page, search page), and 1,000 unclassified web pages from SPIRIT collection (Joho and Sanderson, 2004). The total number of web pages is 2,480. The four BBC genres represent traditional genres adapted to the functionalities of the web, while the seven genres are novel web genres, either unprecedented or showing a loose kinship with paper genres. Proportions are purely arbitrary and based on the assumption that at least half of web users tend to use recognized genre patterns in order to achieve felicitous communication. We consider the sampling distribution of the sample mean as approximately normal, following the Central Limit Theorem. This allows us to make inferences even if the population distribution is irregular or if variables are very skewed or highly discrete. The web corpus is available at: http://www.nltg.brighton.ac.uk/home/Marina.Santini/

4.2 Bayesian Inference: Inferring with Odds-Likelihood

The inferential model is based on a modified version of Bayes' theorem. This modified version uses a form of Bayes' theorem called *odds-likelihood* or *subjective Bayesian method* (Duda and Reboh, 1984) and is capable of solving more complex reasoning problems than the basic version. Odds is a number that tells us how much more likely one hypothesis is than the other. Odds and probabilities contain exactly the same information and are interconvertible. The main difference with original Bayes' theorem is that in the modified version much of the effort is devoted to weighing the contributions of different pieces of evidence in establishing the match with a hypothesis. These weights are confidence measures: Logical Sufficiency (LS) and Logical Necessity (LN). LS is used when the evidence is known to exist (larger value means greater sufficiency), while LN is used when evidence is known NOT to exist (a smaller value means greater necessity). LS is typically a number > 1, and LN is typically a number < 1. Usually LS*LN=1. In this implementation of the model, LS and LN were set to 1.25 and 0.8 respectively, on the basis of previous studies and empirical adjustments. Future work will include more investigation on the tuning of these two parameters.
The steps included in the model are the following:

1) Representation of the web in a corpus that is approximately normal.
2) Extraction, count and normalization of genre-revealing features.
3) Conversion of normalized counts into z-scores, which represent the deviation from the 'norm' coming out from the web corpus. The concept of "gradation" is based on these deviations from the norm.
4) Conversion of z-scores into probabilities, which means that feature frequencies are seen in terms of probabilities distribution.
5) Calculation of prior odds from prior probabilities of a text type. The prior probability for each of the four text types was set to 0.25 (all text types were given an equal chance to appear in a web page). Prior odds are calculated with the formula:

   `prOdds(H)=prProb(H)/1-prProb(H)`

6) Calculation of weighted features, or multipliers (M_n). If a feature or piece of evidence (E) has a

probability >=0.5, LS is applied, otherwise LN is applied. Multipliers are calculated with the following formulae:

```
if  Prob (E)>=0.5 then
    M(E)=1+(LS-1)(Prob(E)-0.5)/0.25
if Prob (E)<0.5 then
    M(E)=1-(1-LN)(0.5-Prob(E))/0.25
```

7) Multiplication of weighted probabilities together, according to the co-occurrence decided by the analyst on the basis of previous studies in order to infer text types. In this implementation the feature co-occurrence was decided following Werlich (1976) and Biber (1988).
8) Posterior odds for the text type is then calculated by multiplying prior odds (step 5) with co-occurrence of weighted features (step 7).
9) Finally, posterior odds is re-converted into a probability value with the following formula:

```
Prob(H)=Odds(H)/1+Odds(H)
```

Although odds contains exactly the same information as probability values, they are not constrained in 0-1 range, like probabilities.

Once text types have been inferred, *if-then* rules are applied for determining genres. In particular, for each of the seven web genre included in this implementation, few hand-crafted rules combine the two predominant text types per web genre with additional traits. For example, the actual rules for deriving a blog are as simple as the following ones:

```
if (text_type_1=descr_narrat_1|argum_pers_1)
if (text_type_2=descr_narrat_2|argum_pers_2)
if (page_length=LONG)
if (blog_words >= 0.5 probabilities)
then good blog candidate.
```

That is, if a web page has description_narration and argumentation_persuasion as the two predominant text types, and the page length is > 500 words (LONG), and the probability value for blog words is >=0.5 (blog words are terms such as *web log, weblog, blog, journal, diary, posted by, comments, archive* plus names of the days and months), then this web page is a good blog candidate.

