AUTOMATIC GENRE CLASSIFICATION OF HOME PAGES ON THE WEB

by
Alistair Kennedy

SUBMITTED IN PARTIAL FULFILLMENT OF THE
REQUIREMENTS FOR THE DEGREE OF
BACHELOR OF COMPUTER SCIENCE WITH HONOURS

AT

DALHOUSIE UNIVERSITY
HALIFAX, NOVA SCOTIA
APRIL 2004

DALHOUSIE UNIVERSITY

DEPARTMENT OF COMPUTER SCIENCE

The undersigned hereby certify that they have read and recommend to the Faculty of Computer Science for acceptance a thesis entitled "**Automatic Genre Classification of Home Pages on the Web**" by **Alistair Kennedy** in partial fulfillment of the requirements for the degree of **Bachelor of Computer Science with Honours**.

Dated: April 2004

Supervisor: _____
Dr. Michael Shepherd

Reader: _____
Dr. Raza Abidi

DALHOUSIE UNIVERSITY

Date: **April 2004**

Author: **Alistair Kennedy**

Title: **Automatic Genre Classification of Home Pages on the Web**

Department: **Computer Science**

Degree: **BCSc** Convocation: **May** Year: **2004**

Table of Contents

List of Tables

List of Figures

Abstract

The World Wide Web contains many web pages, consisting of a wide variety of genres. Finding a method to automatically determine the genre of these web pages could greatly improve the search results of search engines on the web. Knowledge about the genres of web pages could be used to allow a user to search for web pages within particular genres. A web page's genre could also be used for ranking the web page in cases where one genre is know to be more valuable to a particular query than another genre. One of the largest genres of web pages is the home page genre. The home page genre can be thought of as a hierarchy that includes the sub-genres, personal home page, corporate home page and organization home page. A web page's genre can be identified by extracting features from the page and then using those features to determine the web page's genre. In this research, feature sets are used to train and test a neural network for genre classification. The neural network will classify web pages as personal, corporate and organization home pages.

Acknowledgements

I would like to thank Dr. Shepherd for supervising me, as well as providing me with advice and resources for completing this thesis. Also, I would like to thank Dr. Duffy for assistance with hypothesis testing of my results, specifically using the two-sample t test.

Finally I want to thank Bin Tang, and Henry Stern for their assistance in choosing to work with neural networks and explaining some of the theory of how they work. Their assistance with using the Matlab Neural Network Toolbox was also valuable for this research.

Halifax, Nova Scotia Alistair Kennedy
April 5, 2004

Chapter 1

Introduction

1.1 Purpose

There are many genres of communication that currently exist on the World Wide Web. One of the more prominent genres on the web is the home page. Genres form a hierarchy of sub genres. For example home pages have the hierarchy

- Home Page
  - Personal
  - Corporate
  - Organization

Classifying home pages into genres has applications in Information Retrieval. Currently, search engines are able to classify and rank pages based on the topic of the web pages. If web pages could be classified into genres then the search engine could allow users to specify a genre in which to search[7]. Also search engines could rank pages according to genre, putting genres that are more relevant to a given topic ahead of other genres [2].

This thesis attempts to discover a method for automatically classifying home pages into "Personal", "Corporate", and "Organization" home pages. The approach is to automatically extract a set of features from the web page and then use a neural network to classify the web page into one of the three genres based on these features.

1.2 Where Genres Come From

There are many mediums containing many different genres. For example books have many genres, such as mystery, humor and biographical. Many of these genres have

been brought to the web; as well there are numerous genres that are unique to the web. The genre of a document can be thought of as being tied to the purpose of a document [4]. Knowing the purpose of a document could be very useful for information retrieval.

Unlike many other mediums there is no method to control the genres that exist on the web. The genre of a web page is left up to the designers of the web page. Outside of the community that uses a particular genre that genre may not be recognizable, and its purpose may be unknown. As such a method for identifying the genre of a document is needed [4].

In addition to genres simply being reproduced from other mediums or spontaneously created on the web there are numerous genres that have evolved from reproduced genres. Often a genre that is put on the web will go through changes such as adding new functionality to the genre. This can continue until it is an entirely different genre from the one that it originated as. A good example of this is the electronic newspaper genre, which has gone through numerous evolutionary stages and is becoming personalized news [22].

1.3 Applications of Genre Classification

We want to be able to automatically determine the genre of a given web page so that the genre information can be used by search engines to categorize the pages [2]. Ultimately genre could play a role in helping to improve search engine results to help provide more useful information.

Currently search engines classify documents based on key words found in the web pages. By indexing pages based on genres and allowing the user to indicate what types of genres they want to search for, the results of their search may be significantly improved [20]. For example if a user is doing research on an author's life then he/she could specify the search to look for biographies and to exclude fictional work by the author.

