Automatic Identification of Home Pages on the Web

Alistair Kennedy and Michael Shepherd
Faculty of Computer Science
Dalhousie University
Halifax, NS, Canada B3H 1W5
{kennedy | shepherd}@cs.dal.ca

Abstract

The research reported in this paper is the first phase of a larger project on the automatic classification of Web pages by their genres. The long term goal is the incorporation of web page genre into the search process to improve the quality of the search results. In this phase, a neural net classifier was trained to distinguish home pages from non-home pages and to classify those home pages as personal home page, corporate home page or organization home page. Results indicate that the classifier is able to distinguish home pages from non-home pages and within the home page genre it is able to distinguish personal from corporate home pages. Organization home pages, however, were more difficult to distinguish from personal and corporate home pages.

1. Introduction

As the World Wide Web continues to grow exponentially, researchers and search engine companies continue to look for techniques that will improve the quality of search results. One method that has been suggested is to classify web pages by their type of genre and use this information to focus a search more narrowly or to rank search results [8, 13]. Experiments by Dewdney et al. [3] have shown that the inclusion of genre information as part of the query can significantly improve precision, while suffering only a modest reduction in recall.

However, the growth of the World Wide Web has been matched by a similar growth in the variety of cybergenres found on the web [16]. This growth includes the replication of existing genres onto the web, the evolution of existings, and the spontaneous appearance of new genres [15]. This expanding and evolving set of web genres makes it very difficult to identify automatically the genre of a web page, thus making it difficult to use in the improvement of the quality of search results. Additionally, it is difficult to know the boundaries of a genre and to know when one has crossed from one genre into another genre [1] or when a web page represents the emergence of a new genre.

Given the dynamic nature of the growth and evolution of web genres, static categories are inappropriate for the classification of web genres. A classification system that is based on adaptive learning is more appropriate in this environment. This is the focus of our larger research project – to apply machine learning techniques to the development of adaptive models that will classify web pages according to genre and will identify new genres as they emerge.

The research reported in this paper is the first phase of this larger project. This phase has focused on the automatic identification of home pages, and the type of home page (sub-genres). A neural net classifier was trained to distinguish home pages from non-home pages and to classify these home pages as personal home page, corporate home page or organization home page. Personal home pages were defined to be home pages that contain information describing the interests and ambitions of a person, where those ambitions do not include making profit through selling some product or service. Corporate home pages were defined as web pages describing the interests and ambitions of companies whose purpose for existing is to make profit through selling some product or service. Organization home pages were defined to be home pages that contain information describing the interests and ambitions of a group (such as a society or religious organization, etc.), where those ambitions do not include making profit through selling some product or service. Organization home pages appear to fill the role of home pages that do not fall into the personal or corporate categories. Results indicate that the classifier is able to distinguish home pages from non-home pages and within the home page genre it is able to distinguish personal from corporate home pages. Organization home pages, however, were more difficult to distinguish from personal and corporate home pages.

0-7695-2268-8/05/$20.00 (C) 2005 IEEE

Section 2 of this paper discusses the growth and evolution of genre on the web, while Section 3 reviews other research on web genre identification. Section 4 introduces the methodology involved in our research while Section 5 presents and discusses the results from this phase of the research. Section 6 summarizes this paper and points the way to further research.

2. Growth of web genre

Although "genre" has been long recognized as a classifying statement [12], the first research to examine the types of genres on the web was done fairly recently. In 1997, Crowston and Williams [2] examined 100 web pages with the intention of looking for reproduced and emergent genres. On the basis of form and purpose, they identified 48 different genres. They identified no search engine or game genres. Of the 100 sampled pages, they found that 80 of the pages more or less faithfully replicated the genres in the traditional media. This is consistent with McLuhan's [10] observation that, "The objectives of new media have tended, fatally, to be set in terms of the parameters and frames of the older media."

Two years later, Shepherd and Watters [15] classified 96 randomly selected web sites on the basis of content, form and functionality. They used a much coarser grained set of criteria and grouped the 96 sites into 5 major categories consisting of: home page, brochure, resource, catalogue and game. Again, no search engines were among the 96 randomly selected web sites.

