An Examination of Genre Attributes for Web Page Classification

Lei Dong, Carolyn Watters, Jack Duffy, Michael Shepherd
Faculty of Computer Science
Dalhousie University, Halifax, Nova Scotia, Canada
{ldong | watters | shepherd@cs.dal.ca; jack.duffy@dal.ca }

Abstract

In this paper, we describe a set of experiments to examine the effect of various attributes of web genre on the automatic identification of the genre of web pages. Four different genres are used in the data set, namely, FAQ, News, E-Shopping and Personal Home Pages. The effects of the number of features used to represent the web pages (5, 20, or 100) as well as the types of attributes, <content, form, functionality>, singly and in various combinations are examined.

The results indicate that fewer features produce better precision but more features produce better recall, and that attributes in combinations will always perform better than single attributes.

1. Introduction

A genre is a, "classifying statement," [15] 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 ...," [25]. 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 as opposed to being of some other 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 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 [9, 16]. 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. Another study [13] reported 64% of the students in the study regarded automated genre classification as potentially very useful for web search.

However, the growth of the World Wide Web has been matched by a similar growth in the variety of genres found on the web [20]. This growth includes the replication of existing genres onto the web, the evolution of existing genres, and the spontaneous appearance of new genres [18]. This expanding and evolving set of web genres makes it 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.

The research reported in this paper examines the effectiveness of different types and combinations of genre attributes for the classification of web pages by genre. It is one step in a larger research project that applies machine learning techniques to web page classification according to genre. 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 [18]. In this model, features representing the content attribute may include terms extracted from the text, features representing the form attribute may include tables and graphics, and features representing the functionality attribute may include navigation links. It is the effectiveness of combinations of these three types of attributes for the identification of web genre that is reported in this paper.

1530-1605/08 $25.00 © 2008 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. Sections 4 and 5 introduce the data set and the methodology involved in our research. 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. Growth of web genre

Although "genre" has been long recognized as a classifying statement [15], 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. 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 [12] observation that, "The objectives of new media have tended, fatally, to be set in terms of the parameters and frames of the older media."

In 1998, Shepherd and Watters [18] 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. This research did highlight the enormous change that took place on the web the between the studies.

In 2001, Roussinov et al. [16] 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.

A number of researchers [9,17,20] have postulated that the reasons it appears to be so difficult to identify web genre is that it is like "hitting a moving target" [20], i.e., the web and web genre are evolving at a tremendous rate.

These studies reveal three important issues; the number of web genres seems to be growing, existing web genre are evolving, and lastly, it is often difficult to determine the genre of a web page. The research reported in this paper examines what attributes or combinations of attributes of <content, form, functionality> are most effective in identifying web genre automatically, and how many features of each genre attribute.

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 [19], 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. [22] used discriminant analysis on the frequencies of commonly occurring terms and punctuation marks with modest success, whereas Lee and Myaeng [10] had better results using word statistics in sets of Korean and English web pages.

Stamatatos et al. [23] used discriminant analysis on the features generated with SCBD, which is a Natural Language Processing (NPL) tool for unrestricted Greek text. They found that for the web replications of traditional Greek genres of their interest, discriminant analysis's performance is slightly better than multiple regression and features generated through NPL outperforms the most frequent term method and vocabulary richness method.

Karlgren and Cutting [7] 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. [8], also used only form attributes, such as parts-of speech counts, average sentence length, etc. Georg Rehm [14], 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).

Lim et al. [11] et al. investigated the usefulness of a variety of feature sets using a data corpus of 16 genres from a collection of Korean Web pages. They found most useful features included the URL, HTML tags, token information, function words, punctuation marks and chunks.

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 [5,6] 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 web genres themselves are characterized by the <content, form, functionality> triple. However, none of these studies included the features of the functionality attribute. One study [21] that did attempt to differentiate personal home pages from corporate and organizational home pages found that functionality played little effect, but the data set was fairly small.

4. Data set

A data corpus of 1280 web pages was created that included 170 instances of each of four genres, which are FAQ, News, E-Shopping, and Personal Home Page (PHP). The FAQ, E-Shopping, and News data sets were taken from the Santini genre collection.[1] A set of 170 Personal Home Pages were collected using the random Google feature at http://www.mangle.ca. These four genre types were selected as they appear to be quite different from each other. Although using a larger number of genres would probably lower the classification results as presented in Section 6, we were not as interested in the results *per se*, but rather in the effects of various combinations of the content, form and functionality attributes and the number of representative features of each had on the ability to correctly identify a web genre.