Another possible way that search engines can use genre is in the ranking process. If documents from one set of genres are known to be more useful than documents from another set of genres to particular search queries then the documents can be ranked in part according to their genres. For example, home pages may be less useful

for a particular search, as such they may be ranked lower than other pages when the search results are displayed [7].

Automatic information extraction can also be undertaken if you can identify the genre of a document. Different genres contain different information. For example, if a corporate home page is known to almost always have an e-mail address on it then a program can look for an e-mail address if it knows it has a corporate home page [19].

1.4 Methods Used

There are two steps taken in identifying the genre of a web page. The first step is feature extraction. In this step a program extracts information, such as linking structure and key terms, from a set of web pages and records their assigned genres. The second step is to take the feature information that has been extracted and enter it into a neural network. A feed forward neural network with no middle layers is trained and tested with the features extracted from the data set and the results are examined using precision and recall. The Levenberg-Marquardt algorithm for back propagation is used.

This thesis is divided into 5 chapters. The second chapter is the Literature Review where previous work is examined. The third chapter is the Methodology, which describes how this experiment is carried out. The fourth chapter is the Results. The fifth and last chapter is the Summary and Discussion, where the results are discussed.

Chapter 2

Literature Review

Genres of web pages have been the focus of some research in the past. There have been a number of papers identifying web genres as well as properties of these genres, which can be used to identify them. In addition to this, some research has been done on actually classifying them with various Machine Learning algorithms.

2.1 Defining Genre on the Web

Non-digital genres can be defined by the tuple <content, form>, where content is the themes and topics of the document, and the form is the physical features and linguistic features of the document [22]. Genres found in web pages, (also called Cybergenres), can be classified by the triple <content, form, functionality> [22]. The functionality attribute describes how a person interacts with the instance of the Cybergenre. This attribute becomes important when examining web pages that include games, search functions, or other functionality that takes advantage of the interactive nature of the web [23]. Different features from each element of the triple <content, form, functionality> can be used in identifying the Cybergenre of a given web page [22].

The question of how we identify a genre is posed by Crowston and Kwasnik [3]. How do we know the boundaries of a genre, and how do we know when we have crossed from one genre into another [3]? Existing genres can be identified in two ways, top-down and bottom-up. Top down gathers genre names and attempts to classify documents into these genres. Top down has the problem that different social groups create and name genres and so a genre that is recognizable to one community may not be recognizable to another [3]. The top-down method has been most widely used in the past. The bottom up approach involves finding out what characteristics make up genre and then finding how to apply those characteristics in order to define genres. This has the advantage that we can identify new emerging and dynamic

genres. One of the disadvantages of this method is that so far it has only been done on small scale studies and so there is not yet a set of data that has been used to find what genres people recognize [3].

2.2 Evolution of Web Genres

There are two kinds of genres that exist on the web, extant and novel. Extant genres are genres that exist in some other medium; novel genres are unique to the web. Novel genres could have evolved from extant genres or they could be completely original [22]. When a previously existing genre is introduced to the web it often evolves from its original form. Within the group of existent genres there are two sub groups, replicated and variant. Replicated genres are genres that appear as they are in their original medium. Variant genres are those genres that have some changes or added functionality but are still recognizable as their original genre. There are two subgroups of novel genres: emergent genres and spontaneous genres. Emergent genres are genres that have evolved from variant genres to the point where they are no longer recognizable as their original genres. Spontaneous genres are those genres that have spontaneously come to exist on the web [22]. An example of an evolving genre is electronic news, which was originally put on the web as a text version of the original newspaper. This genre has since adapted to take advantage of the web as a medium [10]. Examples of spontaneous genres include home pages and hot lists [22].

The newspaper genre is an example of a genre that has evolved from simply replicating the existing newspaper genre to becoming a new kind of genre that is exclusive to the web [9]. Its evolution started with simply placing the newspaper articles on a web page with some navigation links. The evolution of this genre includes increasing the number of columns on the web page from as few as 2-3 to as many as 5. In addition to this, the electronic newspaper was divided into sections [9]. The positioning of a news article on the web page was also identified as being important to the genre. This suggests that genres (specifically electronic newspapers) could be viewed as the quadruple <content, form, functionality, positioning> where positioning is the location on the page of the specific news article. Positioning is included because the often the location of a news article on the page can indicate it's genre [9].

2.3 Existing Genres on the Web

In Crowston and Williams [4] a set of randomly selected web pages is used to manually identify genres. Most genres, which were found by this experiment, were found to be extant genres. Far fewer were exclusive to the world wide web [4]. Genres were identified based on the purpose of the web page, rather than the actual form of the web page. Crowston and Williams argue that the purpose of the document is more important that the form and as such genres should be examined from this point of view. A total of 48 genres are listed by Crowston and Williams [4].