As this classification was much coarser grained than that of Crowston and Williams, Shepherd and Watters proceeded to map Crowston and Williams' 48 genres into the 5 cybergenres they discovered with the results shown in Table 1. The column headed "S & W" represents the proportion of each cybergenre in Shepherd and Watters' sample of 96 web sites. The column headed "C & W" represents the proportion of each cybergenre after mapping the 48 genres of Crowston and Williams's into the 5 cybergenres.

Table 1. Proportions of cybergenres

Cybergenre	S & W	C & W
Home Page	0.40	0.10
Brochure	0.17	0.06
Resource	0.35	0.82
Catalogue	0.05	0.02
Game	0.03	0.00

Although this was not done statistically, there appears to be significant differences in the proportions of each cybergenre. Shepherd and Watters indicate that while these differences may be due to a number of reasons, they believe the main reason may well be the enormous change that took place on the web over the two years between the studies (1997-1999).

In 2001, Roussinov et al. [13], did a larger study of genre on the web with 184 users. The web pages were tracked and the respondents were asked to report the purpose or task that they were performing when viewing that page. There were 1234 web pages all together. The interviewers coded the web pages with the addition of new genres as needed. There were 116 different genres identified. The respondents were asked to assign their web pages to the appropriate genres. Only 1076 web pages were successfully assigned to genre categories with agreement of only 49.63% between the interviewers and the respondents.

These studies reveal two important issues; firstly, the number of web genres seems to be growing, and secondly, it is often difficult to determine the genre of a web page.

3. Automatic genre identification

In order to apply a machine learning approach to the automatic identification of genres, a feature set must be selected that can be used to distinguish one genre from another and to properly assign a web page or document to a target genre class. The features normally used in genre identification represent the attributes by which genres are normally characterized, i.e., the tuple, <content, form>. However, genres found on the web, cybergenres, may be characterized by the triple, <content, form, functionality>, where functionality is the functionality afforded by the web page [15], and the feature set should also represent the functionality attribute.

The content attribute is normally represented by vectors of terms extracted from the text of the documents. These may be extracted on a statistical basis or they may be extracted on a syntactic basis, such as extracting all noun phrases. The form attribute may be represented by a number of different features including parts-of-speech, punctuation, number of images and positioning on the page. Functionality may be represented by the presence of executable code found in the web page, such as javascript and applets.

Stamatatos et al. [14] used discriminant analysis on the frequencies of commonly occurring terms and punctuation marks with modest success, whereas Lee and Myaeng [9] had better results using word statistics in sets of Korean and English web pages.

Karlgren and Cutting [6] used only form attributes such as parts-of-speech and had good results when the number of target genre categories was only two or four, but achieved only about fifty percent accuracy when the number of target genre categories increased to fifteen. Kessler et al. [7], also used only form attributes, such as

parts-of speech counts, average sentence length, etc. Georg Rehm [11], discusses a series of features for the classification of academic web pages as a genre. These features include such things as: use of logos or graphics of university/departments, alternate version for other languages, home page owners name, pictures or photos of author, contact information (address, phone/fax/e-mail, room number, office hours or secretary phone number).

The literature seems to indicate that results are somewhat better when form and content features are used together. Dewdney et al. [3] found that support vector machines performed equally well when using either content only or form only feature sets, but when the feature sets were combined, the results were significantly better. Their results with a Naïve Bayes classifier showed that performance with a content-based feature set was better than with a form-based feature set but, again, a combined feature set performed best. Finn and Kushmerick [5] examined three feature sets; a bag of words, a part-of-speech vector of ratios of different parts of speech, and a vector of text statistics such as average sentence length and word length. Again, they found that in most cases they had their best results when all three feature sets were used in combination.

The reports of better results when content and form attributes are used in combination makes sense as genres themselves are characterized by the <content, form, functionality> triple. However, none of these studies included the features of the functionality attribute.