In addition, a random set of 600 noise web pages was also generated using http://www.mangle.ca. A noise page was a web page of any web genre that was not one of FAQ, News, E-Shopping, or Personal Home Page as agreed upon by the three raters. Six hundred noise pages were considered sufficient as only 170 of the pages were selected randomly from this set for these experiments.

All of the data sets were reviewed to ensure that there was 100% agreement on the genre of each item, among the three raters. In the case of the noise data set, the agreement was that none of the noise pages was an instance of Personal Home Page, E-Shopping, News or FAQ, i.e., the noise data set was pure noise.

An open source HTML parser issued by SourceForge [2] was used and reprogrammed so that lexical information from each page was retrieved and organized in accordance with the needs of this project. During the parsing, the retrieved terms were stemmed using Porter's algorithm[3]. No stop words were remove as all candidate features were selected by a feature selection measure. After parsing, the set of terms was reduced by eliminating terms that occurred only one time. Those terms are eliminated because: first, they might be misspelled terms; second, they occur too rarely to discriminate among genres; third, there are many of them so that they would affect performance of further data processing.

[1] http://www.itri.brighton.ac.uk/~Marina.Santini

[2] http://htmlparser.sourceforge.net/

[3] http://www.tartarus.org/~martin/PorterStemmer/

5. Methodology

Previously reported results of this research project [4] addressed the issues of classifiers, feature sizes and feature selection measures for this classification problem. It found that in a binary classification scenario, the results using a Naïve Bayes classifier were as good as those obtained using a neural net classifier or a support vector machine. It also found that the Information Gain [24] measure for feature selection was as good as the Chi Statistic for sizes of the features set, and superior to the Mutual Information measure in the case of a feature set of only 5 features.

On this basis [4] the research reported in this paper is based on using the Naïve Bayes classifier and the Information Gain Measure for feature selection. In addition, the results presented in this paper used the Multivariate Bernoulli Model form of the Naïve Bayes classifier as it significantly out performed the Multinomial Model.

A classification architecture was created consisting of 4 two-way classification subsystems, as per [4]. Each subsystem was trained to classify one of the 4 genres: Personal Home Page, FAQ, E-Shopping, and News.

For each subsystem, a data set was prepared independently. It consisted of 170 web pages of the genre it was associated with as a genre category, and 170 web pages from the randomly generated set of noise web pages. This set of 170 noise pages was then used in all further experiments reported in this paper. An independent feature selection procedure using the Information Gain measure was used on each data set, generating a separate feature space for each subsystem.

10-fold cross validation was used in the experiments, which means the data set in each subsystem is stratified and randomized in each run of the experiment, and 90% of the web pages in each category in a subsystem are used to form a training set on which the subsystem is trained independently. The remaining 10% of the web pages in the data set of each subsystem are used to form a test data set for all the subsystems, i.e. they will be run on each subsystem.

The precision and recall are calculated based on a confusion matrix produced after all 10 runs are over, by taking a micro-average over all the results in each run.

5.1. Experiments

The objectives of these experiments are to determine which attribute or combination of attributes drawn from the triple, <content, form, functionality> are required to correctly identify web page genre, and the effect of the number of representative features of each attribute.

Typical features that would be associated with the content, form and functionality attributes are as follows:

- Content
  - Text visible to the user via the web browser
- Form
  - The name, attribute names and values of the HTML tags; title, Head, Font, Bullet, Div, Style, table, tr (table row), td (table data element)
- Functionality
  - The name, attribute names and values of the HTML tags; Applet, Script, Jsp, Link, Form, Select, Option, Textarea, Input

Table 1 shows typical features of the four web genres used in this research in terms of the content, form and functionality attributes.