Georg Rehm has done some research into the genres that come from the Academics personal homepages[19]. A hierarchy of web pages is manually created and sub genres are identified. A total of 35 different genres were identified and put into a hierarchical form. Three other genres are identified, however they did not fit into the hierarchies. The hierarchy of most interest to this thesis is the hierarchy of homepages:

- Personal Homepage
  - Academics Personal Homepage
  - Student Assistants Personal Homepage
  - Virtual Business Card of Staff Member

Other hierarchies such as bibliographies, administrative information at universities and Institute/research units are also identified [19].

Another example of a genre on the web is the weblog. These weblogs or blogs first appeared in 1997 and since mid 1999 have become more and more popular on the web[8]. Blogs tend to be made using one of several programs, the most popular of which is Blogger. Some of the distinguishing characteristics of a blog is the presence of dates. Every entry has a date associated with it. Also they frequently contain titles for each entry. The time of an entry is also there for most posts as well as the authors name and internal links [8].

As was mentioned earlier, the newspaper is an example of a genre which exists on the web [9]. Shepherd and Watters identify six more web genres. These genres include Home Page, Brochure, Resource, Catalogue, Search Engine and Game and identify

different features for each member of the triple <content, form, functionality>, which can be used to distinguish one genre from another [23].

Lee and Myaeng identified a set of genres in their article, "Text Genre Classification with Genre-Revealing and Subject-Revealing Features" [16]. Large data sets of both English and Korean web pages were examined where the genres included; reportage, editorial, research article, review, homepage, Q&A and spec's [16].

2.4 Features of web documents

A web pages genre can be determined by examining certain features from the web page. These features could include things such as the number of links, images or simply the words found in the document or web site.

Crowston and Williams indicate how linking can be used to help identify the form of a web genre. The linking structure of a web page can be extracted by seeing where the links are going to. For example, links can be directed to locations within the same web page, or to other pages on the same server, or to pages on other servers [5].

> "...documents on the Web are sometimes composed of multiple Web pages, suggesting the need to consider how linking affects the documents form." [5]

Linking is examined in FAQ's. It was found that 7 FAQ's had no links, 19 were predominantly links on the same page, 33 were predominantly links within the same site and 11 predominantly had links to other sites. As such, the linking structure of a web page can help identify the form of the web page and can be used to help identify its genre.

Kessler et al. [15] uses digital documents from the Brown corpus to classify features for genre classification into 4 different categories. These features include structural features, which includes counts of the occurrences of syntactic categories such as passive or topicalized sentences. The second type are lexical features, these features include things like the existence of Latin symbols, or terms of address. The third kind of feature is character level features. These character level features mostly include punctuation marks in text. The last types of features are derivative features. These

features are derived from character level, and lexical features. These four feature sets can be broken down into two groups, structural features, which are more difficult to identify because they require a part-of-speech tagger. The second group is called surface features, which are much easier to determine than structural features since they do not require a part-of-speech tagger [15]. Ratios of these features can be used to identify genre. Surface features were tested using a neural network. The results showed that it was accurate for classifying genres anywhere between 67% and 100%.

Four more sets of features from digital documents, are identified and examined by Raüber and Muller-Kögler [18]. These features include:

- text statistics such as average length of words and sentences
- counts of punctuation and special characters
- counts of stop words and key words
- mark-up tag level features, such as the number of image tags on a page

These features were used with a Self Organizing Map to classify the genre of some documents, however the accuracy of the classifier was not discussed.

In Roussinov et al. [20] several web genres are identified by a user study. It was found that the majority of searches on the web are for academic purposes followed by shopping [20]. For each purpose for searching the web a set of genres was compiled and lists of features for each genres was found. It was proposed that the following features could identify genre:

- Format of the URL
- The number of graphics on the web page
- Length of the text
- Number of incoming links
- Is the page structured as a hierarchy
- Existence of phone numbers

- Education related lexicon
- Keywords such as FAQ and Q&A
- Questions followed by short answers

These features were used in manually identifying large groups of genres such as topics, publications, products, educational material and FAQ's. It is proposed that a larger feature set could be used with a machine learning algorithm for classification of web genres, however it is not attempted in this paper. [20].

Genres can also be identified using term frequency and inverse document frequency of terms as features. Lee and Myaeng propose that these measures can be used to identify both the subject and the genre of a given web page [16]. Tests for this were done on large data sets of both English and Korean web pages. From this experiment the inverse document frequency had better results for classifying than term frequency although the best results come when both were used. Document frequency went as high as 87% accuracy while term frequency went only as high as 80% accuracy [16].

In Georg Rehm [19], a series of features are discussed 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). Other features include a C.V., Course information, Research interests, lists of publications, lists of talks or presentations, links to university/department/research group and last updated information. These features were not tested with a machine learning algorithm [19].