For other web genres, the number of rules is higher, but it is worth saying that in the current implementation, rules are useful to understand how features interact and correlate.

One important thing to highlight is that each genre is computed independently for each web page. Therefore a web page can be assigned to different genres (Table 1) or to none (Table 2).
Multi-label and no-label classification cannot be evaluated with standard metrics and their evaluation requires further research. In the next subsection we present the evaluation of the single label classification returned by the inferential model.

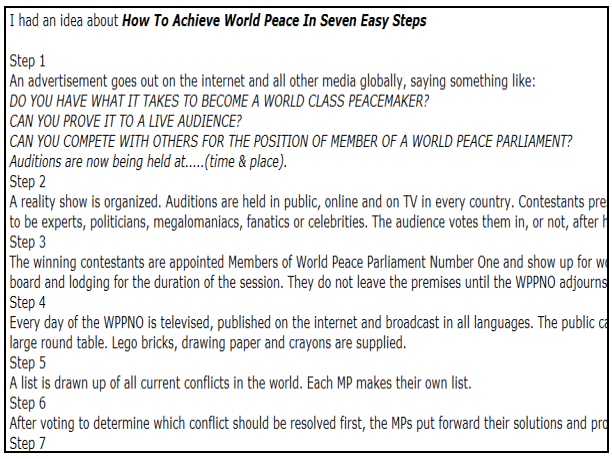
4.3 Evaluation of the Results

Single-label classification. For the seven web genres we compared the classification accuracy of the inferential model with the accuracy of classifiers. Two standard classifiers – SVM and Naive Bayes from Weka Machine Learning Workbench (Witten, Frank, 2005) – were run on the seven web genres. The stratified cross-validated accuracy returned by these classifiers for one seed is ca. 89% for SVM and ca. 67% for Naïve Bayes. The accuracy achieved by the inferential model is ca. 86%.

An accuracy of 86% is a good achievement for a first implementation, especially if we consider that the standard Naïve Bayes classifier returns an accuracy of about 67%. Although slightly lower than SVM, an accuracy of 86% looks promising because this evaluation is only on a single label. Ideally the inferential model could be more accurate than SVM if more labels could be taken into account. For example, the actual classification returned by the inferential model is shown in Table 1. The web pages in Table 1 are blogs but they also contain either sequences of questions and answers or are organized like a how-to document, like in the snippet in Figure 1

blog augustine 0000024	**GOOD blog**	BAD eshop	**GOOD faq**	BAD frontpage	BAD listing	BAD php	BAD spage
blog britblog 00000107	**GOOD blog**	BAD eshop	**GOOD faq**	BAD frontpage	BAD listing	BAD php	BAD spage

Table 1. Examples of multi-label classification


I had an idea about ***How To Achieve World Peace In Seven Easy Steps***

Step 1
An advertisement goes out on the internet and all other media globally, saying something like:
DO YOU HAVE WHAT IT TAKES TO BECOME A WORLD CLASS PEACEMAKER?
CAN YOU PROVE IT TO A LIVE AUDIENCE?
CAN YOU COMPETE WITH OTHERS FOR THE POSITION OF MEMBER OF A WORLD PEACE PARLIAMENT?
Auditions are now being held at.....(time & place).
Step 2
A reality show is organized. Auditions are held in public, online and on TV in every country. Contestants pre
to be experts, politicians, megalomaniacs, fanatics or celebrities. The audience votes them in, or not, after h
Step 3
The winning contestants are appointed Members of World Peace Parliament Number One and show up for w
board and lodging for the duration of the session. They do not leave the premises until the WPPNO adjourns
Step 4
Every day of the WPPNO is televised, published on the internet and broadcast in all languages. The public ca
large round table. Lego bricks, drawing paper and crayons are supplied.
Step 5
A list is drawn up of all current conflicts in the world. Each MP makes their own list.
Step 6
After voting to determine which conflict should be resolved first, the MPs put forward their solutions and pro
Step 7

Figure 1. Snippet *blog_augustine_0000024*

The snippet shows an example of genre colonization, where the vocabulary and text forms of one genre (FAQs/How to in this case) are inserted in another (cf. Beghtol, 2001). These strategies are frequent on the web and might give rise to new web genres. The model also captures a situation where the genre labels available in the system are not suitable for the web page under analysis, like in the example in Table 2.