The features for each attribute type were selected on the basis of the Information Gain measure, as discussed above. Extensive experimentation was done to determine an appropriate feature set size and the results reported here are for 5, 20 and 100 features. Beyond 100 features, the results slowly decreased while interesting results were obtained at 5, 20 and 100 numbers of features. The features selected were those with the highest Information Gain values [4]. Experiments for 5, 20 and 100 features were run for each of the following 7 sets of attribute type combinations:

- Content
- Form
- Functionality
- Content and Form
- Content and Functionality
- Form and Functionality
- Content, Form and Functionality

Table 1. Typical characteristic of content, form and functionality for each genre type

Web Genre	Content	Form	Functionality
Personal Home Page	Information about the site owner	Hierarchy Info. of related sub-topics	Scroll, emails, links to subtopics
FAQ	Pairs of Q. and A.	List	Scroll, search, links
E-shopping	List of products & services with details	Hierarchy	Scroll, search, emails, online inquiry and ordering
News	Multi-media items, poll, chat forum, News items	Hierarchy time-stamp	Dynamic info, navigation. links, search, login, multi-media on-site play, survey

In each combination case, an equal number of features representing that particular attribute were used. For instance, when Content was the sole attribute investigated, the top 5, 20 and 100 content features were selected for the three sets of feature sizes. However, when the combination of Content, Form and Functionality was investigated, the top 5, 20 and 100 features from each of the types of attributes were selected giving total feature sets of 15, 60 and 300, respectively.

6. Results

The results of the experiments are presented in Tables A1 through A5 in the Appendix. An analysis of variance (ANOVA) was performed to determine the effect on precision and recall of the type of genre, the type of attribute and combination of these attributes, and the number of features representing these attribute types. A summary of the results and the results of the ANOVA follow.

6.1. Effect of genre

Table 2 has the summarized results for precision and recall for the effect of the type of genre. The cell values are the means with the standard deviations in parentheses. These results indicate that it is possible to differentiate successfully among different types of genre.

Table 2. Mean precision and recall for genre. Standard deviations in parenthesis.

Genre	Precision	Recall
E-Shopping	0.920 (0.075)	0.902 (0.073)
FAQ	0.992 (0.023)	0.894 (0.088)
News	0.978 (0.930)	0.987 (0.018)
PHP	0.863 (0.073)	0.939 (0.048)

The effect on precision of the type of genre was significant at $p<0.001$. The partial Eta squared[4] was 0.663. The partial Eta squared is the proportion of total variability attributable to a factor, in this case the type of genre. The ANOVA results indicated that means were significantly different and the effect size of genre type was large. The partial Eta squared was 0.663, which means that genre type itself accounted for 66.3% of the overall (effect+error) variance. The precision for the FAQ genre was significantly better than for the News genre, which was significantly better than for the E-Shopping genre, which was significantly better than for the Personal Home Page genre, at $p=0.05$.

The effect on recall of the type of genre was significant at $p<0.001$. The partial Eta squared was

[4] www.linguistics.ucla.edu/facilities/statistics/power.htm

0.442, indicating that genre type was an important factor for recall, accounting for 44.2% of the overall variance. The recall for the NEWS genre was significantly better than for the Personal Home Page genre, which was significantly better than for the E-Shopping or FAQ genre, at p=0.05. There was no significant difference in recall levels between the E-Shopping and FAQ genres.

6.2. Effect of attributes

Table 3 has the summarized results for precision and recall for the effect of the different types of attributes and combinations of those types. These results indicate that combinations of types of attributes are generally better than using only one attribute, such as content.

Table 3. Mean precision and recall for attribute type. Standard deviations in parenthesis.

Attribute	Precision	Recall
Content	0.905 (0.110)	0.838 (0.172)
Form	0.928 (0.084)	0.926 (0.078)
Functionality	0.938 (0.063)	0.944 (0.065)
Content & Form	0.941 (0.067)	0.952 (0.060)
Content & Functionality	0.947 (0.065)	0.931 (0.114)
Form & Functionality	0.955 (0.056)	0.959 (0.047)
Content & Form & Functionality	0.952 (0.058)	0.965 (0.043)

The effect on precision of which attributes or combinations of attributes used was significant at p<0.001. The partial Eta squared was 0.158, indicating that the contribution of combinations of attributes to the variability was modest, accounting for 15.8% of the overall variance. The effect of functionality and of form were significantly better than for content alone (0.905), at p=0.05, but no significant difference between them. However, the combinations of <content, form, functionality>, <form, functionality> and <content, functionality> were all significantly better at p=0.05 than either form or functionality alone, but there were no significant differences among these combinations. The <content, form> combination was not significantly different than functionality alone, at p=0.05.