4. Methodology

4.1. Dataset

The dataset consisted of 321 web pages, 244 of which were classified as home pages and 77 as noise pages (not home pages). Of the 244 homepages, 17 were classified manually as belonging to two of the three home page sub-genres, giving a breakdown of 94 corporate home pages, 93 personal home pages, 74 organization home pages and 77 noise pages. None of the pages was classified as belonging to all three sub-genres.

4.2. Feature selection

In order to classify these pages, an appropriate set of features needed to be determined. The full set of features that were considered included:

- Content
  - Number of Meta tags used.
  - Does the page contain any phone numbers?
  - List of most common words appearing in between 16% and 40% of all documents.
- Form
  - Number of images.
  - Is a Cascading Style Sheet (CSS) included in this page from another file?
  - Is CSS defined in the header?
  - Is CSS defined at the specific tag where it is used?
  - Does the page have its own domain, or is it in a sub-directory within a domain?
  - Size of file in bytes.
  - Number of words in the page.
- Functionality
  - Number of Links in the Web Page.
  - Number of E-mail Links.
  - Proportion of links that are navigational links to other web pages within the same site.
  - Proportion of links that are links to locations within the same page.
  - Proportion of links that are links to other pages on other sites.
  - Is JavaScript included from an external file?
  - Is JavaScript written into the HTML?
  - Are there any forms?
  - Number of form inputs
  - Is the first tag a Script tag?

The data for these features were normalized so that the mean of every feature was zero and the standard deviation was one. Principle Component Analysis (PCA) was used to preprocess the data and remove features with variance below 0.018.

A subset of features was also constructed manually, through trial and error. Different combinations of the above features were evaluated and the subset of features that produced the best results included the features above, but without:

- Is CSS defined in the header?
- Is CSS included in this page from another file?
- Is CSS defined at the specific tag where it is used?
- Is JavaScript imported from a file?
- Is JavaScript written into the HTML?
- Are there any forms?

In addition, the content feature was examined more closely and a term was identified as being good for classifying a genre if it appeared in more than 21 percent of all web pages of that genre and more than 44 percent of all web pages in the dataset (excluding noise pages) with that term are of that genre. The list of terms is in Table 2. Note that the letter, “t”, is a feature term for personal home pages. Authors of such web pages tend to use contractions ending in apostrophe t. The apostrophes are

replaced by spaces during data cleaning, leaving the letter t as a stand-alone term.

Table 2. List of feature terms selected statistically

Class	Terms
Personal Home Page	my, me, i, t
Corporate Home Page	we, services, service, available, fax, our, us, com, contact, copyright, free, amp
Organization Home Page	events, community, organization, 2004, help, its, members, news, information

4.3. Training, testing and evaluation measures

An artificial neural net was used for these experiments. Although Dewdney et al. [3] had quite good results with a support vector model, the support vector model requires training a separate classifier for each target category whereas a neural net model permits the development of either a separate classifier for each target category or the development of a single classifier with multiple target categories. A single classifier with multiple target categories requires less training than the total training effort required for the training of separate classifiers for each category.

All training and testing was done with 10-fold cross-validation. In 10-fold cross validation, the data is divided into 10 different groups, so that each group contains proportionally the same number of instances of each class. The neural net classifier was tested 10 times. For each iteration a different group from the 10 groups was chosen for testing and the other 9 groups were used for training. The advantage of this method is that it eliminates the possibility of the neural network being misrepresented by giving extremely good or extremely bad results, by chance. The 10-fold cross validation was run 10 times and the mean and standard deviation of the recall and precision of the results were determined.

The experiments were conducted to evaluate the effects of PCA feature selection versus manual selection of features, including the set of noise pages versus excluding the noise pages, and constructing separate classifiers for each of the three sub-genres of home pages versus one classifier with three target output classes. Note that there was no feature set associated with the noise pages, i.e., the non-home pages, and no classifier was trained specifically to recognize "noise". Rather, the classifiers were trained to recognize the three sub-genres of home pages and if the classifiers could not classify a page as one of these sub-genres, then it was deemed to be "noise".

