Journal of Web Engineering, Vol. 3, No.3&No.4 (2004) 236-251
© Rinton Press

CYBERGENRE: AUTOMATIC IDENTIFICATION OF HOME PAGES ON THE WEB

MICHAEL SHEPHERD, CAROLYN WATTERS and ALISTAIR KENNEDY

Dalhousie University, Halifax, Canada
{ shepherd | watters | kennedy }@cs.dal.ca

Received July 22, 2004
Revised August 30, 2004

The research reported in this paper is part 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. In order to evaluate the importance of the functionality attribute of cybergenre in such classification, the web pages were characterized by the cybergenre attributes of <content, form, functionality> and the resulting classifications compared to classifications in which the web pages were characterized by the genre attributes of <content, form>. 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. A significant improvement was found in identifying personal and corporate home pages when the functionality attribute was included.

Key words: Web systems, genre, cybergenre, web page genre
Communicated by: A Spink

1 Introduction

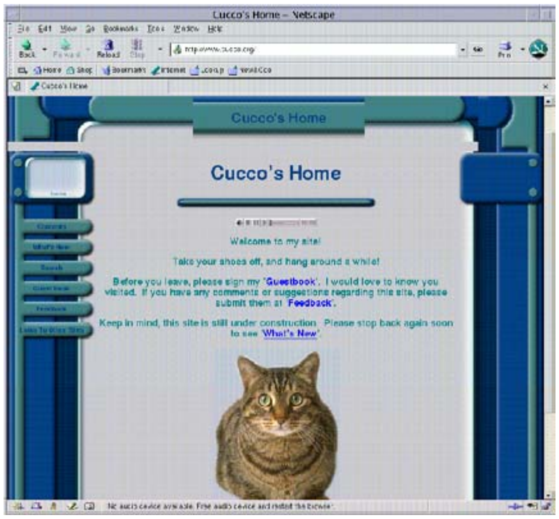
As the World Wide Web continues to grow exponentially, researchers and search engine companies continue to look for techniques that will improve the quality of search results. One method that has been suggested is to classify web pages by their type of genre and use this information to focus a search more narrowly or to rank search results [12]. 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 genre onto the web, the evolution of existing genre, and the spontaneous appearance of new genre [14]. This expanding and evolving set of web genre 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 genre, 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 this research – to apply machine learning techniques to the development of adaptive models that will classify web pages according to genre and will identify new genre 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). As genre are normally characterized by the tuple, <content, form> and cybergenre by the triple, <content, form, functionality> [15], the effect of the functionality attribute on the ability to automatically identify the type of home page was evaluated.

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. Figures 1 through 3 are examples of the three types of home pages.



Cucco's Home

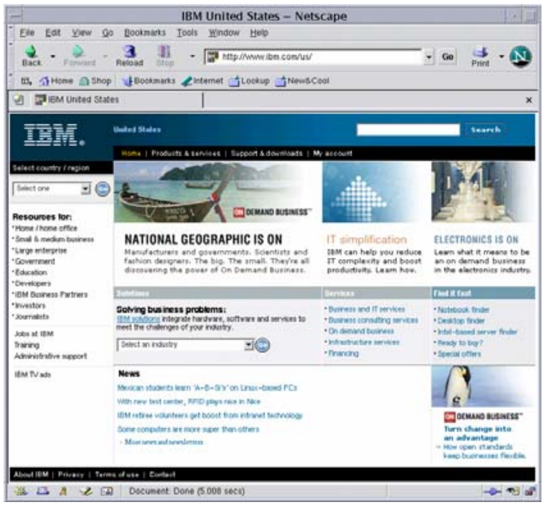
Welcome to my site!

Take your shoes off, and hang around a while!

Before you leave, please sign my 'Guestbook'. I would love to know you visited. If you have any comments or suggestions regarding this site, please submit them at 'Feedback'.

Keep in mind, this site is still under construction. Please stop back again soon to see 'What's New'.

Figure 1. Example of a personal home page.