Ingrid de Saint-Georges [6] has examined how language is used in personal web pages that can be used to distinguish them from other texts. Features can be extracted from the language in the page. For example personal web pages tend to contain a lot of personal pronouns, such as "I" and "you". Also, personal home pages tend to use spatial references, indicating where on the page certain items can be found, for example "Go back to top" or "click here". Other linguistic features of Personal home pages include establishing inter-personal relationships by directly addressing

the reader of the page, indicating that the page is a work in progress and indicating past/present and future activities of the author. This helps to illustrate the value that the text of the web page holds in identifying the class of a web page. These features were not tested in a machine learning algorithm.

Finn and Kushmerick [7] examine three feature sets extracted from a set of web pages. The first feature set is a Bag of Words in which all the words from the documents are entered into the classifier. The second feature set is a Part of Speech vector in which each documents is represented as a vector of ratios for 36 different parts of speech. The third feature set is text statistics, such as average sentence length and word length. In most tests either the Bag of Words or the Part of Speech method worked the best, however in most cases a combination of all three methods produced the best results [7].

The Bag of Words approach can be improved to create a feature set out of just the nouns in a web page. Nouns are the most important part of speech for classification in information retrieval [17]. Different groups of nouns have different effectiveness for classification. For example General Nouns are very good for classification with precision and recall of 0.899 and work well when combined with Personal names where it had 0.905 as precision and recall, however other types of nouns such as Pronouns, Personal names and Special Symbols are fairly poor classifiers when used on their own. Since the noun indicates the subject matter of a page, this means that subject can be important in understanding a web pages genre.

Karlgren and Cutting [14] identified some features in their research on web genres. Some of the features which were examined include adverb count, character count, long word count, preposition count, second person pronoun count, “therefore” count, words/sentence average, chars/sentence average, first person pronoun count, “me” count, sentence count, “I” count, characters per word average, “It” count, Noun count, present verb count, “That” count and “Which” count. These features were tested with data sets of 500 pages containing a hierarchy of sub-genres with different numbers of genres at each level. At the highest level there are two sub-genres. 478 of the 500 pages were classified correctly. At the next level with 4 sub-genres, 366 pages are classified correctly. At the lowest level there are 15 sub-genres. Only 242

were classified correctly, out of the 500 when there are 15 sub-genres.

2.5 Using Features with Machine Learning to Classify Genre

Once a set of features has been established it is beneficial to perform some sort of analysis on them in order to better decide which features to use. In Karlgren and Cutting's [14] article a set of features was examined using discriminate analysis. Discriminate analysis is a method of changing the feature set such that features that are of no value are removed and features which do not help the classifier are dropped and other features are given weights to indicate how valuable they are in classifying genre. This method was tested on data sets with varying numbers of genres. It was initially successful with relatively few genres (i.e., 2 or 4), however did not work as well with 10 or 15 genres [14].

In addition to selecting the best features it is also a good idea to normalize the feature so that they give the best results. One method of normalization is to normalize the mean and standard deviation of each element of the feature set [13].

There are many possible machine-learning algorithms for classification of genre on the web. One method that is commonly used is Naive Bayesian classifiers [16, 17]. A method that Lee and Myaeng compare to naive Bayesian is to use the cosine similarity function to compare vectors of features. First a vector of features is created for every genre that will be classified. Every document can then have a vector of features created for it and that vector is compared using the cosine similarity function to each genre vector to decide what genre it belongs to [16, 17]. These experiments found that the cosine similarity method was superior to the Naive Bayesian classifier method, which had a maximum success rate of 83% while the cosine similarity method had a maximum success rate of 87% on English texts. [16, 17].

Neural networks have also been used for classifying genres. A neural network was used for classifying genres of the Brown corpus by Kessler et al. [15]. For some feature sets the accuracy was as high as 90% or even 100% [15]. Neural networks seem to be the most successful of all the classifiers used.

Some research has been done into using Self Organizing Maps for genre classification of digital documents. Self organizing maps have been successful in other text

classification, however not much work has been done with them for classifying genre [18].

Another machine learning algorithm which has been used for both text classification and the classification of web genres is decision tree algorithms such as C4.5 [7, 13]. Finn and Kushmerick experiment used the C4.5 algorithm for genre classification of web pages with two different tests. One test is where the documents being classified are all of the same subject matter but of different genre. The second test is where they are of different subject matter and different genre. The results were that it was successful in classification when the documents were of the same subject matter. When classifying pages of the same subject matter into genres it was accurate to as much as 90%. When classifying pages of varying subject matter into genres it was only about 77% accurate [7].

Chapter 3

Methodology

In performing the research for this thesis there were several steps and techniques used in creating an automatic genre classifier. These steps are outlined in this chapter.