The quality of each classifier was measured using the *F*-measure, which is based on precision and recall measures. For web genre classification, precision is the proportion of web pages assigned to a genre class that were of that specified genre, while recall is the proportion of web pages of a specified genre that were properly classified. The *F*-measure is calculated as follows:

$$\text{Precision } (G_i) = N / |C_i|$$
$$\text{Recall } (Gi) = N / |Gi|$$
$$F\text{-measure } (Gi) = 2PR / (P + R)$$

where:

$|G_i|$ = number of web pages of genre type personal, corporate or organization home page

$|C_i|$ = number of web pages assigned to class labeled personal, corporate or organization home page

N = number of web pages of genre type G_i assigned to class labeled C_i

P = precision

R = Recall

F-measure(G_i) = the quality of the classifier with respect to web pages of genre type G_i

5. Results and discussion

The results of the experiments are shown in Tables 3 and 4. Table 3 shows the *F*-measure values achieved when using the Principle Component Analysis (PCA) for the selection of the features, while Table 4 shows the values achieved when using the manually selected set of features. Each table contains values for experiments when the noise pages (non-home pages) were included and when they were excluded, and for a single classifier with multiple output targets versus a separate classifier for each target output.

Table 3. *F*-measures for PCA selected features

	Noise		No Noise	
	Single Classifier	**Separate Classifier**	**Single Classifier**	**Separate Classifier**
Personal Home Page	0.71	0.72	0.74	0.75
Corporate Home Page	0.51	0.55	0.65	0.68
Organization Home Page	0.32	0.33	0.39	0.41

Table 4. *F*-measures for manually selected features

	Noise		No Noise	
	Single Classifier	**Separate Classifier**	**Single Classifier**	**Separate Classifier**
Personal Home Page	0.71	0.71	0.80	0.79
Corporate Home Page	0.65	0.67	0.68	0.70
Organization Home Page	0.54	0.55	0.61	0.62

From examining Tables 3 and 4, one can make the following observations:

1. The personal home pages were classified the most correctly, under all conditions.
2. While it was possible to classify correctly the personal and corporate home pages, it was significantly more difficult to classify correctly the organization home pages under any of the conditions imposed.
3. The introduction of noise (non-home pages) decreased the accuracy of the classifiers.
4. In general, the classifiers performed better with manually selected features than with PCA selected features. The exceptions to this were that there was no significant difference between the results for manual versus PCA selected features for personal home pages when noise was introduced and, for some reason, for corporate home pages with no noise.
5. Surprisingly, there were no significant differences between results obtained with a single classifier with multiple target output classes and with multiple classifiers, one for each specific output target class.

5.1. Misclassifications

The misclassification tables were examined in order to understand better the resulting classifications and problems in the classifications. The tables are presented in Tables 5 through 12. In each table, the rows represent the known genres and the columns represent the target classes. The target classes are represented by the letters P for personal home page, C for corporate and O for organization home page.

The diagonal of each table represents the number of web pages of that genre type that were correctly classified. Across the rows, one can see the classes across which that genre was distributed by the classifier. Down the columns, one can see how many of each known genre was classified as belonging to the class represented by that column.

The numbers in each table represent the averages of having run the 10 iterations of the classifer (10-fold cross-validation, run 10 times). The classifier evaluated each web page against each target class. If the calculated value fell below the threshold for all three of the target classes, then the web page was deemed not be a home page of any of the three types and was classed as a "non-home" page. However, it is also possible for a web page to be placed into more than one of the three target classes, thus reducing the precision calculation for those classes in which the page does not belong.

From these tables, one can see that the personal home pages are generally well identified by the various classifiers, under all conditions. The problem seems to be in the appropriate classification of the organization home pages. There does seem to be some confusion between the organization home pages and the corporate home pages when the features are selected using PCA. When noise pages are introduced, the classifiers do not perform as well, but when the manually selected features are used the performance seems to be slightly better than when PCA selected features are used. There seems to be no difference between using a single classifier with three

target output classes and using a separate classifier for each target output class.