IBM United States – Netscape
File Edit View Go Bookmarks Tools Window Help
Back Forward Reload Stop
http://www.ibm.com/us/
Go Print
Home Shop Bookmarks Internet Lookup New&Cool
IBM United States
United States
Search
Home | Products & services | Support & downloads | My account
Select country / region
Select one
ON DEMAND BUSINESS
Solutions
Services
Find it fast
Select an industry
About IBM | Privacy | Terms of use | Contact

Resources for:
- Home / home office
- Small & medium business
- Large enterprise
- Government
- Education
- Developers
- IBM Business Partners
- Investors
- Journalists

Jobs at IBM
Training
Administrative support

IBM TV ads

NATIONAL GEOGRAPHIC IS ON
Manufacturers and governments. Scientists and fashion designers. The big. The small. They're all discovering the power of On Demand Business.

IT simplification
IBM can help you reduce IT complexity and boost productivity. Learn how.

ELECTRONICS IS ON
Learn what it means to be an on demand business in the electronics industry.

Solving business problems:
IBM solutions integrate hardware, software and services to meet the challenges of your industry.

- Business and IT services
- Business consulting services
- On demand business
- Infrastructure services
- Financing

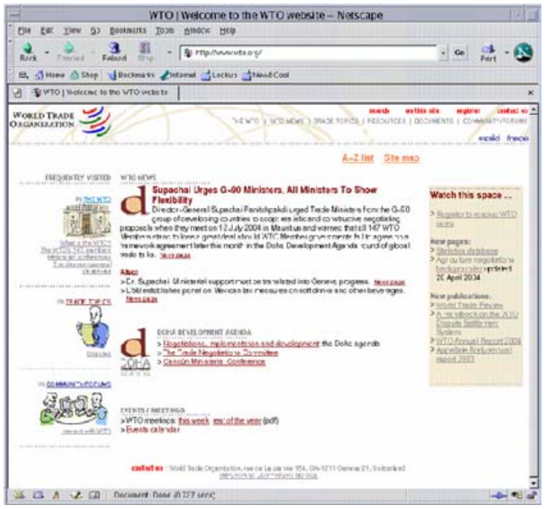
- Notebook finder
- Desktop finder
- Intel-based server finder
- Ready to buy?
- Special offers

News
Mexican students learn "A-B-Si's" on Linux-based PCs
With new test center, RFID plays nice in Nice
IBM retiree volunteers get boost from intranet technology
Some computers are more super than others
› More news and newsletters

ON DEMAND BUSINESS
Turn change into an advantage
→ How open standards keep businesses flexible.

Document: Done (5.008 secs)

Figure 2. Example of a corporate home page.


WTO | Welcome to the WTO website – Netscape
File Edit View Go Bookmarks Tools Window Help
Back Forward Reload Stop
http://www.wto.org/
Go Print
Home Shop Bookmarks Internet Lookup New&Cool
WTO | Welcome to the WTO website
WORLD TRADE ORGANIZATION
search   on this site   register   contact us
THE WTO | WTO NEWS | TRADE TOPICS | RESOURCES | DOCUMENTS | COMMUNITY/FORUMS
español   français
A-Z list   Site map
FREQUENTLY VISITED
WTO NEWS
IN THE WTO
IN TRADE TOPICS
IN COMMUNITY/FORUMS
DOHA DEVELOPMENT AGENDA
EVENTS / MEETINGS

Supachai Urges G-90 Ministers, All Ministers To Show Flexibility
Director-General Supachai Panitchpakdi urged Trade Ministers from the G-90 group of developing countries to adopt realistic and constructive negotiating proposals when they meet on 12 July 2004 in Mauritius and warned that all 147 WTO Member nations [illegible] agree on a framework agreement later this month in the Doha Development Agenda round of global trade talks. More page

Also:
> Dr. Supachai: Ministerial support must be translated into Geneva progress. More page
> [illegible] establishes panel on Mexican tax measures on soft drinks and other beverages. More page

> Negotiations, implementation and development: the Doha agenda
> The Trade Negotiations Committee
> Cancún Ministerial Conference

> WTO meetings: this week, rest of the year (pdf)
> Events calendar

Watch this space ...
> Register to receive WTO news
New pages:
> Statistics database
> Agriculture negotiations backgrounder updated 20 April 2004
New publications:
> World Trade Review
> A handbook on the WTO Dispute Settlement System
> WTO Annual Report 2004
> Appellate Body annual report 2003

contact us World Trade Organization, rue de Lausanne 154, CH-1211 Geneva 21, Switzerland

Document: Done (0.377 secs)

Figure 3. Example of an organization home page.