3.1 How the Genre Classification is Accomplished

There are several steps taken in classifying the web pages from the data set into genres. The first step is to take the data set and deciding whether to extract the feature set from Appendix 1 or Appendix 2. A program has been written to extract the selected feature set and save it into files. The features extracted are then normalized by applying the Matlab function "prestd". If the feature set from appendix 1 is uses then Principal Component analysis is used to preprocess the data set. 10-fold cross validation is then used to partition the data into training and testing data and then enter them into the neural network. The 10-fold cross validation is taken 10 times and the mean precision and recall as well as their standard deviations are taken. Figure 1 shows this process as a flow chart.

3.2 Choosing a Set of Genres

The World-Wide-Web contains many different web pages covering a wide range of subject matter and genres. For this thesis a subset of these genres must be manually selected from the web, which will be classified automatically. Previous researchers have identified many different genres that are common to the web. Some of these genres include Home pages, articles/publications, Reviews, News bulletins and FAQs to name a few[4, 20, 16].

The first step taken was to identify a set of genres that are common to the web. To accomplish this a random cross section of web pages was needed. There are

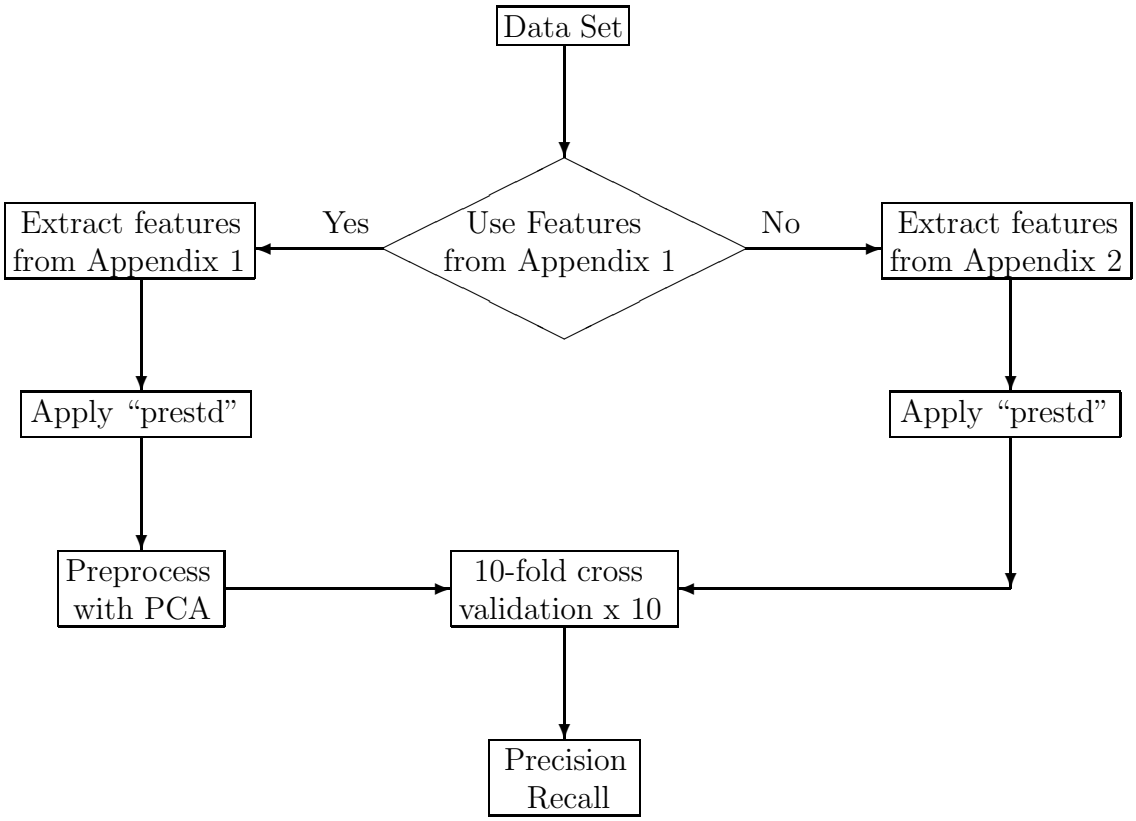

Figure 3.1: Steps taken to learn and classify the genres.

several resources that can be used to randomly select pages from the web. Two such resources are: http://www.uroulette.com and http://random.yahoo/fast/ryl/, both of which were used for this thesis. A set of 25 web pages was randomly selected and their genres were analyzed. The genres, sub genres and frequency of each sub genre are listed in Table 3.1.