Tables 5 and 6. Misclassification tables, PCA selected features, no noise pages

Single Classifier

Class	P	C	O	Non-home
Personal	66.2	13.6	4.9	14.9
Corporate	6.3	59.0	26.1	14.3
Organization	12.3	27.4	25.8	17.0

Separate Classifiers

Class	P	C	O	Non-home
Personal	69.0	16.7	2.6	13.0
Corporate	8.4	61.8	31.4	13.5
Organization	12.9	27.7	31.1	15.6

Tables 7 and 8. Misclassification tables, PCA selected features, with noise pages

Single Classifier

Class	P	C	O	Non-home
Personal	61.4	7.1	3.5	23.6
Corporate	2.2	44.3	24.2	31.7
Organization	7.2	17.3	21.3	33.5
Noise Pages	10.7	10.1	8.6	49.0

Separate Classifiers

Class	P	C	O	Non-home
Personal	60.0	8.6	.3.2	24.1
Corporate	0.4	46.4	29.0	30.0
Organization	6.6	18.5	23.7	31.3
Noise Pages	8.2	10.1	10.2	51.5

Tables 9 and 10. Misclassification tables, manually selected features, no noise pages

Single Classifier

Class	P	C	O	Non-home
Personal	71.4	7.4	11.1	8.0
Corporate	7.3	63.2	16.5	16.2
Organization	10.1	17.3	42.8	12.2

Separate Classifiers

Class	P	C	O	Non-home
Personal	70.9	4.5	8.6	12.0
Corporate	4.5	65.0	13.3	18.6
Organization	8.1	21.1	41.9	12.5

Tables 11 and 12. Misclassification tables, manually selected features, with noise pages

Single Classifier

	P	C	O	Non-home
Personal	62.2	3.1	8.2	22.2
Corporate	3.7	56.5	14.8	25.4
Organization	4.8	12.2	36.5	25.9
Noise Pages	11.1	7.4	6.7	52.9

Separate Classifiers

	P	C	O	Non-home
Personal	61.1	1.7	6.5	24.5
Corporate	4.1	58.4	10.3	27.0
Organization	4.3	11.9	36.0	27.5
Noise Pages	11.5	6.6	4.9	55.1

5.2. *k*-means clustering

In this first phase of the research project, all the classifiers were neural networks. There was some concern that the results might be biased because of the type of classifier and that performance might be different with other models of classifiers. Therefore, the dataset was clustered using the *k*-means algorithm and the resulting clusters examined to see the distribution of the various sub-genres of home pages across the clusters. The assumption is that if we see the same types of distributions in the results of the *k*-means clustering as we do in the misclassification tables, then the problems have more to do with the feature set selection than with the type of classifier used.

The *k*-means text clustering algorithm [18] is a top-down or divisive algorithm that partitions the dataset into a non-hierarchical set of clusters. The basic *k*-means algorithm is:

1. Randomly select *k* data objects from the whole dataset;
2. Treat these data objects as the initial cluster centroids;
3. Assign each of the remaining data objects to the most similar cluster, based on the similarity of the object with the cluster centroid;
4. Update the centroid of the cluster;
5. Repeat steps 3 and 4 until the centroids are stable.

Tables 13 through 16 present the results of the *k*-means clustering, where P means personal home page, C means corporate home page and O means organization home page. Recall that 17 web pages were manually assigned to two different genres and this is represented in these tables. For Tables 13 and 15, *k* was set to 3 as we wanted to generate 3 clusters representing personal, corporate and organization home pages. For Tables 14 and 16, *k* was set to 4 as we wanted to identify an additional cluster of noise pages.

Table 13. *K*-means, PCA features, no noise

	P&C	P&O	C&O	P	C	O
Cluster_1	2	0	0	64	3	11
Cluster_2	4	2	3	6	26	28
Cluster_3	3	1	2	11	51	27

Table 14. *K*-means, PCA features, with noise

	P&C	P&O	C&O	P	C	O	Noise
Cluster_1	0	0	0	41	0	10	14
Cluster_2	5	1	1	26	13	11	15
Cluster_3	2	1	3	11	52	31	17
Cluster_4	2	1	1	3	15	14	31