SPRT_010_049_112_0055685	BAD blog	BAD eshop	BAD faq	BAD frontpage	BAD listing	BAD php	BAD spage

Table 2. Example of zero label classification

This web page (shown in Figure 2) from the unannotated SPIRIT collection (see Section 4.1) does not receive any of the genre labels currently available in the system.

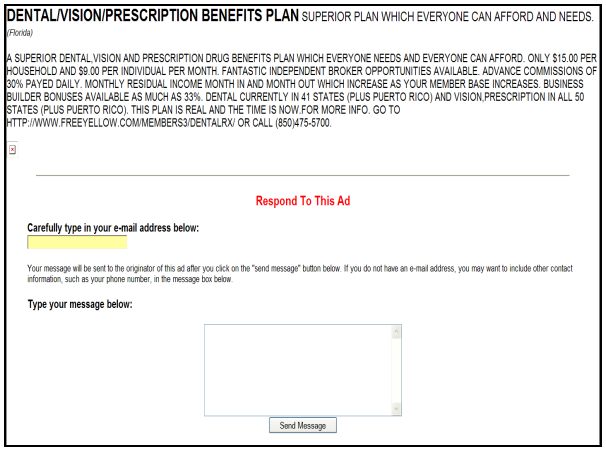


Figure 2. SPRT_010_049_112_0055685

If the pattern shown in Figure 2 keeps on recurring even when more web genres are added to the system, a possible interpretation could be that this pattern might develop into a stable web genre in future. If this happens, the system will be ready to host such a novelty. In the current implementation, only a few rules need to be added. In future implementations hand-crafted rules can be replaced by other methods. For example, an interesting adaptive solution has been explored by Segal and Kephart (2000).

Predictions. Precision of predictions on one web genre is used as an additional evaluation metric. The predictions on the eshop genre issued by the inferential model are compared with the predictions returned by two SVM models built with two different web page collections, Meyer-zu-Eissen collection and the 7-web-genre collection (Santini, 2006). Only the predictions on eshops are evaluated, because eshop is the only web genre shared by the three models. The number of predictions is shown in Table 3.

Models	Total Predictions	Correct Predictions	Incorrect Predictions and Uncertain
Meyer-zu-Eissen and SVM	6	3	3
7-web-genre and SVM	11	3	8
Web corpus and inferential model	17	6	11

Table 3. Predictions on eshops

The number of retrieved web pages (Total Predictions) is higher when the inferential model is used. Also the value of precision (Correct Predictions) is higher. The manual evaluation of the predictions is available online at: http://www.nltg.brighton.ac.uk/home/Marina.Santini/

5 Conclusions and Future Work

From a technical point of view, the inferential model presented in this paper is a simple starting point for reflection on a number of issues in automatic identification of genres in web pages. Although parameters need a better tuning and text type and genre palettes need to be enlarged, it seems that the inferential approach is effective, as shown by the preliminary evaluation reported in Section 4.3.

More importantly, this model instantiates a theoretical characterization of genre that includes hybridism and individualization, and interprets these two elements as the forces behind genre evolution. It is also worth noticing that the inclusion of the attribute 'text types' in the tuple gives flexibility to the model. In fact, the model can assign not only a single genre label, as in previous approaches to genre, but also multiple labels or no label at all. Ideally other computationally tractable attributes can be added to the tuple to increase flexibility and provide a multi-facetted classification, for example register or layout analysis.