The effect on recall of which attributes or combinations of attributes used was significant at p<0.001. The partial Eta squared was 0.470 indicating that the effect of combinations of the attributes to the variability was large, accounting for 47% of the overall variance. Of the single attributes, functionality had significantly better recall than did the form attribute, which had significantly better recall than did the content attribute, at p=0.05. The combinations of <content, form, functionality> and <form, functionality> were both significantly better than the attributes used alone, at p=0.05, but not significantly different from each other. The <content, form> combination is between the single attributes of form and functionality.

6.3. Effect of number of features

Table 4 has the summarized results for precision and recall for the effect of the number of features. The results indicate that the number of features of each attribute is important and that fewer features may be better than more features.

Table 4. Mean precision and recall for number of features. Standard deviations in parenthesis.

No. Features	Precision	Recall
5	0.941 (0.069)	0.916 (0.0108)
20	0.944 (0.076)	0.919 (0.122)
100	0.929 (0.081)	0.938 (0.075)

The effect on precision of the number of features used was significant at p<0.001. Although the means were significantly different, the effect of the number of features was small to modest. The partial Eta squared was just 0.033, which means that the number of features by itself accounted for only 3.3% of the overall (effect+error) variance. The precision for features sets of size 5 and 20 were significantly better than for size 100, at p=0.05.

The effect on recall of the number of features used was significant at $p<0.001$. The partial Eta squared was 0.149, indicating that the number of features had a modest effect, accounting for 14.9% of the overall variance. The recall for the features set of size 100 was significantly better than either size 5 or 20, at $p=0.05$. The difference in recall for feature sets of size 5 and 20 were not significant at $p=0.05$.

7. Summary

This examination of the effect on the automatic classification of web pages by genre has shown that the type of genre, the type of attributes and combination of those attributes, and the number of features representing each type of attribute used have significant effects on the effectiveness of the classification, measured in terms of precision and recall.

Although we cannot generalize for both recall and precision, we can generalize for either recall or for precision. The type of genre significantly affects the recall and the precision. A smaller feature set is more effective in increasing precision than it is in increasing recall, while a larger feature set is more effective in increasing recall.

In examining the results for the types of attributes and combinations of attributes, combinations of attributes improve both recall and precision, and the functionality attribute appears to be important in the combinations.

This research is continuing and experimentation is ongoing with respect to the effect of noise and the combinations of known genres and the identification of web pages that may belong to multiple genres.

8. Acknowledgment

We would like to thank Marina Santini for making her web genre data set available to us for this research.

9. 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, Hawaii, 5-8 January 2004.

[2] Crowston, K. and M. Williams, "Reproduced and Emergent Genres of Communication on the World-Wide-Web". Proc. of the 30th Hawaii International Conf. on Systems Sciences, Vol. 6, 1997, pp. 30-39

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

[4] Dong, L., C. Watters, J. Duffy, and M. Shepherd, "Binary Cybergenre Classification Using Theoretic Feature Measures" .IEEE / WIC / ACM International Conference on Web Intelligence (WI 2006). 2007

[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] Finn, A. and N. Kushmerick, Learning to classify documents according to genre. *Journal of the American Soc. for Inf. Science and Technology*, 2006

[7] 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

[8] Kessler, B., G. Nunberg, 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, Assoc. for Computational Linguistics, Somerset, New Jersey, 1997, pp. 32–38

[9] Kwasnik, B.H., K. Crowston, M. Nilan, 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

[10] Lee, Y-B. and S.H. Myaeng. "Automatic Identification of Text Genres and Their Roles in Subject-Based Categorization", Proc. 37th Hawaii Int. Conf. on System Sciences, Hawaii, 2004

[11] Lim, C.S., K.J. Lee, J.C. and Kim, "Multiple sets of features for automatic genre classification of web documents". *Information Processing & Management* Volume 41, Issue 5, 2005, Pages 1263-1276

[12] 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.