Section 2 of this paper discusses the concept of cybergenre while Section 3 discusses the growth and evolution of genre on the web, which makes the automatic identification of web genre such a difficult task. Section 4 reviews other research on web genre identification. Section 5 introduces the methodology involved in our research while Section 6 presents and discusses the results from this phase of the research. Section 7 summarizes this paper and points the way to further research.

2 Cybergenre

A genre is a "classifying statement," [11] and is characterized by having similar *content* and *form* where content refers to themes and topics and form refers to, "... observable physical and linguistic features ...," [18]. It allows us to recognize items that are similar even in the midst of great diversity. For instance, the detective novel is a particular genre and we are able to recognize novels as members of that genre, even though the novels themselves may be very different. Once recognized as being of the same genre, we can then more easily compare the individual novels.

As Yates and Orlikowski [18] have shown in their study of the evolution of the business letter of the late 19th century into the electronic mail of today, genres evolve over time in response to institutional changes and social pressures. In some cases, the changes to an existing genre are so extensive that they lead to the emergence of a new genre. One of the triggers for the emergence of variants of existing genres or of new genres is the introduction of a new communications medium [19].

Shepherd and Watters [14] argue that the World Wide Web has been such a powerful trigger that it has resulted in the emergence of cybergenre, a new superclass of genre. While genre are characterized normally by the tuple, <content, form>, cybergenre can be characterized by the triple, <content, form, functionality>, where functionality refers to the capabilities afforded by this new medium.

In much the same way that media like print, radio, television, and movies enable the distribution of information and entertainment to large segments of the population, the web has emerged as a mass medium [17]. The web has become part of the popular culture and the focus of both communities of practice and communities of interest. From the McLuhan perspective [9], the personal and social consequences of a medium result from the new functionalities that the medium affords the user, for example the visual input that movies add to radio, the control that television adds to movies, and the interactivity we gain from the Internet. Typically, the content of a new medium is, at least initially, derived from an earlier one. That is first we try to replicate the successes of the known medium on the new one and then we experiment with the new medium. For example, the early radio shows were patterned after the vaudeville show, early movies the stage play, and early television the radio show. Like other media before it, the Web has developed to the point of having classes of well described genre [17], which merge and transform characteristics of previous media. For example, news web sites merge news content and form of newspapers with the personalization, video and audio, interaction, direct communication, and feedback found in other media.

McLuhan makes an important (if *sixtyish*) distinction between *hot* media like radio and movies and *cool* ones like the telephone or television [9]. In this sense, a hot medium is one that provides "high definition" in terms of information, where high definition means that not much interpretation is required to understand the message. For example, a photograph is a hot medium providing a great deal of visual information while a cartoon is cooler since relatively little visual information is given and the user needs to work harder to provide a context and meaning to that information. Hot media leave little information to the imagination and typically demand little attention and participation from the user while the information content of cool media provides more opportunity for interpretation and participation by the recipient. We see that hot and cool media also exist in digital form; the digitized newspaper is hot and the interactive multimedia news site is cool. The evolution of cybergenre is driven, in part, by increases in functionality afforded by web sites.

3 Genre Identification – A Moving Target

The growth in the number of new and evolved genre on the web makes the automatic identification of web genre a very difficult task. Figure 4 presents a (fuzzy) taxonomy of the classes of subgenres of the class of cybergenres, where the dotted lines represent evolutionary paths between subgenre [14]. The taxonomy is fuzzy as the distinctions among the classes are not clearly defined.