Genre	Sub-Genre	# of Instances
Home Pages	Government Web Site	2
	Corporate Home Page	2
	University Home Page	1
	University Faculty Description	1
	School Home Page	1
	Union Home Page	1
	Celebrity Home Page	2
	Personal Home Page	1
	Fan Club	1
	Sports Team	1
Navigation	Dictionary/Table of contents	2
	Hot List	1
	News Hub	1
Information	Religion	3
	Production Review	2
	Brochure	3

Table 3.1: Genres, sub-genres and number of occurrences of each sub-genre from a randomly selected cross section of 25 web pages.

In the Table 3.1, 13 of the 25 web pages examined are home pages of some variety. Since this cross section was randomly selected this suggests that the home page is likely to be one of the largest genres of web pages. From the set of home pages most of the pages can be either classified as personal home pages, organization home pages or corporate home pages. Personal home page were defined to be home pages that contains information describing the interests and ambitions of a person, where those ambitions do not include making profit selling some product or service. Organization home page were defined to be home pages that contains 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 selling some product or service. A corporate home page was defined, as a web page describing the interests and ambitions of a company whose purpose for existing is to make profit through selling some product or service. Organization home pages basically filled the role of other home pages that did not fall into personal or corporate. It was decided that the classification of personal, organization and corporate home pages would be the focus of this research.

3.3 Features of Web Home Pages

The next step carried out was to identify a set of features that could be used to distinguish between corporate, organization and personal home pages. To accomplish this a larger set of home pages was gathered using http://random.yahoo/fast/ryl. When a web page was randomly selected it was decided if the page was in fact a home page or some other genre of web page. If it was a home page it was downloaded and classified as a personal, organization or corporate home page. The URL of the page as well as the genre of the page was recorded along with the pages source.

The newly created set of pages was examined for characteristics that seemed likely to distinguish corporate from personal home pages. This list of features contained statistics about the markup tags as well as lists of the most frequent terms. The complete list of features is as shown in Appendix 1.

3.4 Fuzzy Genre

Classifying a web page into just one genre can often be difficult. In many cases a web page may have qualities that make it seem to be a member of multiple genres. The author of this thesis hypothesizes that this is because the creator(s) of web pages have almost limitless control over the <content, form, functionality> of the web page, which decide the genre of the web page[22]. The web page creators control over the <content, form, functionality> of the web page is much greater than the control an author has over the appearance of a newspaper article or a journal article. As a result it is very easy to merge/change genres of web pages to create new genres. This means that some pages may in fact be instances of several genres. To solve this problem it

was decided that a classifier that is able to classify a single page into multiple possible genres should be used.

Neural networks were chosen as the classification technology for this thesis. A neural network can be used as a fuzzy classifier. In the neural network each possible class has an output node. For each output node, if the result is above some threshold the input can be said to be of that class. The neural network used in this thesis had outputs between 0 and 1. If the output was above 0.5 then it the input data was considered to be of that genre. In our case every output node would represent one genre and every web page would be classified as all genres for which the output value is greater than the threshold.

A set of 321 web pages was created. Initially this set was created using http://random.yahoo.com/fast/ryl/. It was discovered that a disproportionately high number of the pages found were corporate web pages, with some organization home pages and relatively few personal home pages. As a result many of the personal home pages were selected from Yahoo's directory of personal home pages found at http://dir.yahoo.com/Society_and_Culture/People/Personal_Home_Pages/. Pages with frames were not included in this study. The reason for this is that frames do not contain any content and since genre on the web is made up of <content, form, functionality>, frames alone cannot classify a genre. Also, only English web pages were used in this thesis. This is because terms were being used as features, which would not work well if the terms come from multiple languages.

Of the 321 web pages 244 were of the home pages genre and 77 were of other genres that are used as noise in this experiment. Of the 244 home pages collected 17 were manually classified as belonging to two of the three genres and none were classified as belonging to all three genres. With this in mind there were 94 corporate home pages, 93 personal home pages and 74 organization home pages.

3.5 Evaluating a Classifier

To decide if the classifier is good or not there must be a method of evaluating its results. One possible measurement is to measure the proportion of correctly classified documents.

Precision and recall are another method of measuring the quality of a classifier. Precision measures the proportion of retrieved and relevant documents $|Ra|$ in the set of retrieved documents $|A|$ [1]. For web genre classification precision is the proportion of web pages assigned to a genre classification that were of that genre.

$$Precision = \frac{|Ra|}{|A|}[1] \tag{3.1}$$

Recall measures the proportion of retrieved and relevant documents from the set of all relevant documents $|R|$ [1]. For web genre classification recall is the proportion of web pages of a genre that were properly classified.

$$Recall = \frac{|Ra|}{|R|}[1] \tag{3.2}$$

The classifier was rated using measurements for precision and recall.

3.6 Feature Set Selection

As the data set was being collected the web pages were examined to see what features could be used to distinguish one genre from the other. Different features were selected to represent the different elements of <content, form, functionality>. Using the terms from the web page as features could represent the content. The form would include features such as the number of images the size of the page and where the links on the page are linking to. Functionality features would include such things as forms and JavaScript.