Table 15. *K*-means, manually selected features, no noise

	P&C	P&O	C&O	P	C	O
Cluster_1	1	1	0	58	0	8
Cluster_2	5	1	4	5	44	37
Cluster_3	3	1	1	18	36	21

Table 16. *K*-means, manually selected features, with noise

	P&C	P&O	C&O	P	C	O	Noise
Cluster_1	2	1	0	3	7	8	33
Cluster_2	3	1	4	4	42	36	3
Cluster_3	3	0	1	17	31	18	25
Cluster_4	1	1	0	57	0	4	16

As shown in Tables 13 and 15, without noise pages, the personal home pages fell primarily in one cluster only and there are few web pages in this cluster from other genres. However, there does seem to be some confusion in clusters 2 and 3 in these tables with respect to corporate and organization home pages.

When noise pages are introduced in Tables 14 and 16, personal home pages are still fairly well identified in the system with manually selected features, but not so with the PCA selected features. Again, there is confusion between the corporate and organization home pages. Overall, the organization home pages tend to be dispersed over the four clusters.

Again, the clustering with the manually selected features seems to be better than the clustering with the PCA selected features.

These results are similar to those shown in the misclassification tables, indicating that the resulting classifications are more dependent on the feature set selection than on the choice of the neural net model as the classifier.

6. Discussion and future research

This first phase of the research has shown that home pages can be distinguished from non-home pages with some degree of effectiveness for personal and corporate home pages and that they can be distinguished from each other. However, organization home pages do seem to be more difficult to identify correctly.

It appears that organization home pages do not have a specific style that is unique to them, whereas personal and corporate home pages each have a (more) unique style. Organization home pages can look like either a personal or a corporate home page, depending on who creates the page. When evaluations (not shown in this paper) were conducted using only personal and corporate home pages, the respective *F*-measures ranged from 0.78 to 0.85.

There are a number of open research questions yet to be investigated in this area. One open question is which machine learning model is most appropriate. Dewdney et al. [3] found that the support vector machine model performed somewhat better than the Naïve Bayes model, but the support vector model requires training a separate classifier for each target category. Our work with the neural net model suggests that for a limited number of target categories, a single classifier is sufficient. However, it is still unknown as to whether the neural net model will scale to possibly hundreds of target output classes.

Perhaps the most important open question is the selection of an appropriate feature set. As shown in our discussions of the misclassification tables and the *k*-means clustering results, genre classification is highly dependent on the feature set selected. In order to scale up to many different genres, appropriate features from the genres must be identified. However, with the exception of Dillon and Gushrowski [4], most researchers (including ourselves) tend to simply identify each and every feature they see in the set of web pages. Dillon and Gushrowski performed a user study to identify those features of personal home pages that users thought made for good personal home page design, i.e., which features the users identified as characterizing the genre. This approach would require quite an effort for hundreds of different genres and many different user groups.

Current research tends not to classify the features selected as to <content, form, functionality> attributes. Although Dewdney et al. [30] and Finn and Kushmerick [5] have shown that the combination of content and form is more effective than either just content or form, features that might be classified as functionality attributes are not identified as such, they are simply mixed in with the other features. A follow up [17] to the study reported in this paper, using the feature sets as described in this research, showed that a significant improvement was found in identifying personal and corporate home pages when the functionality attribute was included. A more thorough investigation is warranted.

With the growing importance of the web as the repository of information, it is important to develop mechanisms to improve the quality of search engine results, and the incorporation of genre into the search equation may be one way of doing this [8, 13]. The evolutionary nature of web genre and the fuzzy boundaries of these genres make it difficult to recognize web genres automatically and the static classes used in this preliminary stage of our project are inappropriate. To be truly adaptive in this environment, a classifier would have to:

- Track a recognized genre as it evolves
- Recognize the introduction of a novel genre not seen previously

This first requirement would entail continuous learning while the second requirement would entail an examination of the set of web pages identified as "noise" with possible clustering of this set to identify new classes of genre. While we have not yet addressed this, an examination of the topic detection and tracking literature may provide useful insights into this problem.