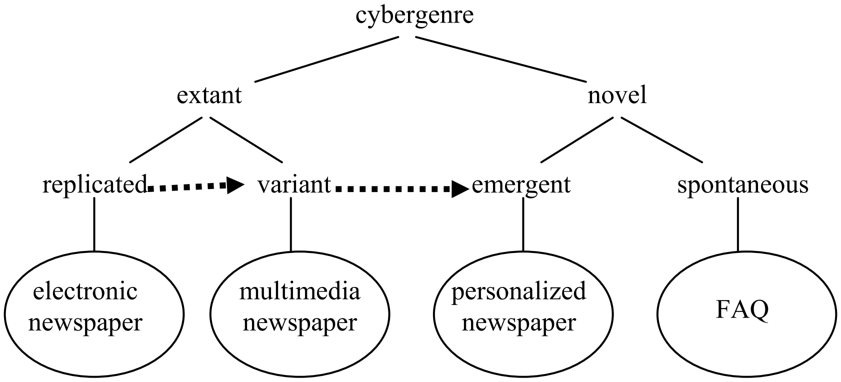


Figure 4. The Evolution of Cybergenres.

According to this taxonomy, cybergenres may be *extant* (i.e., based on existing genres) or they may be *novel* (i.e., not like any existing genre in any other medium). Replicated genres are those based on genres existing in other media. An example of a replicated genre is the early electronic newspaper that attempted to replicate the content and form of the ink-on-paper newspapers with little or no added functionality. This evolved into the multimedia newspaper that incorporated other media such streaming video and eventually into the personalized electronic newspaper where the content, form and functionality of the newspaper are determined by the users themselves and may be driven by adaptive user profiles.

The functionality afforded by the new medium drives the evolution of replicated genres through variants of those genres until novel genres emerge that are significantly different from the original genres. As Erickson [4] points out, "… on-line interaction has the potential to greatly speed up the evolution of genres." In a relatively few years, the replicated e-newspaper has evolved into Web-based, personalized, multi-media electronic news that permits users to link directly to other Web sites and to purchase various commodities.

In addition, the new medium supports the spontaneous creation of new genres that have never existed in other media, such as the Web home page or the FAQ.

Instantiations of novel cybergenres, both emergent and spontaneous, may be either persistent or virtual. Virtual instantiations rely on processes to generate content and/or form as needed and so may be different for different users or even different for the same user at different times. Persistent instantiations rely on stored data and/or forms so that, at any given point in time, simultaneous users could expect the same content and/or form.

Although "genre" has been long recognized as a classifying statement [11], the first research to examine the types of genre 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 genre. 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 [9] 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 [14] 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 genre into the 5 cybergenre 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 genre of Crowston and Williams's into the 5 cybergenre.

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. Although Shepherd and Watters indicate that these differences may be due to a number of reasons, they believe that 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. [12], 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 show two things; firstly, the number of web genres seems to be growing, and secondly, it is often difficult to determine the genre of a web page.

4 Automatic Genre Identification

In order to apply a machine learning approach to the automatic identification of genre, 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., content and form. However, the characterization of genres found on the web may also include the functionality attribute [15], and the feature set used in the machine learning should include all three attributes.

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, applets and hyperlinks.

Stamatatos et al. [13] used discriminant analysis on the frequencies of commonly occurring terms and punctuation marks with modest success, whereas Lee and Myaeng [8] 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 [10], 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 genre themselves are characterized by the <content, form> tuple.. However, none of these studies included the features of the functionality attribute as defined for cybergenre.

5 Methodology

5.1 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.

5.2 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, grouped by attribute type, 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 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. Different combinations of the above features were evaluated on a trial-and-error basis and the subset of features that produced the best results included the features above, but without:

Form

Is 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?

Functionality

Is JavaScript included from an external file?

Is JavaScript written into the HTML?

Are there any forms?

It had been expected that the CSS feature would have appeared in corporate home pages much more often than in either private or organization home pages as it is important to most corporations that their pages have the same look and feel. It is suspected that as this was not the case that the sample size was probably too small to catch this feature. It was also expected that the presence of JavaScript would also have served to differentiate corporate from personal home pages. Again, it is suspected that the sample size was too small and these two features should be evaluated again on a larger sample.