However, other issues remain open. First, the possibility of a comprehensive evaluation of the model is to be explored. So far, only tentative evaluation schemes exist for multi-label classification (e.g. McCallum, 1999). Further research is still needed.

Second, in this model the detection of emerging genres can be done indirectly through the analysis of an unexpected combination of text types and/or genres. Other possibilities can be explored in future. Also the objective evaluation

of emerging genres requires further research and discussion.

More feasible in the short term is an investigation of the scalability of the model, when additional web pages, classified or not classified by genre, are added to the web corpus.

Also the possibility of replacing hand-crafted rules with some learning methodology can be explored in the near future. Apart from the approach suggested by Segal and Kephart (2000) mentioned above, many other pieces of experience are now available on adaptive learning (for example those reported in the EACL 2006 on Workshop on Adaptive Text Extraction and Mining).

References

Bathia V. 1993. *Analysing Genre. Language Use in Professional Settings*. Longman, London-NY.

Beghtol C. 2001. The Concept of Genre and Its Characteristics. *Bulletin of The American Society for Inform. Science and Technology*, Vol. 27 (2).

Biber D. 1988. *Variations across speech and writing*. Cambridge University Press, Cambridge.

Blood, R. 2000. Weblogs: A History and Perspective, Rebecca's Pocket.

Boese E. and Howe A. 2005. Effects of Web Document Evolution on Genre Classification. *CIKM 2005*, Germany.

Crowston K. and Williams M. 2000. Reproduced and Emergent Genres of Communication on the World-Wide Web, *The Information Society*, 16(3), 201-216.

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

Duda R. and Reboh R. 1984. AI and decision making: The PROSPECTOR experience. In Reitman, W. (Ed.), *Artificial Intelligence Applications for Business*, Norwood, NJ.

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

Kessler B., Numberg G. and Shütze H. (1997), Automatic Detection of Text Genre, *Proc. 35 ACL and 8 EACL*.

Kwasnik B and Crowston K. 2004. A Framework for Creating a Facetted Classification for Genres: Addressing Issues of Multidimensionality. *Proc. 37 Hawaii Intern. Conference on System Science*.

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

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

Meyer zu Eissen S. and 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.

McCallum A. 1999. Multi-Label Text Classification with a Mixture Model Trained by EM, *AAAI'99 Workshop on Text Learning*.

Rehm G. 2006. Hypertext Types and Markup Languages. In Metzing D. and Witt A. (eds.), *Linguistic Modelling of Information and Markup Languages*. Springer, 2006 (in preparation).

Roberts, G. 1998. The Home Page as Genre: A Narrative Approach, *Proc. 31 Hawaii Intern. Conference on System Sciences*.

Roussinov D., Crowston K., Nilan M., Kwasnik B., Cai J., Liu X. 2001. Genre Based Navigation on the Web, *Proc. 34 Hawaii Intern. Conference on System Sciences*.

Santini M. 2006a. Identifying Genres of Web Pages, *TALN 06 - Actes de la 13 Conference sur le Traitement Automatique des Langues Naturelles*, Vol. 1, 307-316.

Santini M. 2006b. Some issues in Automatic Genre Classification of Web Pages, *JADT 06 – Actes des 8 Journées internationales d'analyse statistiques des donnés textuelles*, Vol 2, 865-876.

Segal R. and Kephart J. 2000. Incremental Learning in SwiftFile. *Proc. 17 Intern. Conf. on Machine Learning*.

Shepherd M. and Watters C. 1998. The Evolution of Cybergenre, *Proc. 31 Hawaii Intern. Conference on System Sciences*.

Shepherd M., Watters C., Kennedy A. 2004. Cybergenre: Automatic Identification of Home Pages on the Web. *Journal of Web Engineering*, Vol. 3(3-4), 236-251.

Swales, J. *Genre Analysis. English in academic and research settings*, Cambridge University Press, Cambridge, 1990.

Werlich E. (1976). *A Text Grammar of English*. Quelle & Meyer, Heidelberg.