It was noticed that home pages tend to link to pages on other web sites or on the same page, while corporate home pages tended to link to pages within their own site. These observations were turned into the features examining the proportion of pages linking to pages within the same site or to other sites, etc. Personal home pages were noticed to be longer and contain more words than corporate or organization home pages and so the number of terms on the page, and the size of the page in bytes were used as features. It was initially observed that corporate home pages tended to use more CSS and JavaScript than personal home pages. Also corporate home pages were more likely to have JavaScript or CSS written in another file, than in the HTML of

the page. These feature were later found to be less significant than the destination of the links, or the number of words. Corporate and organization home pages also were more likely to have contact information such as phone numbers or e-mail addresses and contained more meta tags. It was also noticed that corporate home pages were more likely than either personal or organization home pages to have a form on their page. Corporate and organization home pages were also more likely to have their own domain name than personal home pages. These observations came to create the feature set shown in Appendix 1.

The initial test used all the feature set from Appendix 1, except the list of most common words to classify it. Also, originally the number of navigation links, links to external sites and links within the page were used instead of the proportion of these links. The initial results were not very good. Several change to the features were tested. The three features: number of links to pages on external sites, pages within the site and links to locations within the same web page, were all changed to proportion of links to all three of these destinations. Another changes was to add the terms which appear in the most web pages, after removing a list of 27 stop words, as features. Some trial and error was done on selecting the terms that could be used to classify the web pages. It was found that terms that occurred in between 16% and 40% of all documents were the best for classification. Initially the terms were represented as Booleans indicating if that term exists in a given web page. This was later changed to the proportion of all terms in a web page that were that term.

The next step was to normalize the data. A function from matlab called "prestd" was used to normalize the data. "prestd" normalizes the data in such a way that the mean of every feature is zero and the standard deviation of every feature is one [12]. Once this was finished a single classifier was constructed that could classify the personal, organization, and corporate web pages.

The last step in creating a classifier for the chosen genres was to determine which features are best to train the neural network with. This can be difficult to do because certain features may be very good for distinguishing genre A from genre B but may not be helpful in distinguishing genre A from genre C. This became apparent later when removing some features improves results for some genres but made results worse

in other genres.

There are two separate ways that were used to reduce the size of the feature set. The first method is to sequentially remove some number of features from the feature set and examine how they affect the performance of the classifier. The second method is to employ Principle Component Analysis (PCA) to preprocess the data and remove features who's variance is below some threshold [11].

In the method of sequentially removing features, 5% of the features were removed at a time in order to determine which features were best. Often removing a feature caused some genres to be classified better while other genres got classified worse and so only features, who's removal, had no negative impact on any genre were removed. The following features were found to be mostly useless:

- 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?

The complete set of manually selected features can be found in Appendix 2.

It was found that terms that occur in more that 16% of all documents and in less that 40% of all documents make the best terms for classification. The exact numbers, 16% and 40% were found after trial and error. Since terms are being used to identify the content of the web pages it is useful that we remove terms that are too common, or too rare, as they do not help us to classify the web pages. Later it was decided that specific terms could be selected to help identify the genre of a document. To do this a script was written which analyzed the terms in the pages in the data set (excluding noise pages) to determine which terms would be good for identifying specific genres. A term is considered a good term for classification of genre X if $\geq 45\%$ of the web pages it occurs in are of genre X and if it appears in $\geq 22\%$ of

web pages of genre X. These numbers were decided upon after much trial and error. Terms which occurred in fewer than 22% of web pages of genre X were too few to be much use. If a term occurs in less that 45% of documents of genre X then it is probably too evenly distributed across the three genres and so it will not be a very good classifier. These terms were selected from the data set with personal, corporate and organization home pages.

When using the feature set from Appendix 1, the Matlab command "prepca" uses Principle Component Analysis to find and remove features with low variance. In some cases using "prepca" improved results, especially with Personal and Corporate home pages.

3.7 Training and Testing

The measurements found in equations 3.1 and 3.2 are applied to the results from the testing data on the neural network. There are several methods of training and testing a neural network. One such method is to select some large percentage of the data as training data and the rest as testing data. This will work well on large data sets where instances of each class will be equally distributed, however with small data sets it is much more likely that one class will be better represented than the other in the testing set. Since we are using a relatively small data set this method is prone to error. To avoid this problem it was decided that V-fold cross validation would be used, more specifically 10-fold cross validation was used. In V-fold cross validation, the data is divided into V different groups, so that each group contains proportionally the same number of instances of each class. The network is tested V times, for every iteration a different group from the V groups is chosen for testing and the other V-1 groups are used for training. The results are measured using equations 3.1 and 3.2. The advantage of this method is that it eliminates the possibility of the neural network being misrepresented by giving extremely good, or bad results by chance. The 10-fold cross validation is run 10 times and the mean and standard deviation of the results are taken.