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 shown in Table 2. The letter "t" was found to characterize personal home pages as the result of page authors using contractions that end in apostrophe "t". During indexing, the apostrophe is removed 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

5.3 Training, Testing and Evaluation Measures

An artificial neural net was used for these experiments and 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 effectiveness of the classification of home pages into personal, corporate and organization home pages under the following conditions:

With and without the inclusion of non-home pages (noise pages)

Constructing separate classifiers for each of the three sub-genres of home pages versus one classifier with three target output classes

It is possible for both the separate classifiers for each of the three sub-genres and the one classifier with three target output classes to classify a web page as belonging to more than one of the three target classes

Using the features associated with <content, form, functionality> versus only <content, form> features

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:

Precision (G_i) = N / $|C_i|$

Recall (Gi) = N / |Gi|

F-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

6 Results and Discussion

The results of the experiments are shown in Tables 3-6. Although there are many different combinations of the various conditions evaluated, the information in these tables is presented in such a way as to highlight any significant differences between the resulting classifications when the features associated with the functionality attribute are included versus when they are excluded.

Tables 3 and 4 present the results for the classifications using a separate classifier for each of the three target classes, with and without noise pages (non-home pages) included in the data set. Tables 5 and 6 present the results using a single classifier with three target outputs, with and without noise in the data set.

In each table, the *F*-measure is reported for feature sets that include features associated with the <content, form, functionality> attributes and for feature sets associated with only the <content, form> attributes. If there was a significant difference between these values at the $p \geq .05$ level, there is an entry in the significant difference column. Otherwise there is no entry.

Table 3. F-measures using Separate Classifiers with NoisePpages

	<content, form, functionality>	**<content, form>**	**Significant Difference**
Personal Home Page	.711	.702	-
Corporate Home Page	.666	.637	.005
Organization Home Page	.553	.555	-

Table 4. *F*-measures using Separate Classifiers without Noise Pages

	<content, form, functionality>	**<content, form>**	**Significant Difference**
Personal Home Page	.794	.793	-
Corporate Home Page	.702	.682	.05
Organization Home Page	.623	.608	-

Table 5. *F*-measures using Single Classifier with Noise Pages

	<content, form, functionality>	**<content, form>**	**Significant Difference**
Personal Home Page	.712	.698	.05
Corporate Home Page	.650	.644	-
Organization Home Page	.537	.536	-

Table 6. F-measures using Single Classifier without Noise Pages

	<content, form, functionality>	**<content, form>**	**Significant Difference**
Personal Home Page	.801	.789	.05
Corporate Home Page	.681	.676	-
Organization Home Page	.606	.600	-

From examining Tables 3 through 6, one can make the following observations:

1. Although the results in which the functionality attribute is included were slightly better in almost every instance, the differences were significant in only a few cases. The inclusion of the functionality attribute significantly improved identification of corporate home pages when each class had its own classifier, and of personal home pages when a single classifier with three target outputs was used.
2. The personal home pages were classified the most correctly, under all conditions.
3. 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.
4. The introduction of noise (non-home pages) significantly decreased the accuracy of the classifiers.
5. Surprisingly, in most cases, 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.

6.1 Misclassifications

The misclassification tables were examined in order to understand better the resulting classifications and problems in the classifications. As significant differences between the inclusion/exclusion of the functionality attribute features occurred in only a few instances, the misclassification tables are presented only for the classifications in which the full set of <content, form, functionality> attributes were used.

The tables are presented in Tables 7 through 10. 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, in that a number of the organization home pages are misclassified as corporate home pages. When noise pages are introduced, the classifiers do not perform as well. 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.

Table 7. Misclassification Table, Single Classifier, No Noise Pages, <content, form, functionality>

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

Table 8. Misclassification Table, Separate Classifiers, No Noise Pages, <content, form, functionality>

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

Table 9. Misclassification Table, Single Classifier, with Noise Pages, <content, form, functionality>

Class	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

Table 10. Misclassification Table, Separate Classifiers, with Noise Pages, <content, form, functionality>

Class	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

7 Summary 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 with all machine learning problems, 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, current research tends not to classify the features selected as to <content, form, functionality> attributes. Dewdney et al. [3] and Finn and Kushmerick [5] have shown that the combination of content and form is more effective than either just content or form, but we believe that this is the first study that has addressed the issue of the effectiveness of features representing the functionality attribute. Although the effect of the functionality attribute was significant only for personal and corporate home pages, we believe that more significant results can be attained with more attention paid to the features classified as functionality features.