7. References

[1] Crowston, K. and B.H. Kwasnik. "A Framework for Creating a Facetted Classification for Genres: Addresssing Issues of Multidimensionality", *Proc. of the 37th Hawaii International Conference on System Sciences*, IEEE Computer Society, Hawaii, 5-8 January 2004.

[2] Kevin Crowston and Marie Williams. "Reproduced and Emergent Genres of Communication on the World Wide Web", *Proc. of the 30th Hawaii International Conference on System Sciences*, IEEE Computer Society, Hawaii, 1997.

[3] Dewdney, N., VanEss-Dykema, C. and R. MacMillan, "The Form is the Substance: Classification of Genres in Text," [http://www.elsnet.org/km2001/dewdnew.pdf] Available 14 June 2004.

[4] Dillon, A. and B. Gushrowski, "Genres and the Web - is the Home Page the First Digital Genre?" *Journal of the American Society for Information Science*, 51, 2, 2000, pp. 202-205

[5] Finn, A. and N. Kushmerick. "Learning to Classify Documents According to Genre" *IJCAI-03 Workshop on Computational Approaches to Style Analysis and Synthesis*, 2003.

[6] Karlgren, J. and D. Cutting. "Recognizing Text Genres with Simple Metrics using Discriminant Analysis", *Proc. of the 15th International Conference on Computational Linguistics (Coling 94),* volume II, Kyoto, Japan, 1994., pp. 1071 – 1075

[7] Kessler, B. Nunberg, G. and H. Schutze. "Automatic Detection of Text Genre", In Philip R. Cohen and Wolfgang Wahlster, (eds.) *Proc. of the Thirty-Fifth Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics,* Association for Computational Linguistics, Somerset, New Jersey, 1997, pp. 32–38.

[8] Kwasnik, B.H., Crowston, K., Nilan, M. and D. Roussinov, "Identifying Document Genre to Improve Web Search Effectiveness". ***Bulletin of The American Society for Information Science and Technology Vol. 27, No. 2 December/January 2001.***

[9] Lee, Y-B. and S.H. Myaeng. "Automatic Identification of Text Genres and Their Roles in Subject-Based Categorization", *Proc. 37th Annual Hawaii International Conference on System Sciences*, IEEE Computer Society, Hawaii, 2004.

[10] McLuhan, M., "Is it natural that one medium should appropriate and exploit another?" In Gerald E. Stern (ed.), *McLuhan: Hot and Cool.* New American Library, Signet Books, New York, 1967. Reprinted in, Eric McLuhan and Frank Zingrone (eds.), *Essential McLuhan*, House of Anansi Press Limited, Concord, Ontario, 1995.

[11] Rehm, G. "Towards Automatic Web Genre Identification", *Proc. of the 35th Annual Hawaii International Conference on System Sciences*, IEEE Computer Society, Hawaii, 2002.

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

[13] Roussinov, D., Crowston, K., Nilan, N., Kwasnik, B., Cai, J. and X. Liu, "Genre Based Navigation on the Web", *Proc. of the 34th Annual Hawaii International Conference on System Sciences*, IEEE Computer Society, Maui, Hawaii, 2001.

[14] Satamatatos, E., Fakotakis, N. and G. Kokkinakis, "Text Genre Detection Using Common Word Frequencies", *Proc. Of the 18th International Converence on Computational Linguistics,* 2000.

[15] Shepherd, M. and C. Watters, "The Evolution of Cybergenres", *Proc. of the 31st Annual Hawaii International Conference on System Sciences*, Maui, Hawaii, 1998.

[16] Shepherd, M. and C. Watters. "Identifying Web Genre: Hitting A Moving Target", *Proc. of the WWW2004 Conference. Workshop on Measureing Web Searach Effectiveness: The User Perspective,* New York, 18 May 2004.

[17] Shepherd, M., Watters, C. and A. Kennedy. "Cybergenre: Automatic Identification of Home Pages on the Web". *Journal of Web Engineering*. To appear.

[18] Steinbach, M., Karypis, G. and V. Kumar, "A Comparison of Document Clustering Techniques", *Proc. of Text Mining Workshop, KDD,* 2000.