[13] Meyer zu Eissen S. and B. Stein, "Genre classification of web pages". In Proc. of the 27th German Conference on AI (KI-2004). Sept. 2004

[14] Rehm, G. "Towards Automatic Web Genre Identification", Proc. of the 35th Hawaii International Conference on System Sciences, Hawaii, 2002

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

[16] Roussinov, D., K. Crowston, N. Nilan, B.H. Kwasnik, J. Cai, and X. Liu, "Genre Based

Navigation on the Web", Proc. of the 34th Hawaii Int. Conf. on System Sciences, Maui, Hawaii, 2001

[17] Santini, M. 2005. "Genres In Formation? An Exploratory Study of Web Pages using Cluster Analysis", Proc. of the 8th Annual Colloquium for the UK Special Interest Group for Comp. Linguistics (CLUK 05), Uni. of Manchester (UK)

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

[19] Shepherd, M. and C. Watters, "The Functionality Attribute of Cybergenres". In Proceedings of the 32nd Hawaii Int. Conf. on System Sciences. 1999.

[20] 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

[21] Shepherd, M., C. Watters, and A. Kennedy. "Cybergenre: Automatic Identification of Home Pages on the Web". *J. of Web Engineering*. 2004.

[22] Stamatatos, E., N. Fakotakis and G. Kokkinakis, "Text genre detection using common word frequencies". Int. Conf. On Computational Linguistics archive Proc. of the 18th conference on Computational linguistics - Volume 2. 2000a.

[23] Stamatatos, E. N. Fakotakis, and G. Kokkinakis, "Automatic text categorization in terms of genre and author," *Computational Linguistics,* Vol. 26, pp. 471-498, 2000b.

[24] Yang, Y. and O. Pedersen. "A comparative study on feature selectionin text categorization." In Proceedings of ICML-97. 14th International Conf. on Machine Learning. 1997. Pp. 412-420.

[25] J. Yates and W. Orlikowski. "Genres of Organizational Communication: A Structurational Approach to Studying Communication and Media", *Academy of Management Review*, 17(2), 1992, pp. 299-326.

Appendix

Table A1. Average recall and precision results over all genres

Genre Attributes	Feature size						Average (STD)	
	5		20		100			
	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall
Content(Cn)	0.902 (0.103)	0.786 (0.178)	0.905 (0.121)	0.801 (0.200)	0.909 (0.108)	0.926 (0.079)	0.905 (0.110)	0.838 (0.172)
Form(Fm)	0.938 (0.063)	0.903 (0.093)	0.934 (0.078)	0.932 (0.063)	0.912 (0.106)	0.942 (0.073)	0.928 (0.084)	0.926 (0.078)
Functionality(Fn)	0.934 (0.064)	0.929 (0.087)	0.951 (0.057)	0.947 (0.046)	0.930 (0.069)	0.955 (0.052)	0.938 (0.063)	0.944 (0.065)
Cn&Fm	0.942 (0.065)	0.929 (0.076)	0.947 (0.068)	0.957 (0.055)	0.934 (0.070)	0.970 (0.036)	0.941 (0.067)	0.952 (0.060)
Cn&Fn	0.951 (0.065)	0.948 (0.050)	0.946 (0.070)	0.866 (0.172)	0.946 (0.061)	0.968 (0.038)	0.947 (0.065)	0.931 (0.114)
Fm&Fn	0.967 (0.039)	0.947 (0.045)	0.968 (0.042)	0.968 (0.045)	0.931 0.073)	0.961 (0.050)	0.955 (0.056)	0.959 (0.047)
Cn&Fm&Fn	0.957 (0.049)	0.963 (0.048)	0.960 (0.062)	0.965 (0.041)	0.939 (0.063)	0.967 (0.040)	0.952 (0.058)	0.965 (0.043)
Overall Average (STD)	0.941 (0.069)	0.916 (0.108)	0.944 (0.076)	0.919 (0.122)	0.929 (0.081)	0.956 (0.056)	0.938 (0.075)	0.930 (0.101)

Table A2. Recall and precision results for E-Shopping genre

Genre Attributes	Feature size						Average(STD)	
	5		20		100			
	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall
Content(Cn)	0.957	0.709	0.941	0.766	0.934	0.828	0.944 (0.058)	0.768 (0.099)
Form(Fm)	0.872	0.776	0.935	0.918	0.755	0.995	0.821 (0.081)	0.896 (0.106)
Functionality(Fn)	0.890	0.814	0.953	0.933	0.867	0.990	0.903 (0.065)	0.912 (0.083)
Cn&Fm	0.932	0.843	0.960	0.900	0.893	0.961	0.928 (0.565)	0.901 (0.073)
Cn&Fn	0.974	0.895	0.979	0.919	0.947	0.976	0.967 (0.037)	0.930 (0.054)
Fm&Fn	0.974	0.909	0.960	0.966	0.849	1	0.928 (0.070)	0.958 (0.046)
Cn&Fm&Fn	0.960	0.904	0.989	0.952	0.897	0.990	0.949 (0.055)	0.949 (0.051)
Overall Average (STD)	0.937 (0.06)	0.836 (0.919)	0.945 (0.065)	0.908 (0.081)	0.877 (0.081)	0.963 (0.067)	0.920 (0.075)	0.902 (0.073)