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, as suggested by Roussinov et al. [12] and shown by Dewdney et al. [3], the incorporation of genre into the search equation may be one way of doing this. As this research project progresses and classifiers are built to identify automatically more different types of cybergenre, the effectiveness of incorporating cybergenre classes into web searching will be measured and reported.

References

1. Crowston, K. and Kwasnik, B.H., A Framework for Creating a Facetted Classification for Genres: Addresssing Issues of Multidimensionality. in Proc. of the 37th Hawaii International Conference on System Sciences, (IEEE Computer Society, Hawaii, 5-8 January 2004).
2. Crowston, K. and Williams, M., Reproduced and Emergent Genres of Communication on the World Wide Web. in Proc. of the 30th Hawaii International Conference on System Sciences, (IEEE Computer Society, Hawaii, 1997).
3. Dewdney, N., VanEss-Dykema, C. and MacMillan, R., The Form is the Substance: Classification of Genres in Text, [http://www.elsnet.org/km2001/dewdnew.pdf] Available 14 June 2004.
4. Erickson, T., Social Interaction on the Net: Virtual Community as Participatory Genre. In Proceedings of the Thirtieth Annual Hawaii International Conference on System Sciences, (Maui, Hawaii, 1997, Vol. 6, pp. 13-21).
5. Finn, A. and Kushmerick, N., Learning to Classify Documents According to Genre. IJCAI-03 Workshop on Computational Approaches to Style Analysis and Synthesis, (2003).
6. Karlgren, J. and Cutting, D., Recognizing Text Genres with Simple Metrics using Discriminant Analysis. In 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 Schutze, H., 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. Lee, Y-B. and Myaeng, S.H., Automatic Identification of Text Genres and Their Roles in Subject-Based Categorization. In Proc. 37th Annual Hawaii International Conference on System Sciences, (IEEE Computer Society, Hawaii, 2004).
9. 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).

10. Rehm, G., Towards Automatic Web Genre Identification. In Proc. of the 35th Annual Hawaii International Conference on System Sciences, (IEEE Computer Society, Hawaii, 2002).
11. Rosmarin, A., The Power of Genre, (University of Minneapolis Press, Minneapolis, 1985).
12. Roussinov, D., Crowston, K., Nilan, N., Kwasnik, B., Cai, J. and Liu, X., Genre Based Navigation on the Web. In Proc. of the 34th Annual Hawaii International Conference on System Sciences, (IEEE Computer Society, Maui, Hawaii, 2001).
13. Satamatatos, E., Fakotakis, N. and Kokkinakis, G., Text Genre Detection Using Common Word Frequencies. In Proc. Of the 18th International Converence on Computational Linguistics, (2000).
14. Shepherd, M. and Watters, C., The Evolution of Cybergenres. In Proc. of the 31st Annual Hawaii International Conference on System Sciences, (Maui, Hawaii, 1998).
15. Shepherd, M. and Watters, C., The Functionality Attribute of Cybergenres. In Proc. of the 32nd Annual Hawaii International Conference on System Sciences, (Hawaii, 1999).
16. Shepherd, M. and Watters, C., Identifying Web Genre: Hitting A Moving Target. In Proc. of the WWW2004 Conference. Workshop on Measureing Web Searach Effectiveness: The User Perspective, (New York, 18 May 2004).
17. Wolf, M.J.P. The Medium and the Video Game. (University of Austin Press, Austin, Texas, 2001).
18. Yates, J. and Orlikowski, W., Genres of Organizational Communication: A Structurational Approach to Studying Communication and Media. In Academy of Management Review, 17(2), 1992, pp. 299-326.
19. Yates, J., Orlikowski, W. and Rennecker, J., Collaborative Genres for Collaboration: Genre Systems in Digital Media. In Proceedings of the Thirtieth Annual Hawaii International Conference on System Sciences, (Maui, Hawaii, 1997, Vol. 6, pp. 50-59).