Table A3. Recall and precision results for FAQ genre

Genre Attributes	Feature size						Average(STD)	
	5		20		100			
	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall
Content(Cn)	0.932	0.561	1	0.504	1	0.933	0.977 (0.043)	0.666 (0.198)
Form(Fm)	0.961	0.933	1	0.990	1	0.990	0.987 (0.030)	0.971 (0.038)
Functionality(Fn)	0.986	0.990	0.995	0.933	1	0.933	0.993 (0.016)	0.952 (0.043)
Cn&Fm	0.989	0.904	1	0.961	1	0.957	0.996 (0.013)	0.941 (0.059)
Cn&Fn	0.991	0.990	1	0.595	1	0.933	0.997 (0.011)	0.839 (0.187)
Fm&Fn	0.995	0.933	1	0.961	1	0.937	0.998 (0.008)	0.944 (0.048)
Cn&Fm&Fn	0.995	0.980	1	0.933	1	0.933	0.998 (0.008)	0.948 (0.046)
Overall Average (STD)	0.978 (0.036)	0.899 (0.147)	0.999 (0.005)	0.839 (0.195)	1 (0)	0.945 (0.043)	0.992 (0.023)	0.894 (0.088)

Table A4. Recall and precision results for News genre

Genre Attributes	Feature size						Average(STD)	
	5		20		100			
	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall
Content(Cn)	0.968	0.995	0.964	1	0.960	0.995	0.964 (0.034)	0.996 (0.012)
Form(Fm)	0.953	0.995	0.991	0.890	0.984	0.881	0.976 (0.042)	0.922 (0.077)
Functionality(Fn)	0.977	1	0.982	1	0.972	0.985	0.977 (0.028)	0.995 (0.014)
Cn&Fm	0.991	1	0.977	0.995	0.977	1	0.982 (0.025)	0.998 (0.008)
Cn&Fn	0.986	1	0.964	1	0.977	1	0.976 (0.027)	1 (0)
Fm&Fn	0.986	1	0.986	1	0.991	0.985	0.988 (0.020)	0.995 (0.014)
Cn&Fm&Fn	0.991	1	0.982	1	0.977	1	0.983 (0.024)	1 (0)
Overall Average (STD)	0.979 (0.032)	0.998 (0.008)	0.978 (0.026)	0.983 (0.047)	0.977 (0.030)	0.978 (0.049)	0.978 (0.030)	0.987 (0.018)

Table A5. Recall and precision results for PHP (Personal Home Page) genre

Genre Attributes	Feature size						Average (STD)	
	5		20		100			
	Precision	Recall	Precision	Recall	Precision	Recall	Precision	Recall
Content(Cn)	0.752	0.881	0.714	0.933	0.742	0.947	0.736 (0.060)	0.920 (0.060)
Form(Fm)	0.962	0.909	0.912	0.928	0.911	0.904	0.928 (0.045)	0.914 (0.060)
Functionality(Fn)	0.883	0.914	0.877	0.923	0.880	0.914	0.880 (0.044)	0.917 (0.061)
Cn&Fm	0.854	0.971	0.850	0.971	0.866	0.961	0.856 (0.048)	0.968 (0.028)
Cn&Fn	0.852	0.947	0.840	0.952	0.861	0.966	0.851 (0.042)	0.955 (0.041)
Fm&Fn	0.914	0.947	0.927	0.947	0.884	0.923	0.908 (0.037)	0.939 (0.051)
Cn&Fm&Fn	0.883	0.966	0.869	0.976	0.882	0.947	0.878 (0.040)	0.963 (0.034)
Overall Average (STD)	0.871 (0.072)	0.933 (0.056)	0.856 (0.081)	0.947 (0.046)	0.861 (0.064)	0.937 (0.056)	0.863 (0.073)	0.939 (0.048)