For our test the data was divided into four sets, based on their genres: corporate, personal, organization home pages and non-home pages. To classify the 17 web pages

that were classified as having 2 genres, just one of their genres was selected and the web page was placed into that group. 10-fold cross validation was used, so from each of these groups, 10% was selected for testing and the rest was used for training. The training data was selected sequentially from the sets. The 10-fold cross validation was taken and the mean and standard deviation of the precision and recall of the three different genres was taken. To decide if the different results from the different tests are significantly different the t test is employed.

The neural network that was used contained an input node for every feature and three output nodes, one for each genre (personal, corporate and organization home pages). Trial and error was used to find the optimal size of the middle layer of the neural network. It was discovered that the best results were found using no middle layers. Different sizes of middle layers could be added to improve classification of one of the genres, but this would also reduce the effectiveness for classifying the other genres. Part of the reason for this may be to do with the size and amount of noise in the data set itself. Often in cases where the data set is relatively small, or where the data is noisy having a middle layer to the neural network will not help the results. There may only be enough good data to generalize a linear model and not enough to generalize a curved model [21]. The transition function used was log.

Chapter 4

Results

This section outlines the results from this experiment. To calculate these results the 10-fold cross validation was run 10 times and the mean and standard deviation are taken for the Precision and Recall for all the genres. Two different feature sets were tested. In one set (found in Appendix 1) every feature is used including all terms occurring in between 16% and 40% of all documents. Principal component analysis is used to reduce the feature set's size. The second features set is manually selected to insure the best results (found in Appendix 2). The data set consists of 321 home pages, 94 corporate, 93 personal, and 74 organization home pages. For one of the experiments a set of 77 web pages of noise were added. The noise pages were classified as non-home pages. These web pages could be any web page that is not a home page.

4.1 Manually Selected Features

When manually selecting features the best terms are selected for classifying web pages. Some terms are better than others for classifying documents of different genres. Table 4.1 shows the genres and which terms could be used to help positively identify these genres. A script was written to determine what terms could be used as features for classifying the genres. Terms were identified as being good for classifying a genre if they appeared in $\geq 22\%$ of all web pages of that genre and $\geq 45\%$ of all web pages in the data set (excluding noise pages) with that term are of that genre. No terms are selected for noise home pages. The terms are represented with Booleans to tell if they exist in the document. These terms are shown in table 4.1

The last term used for identifying personal home pages is the letter "t". Obviously this is not a word, however for this experiment all punctuation is ignored, and so contractions such as "don't" or "couldn't" were treated as two words, the second of the two words being "t". The fact that the letter "t" appears frequently on its own

in personal home pages may indicate high use of contractions. Also, corporate home pages have the term “amp”. This is most likely from “&”, which creates a “&” in html.

4.2 Three Genres

The classifier was tested with several different data sets. In this test web pages of Personal, Corporate and Organization pages were used, no noise pages were included in this experiment. The precision, recall and their standard deviations (STD) are recorded for each of the three genres.

4.2.1 Classifying with All Features and Principal Component Analysis

The results in this section show the precision and recall when the feature set from Appendix 1 is used and principal component analysis is used to preprocess the feature set. A threshold of 0.018 is used for “prepca”. The precision and recall and their standard deviations for all three genres can be found in table 4.2.

4.2.2 Manually Selecting Features

The results found from using the set of manually selected features, found in Appendix 2, for classifying web pages is shown in table 4.3. These results do not use any kind of Principal Component Analysis.

4.3 Just Personal and Corporate Pages

It was noticed that organization pages tend to give the worst results. It was hypothesized that many of the organization pages could be misclassified as either personal or corporate pages. Part of the reason for this is that organization home pages do not have a specific style that is unique to them, where as personal and corporate home pages each have a unique style. Organization home pages could look like either a personal or a corporate home page depending on who make the page. This section shows the results when classifying just the set of personal and corporate home pages are classified. No organization pages or noise pages are included. The precision, recall

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

Table 4.1: The terms selected to represent the genres.

and their standard deviations are recorded for the genres of personal and corporate pages.

4.3.1 Using all Features and Principal Component Analysis

Table 4.4 shows the results of the experiment when only personal and corporate home pages are examined. Organization home pages and noise are not included in this test. The feature set from Appendix 1 is used in this experiment and principal component analysis is employed to preprocess the feature set.

Class	Recall	STD of Recall	Precision	STD of Precision
Personal	0.71	0.016	0.79	0.011
Corporate	0.69	0.000	0.68	0.004
Organization	0.23	0.040	0.44	0.040

Table 4.2: Three Genres: Precision, recall and their standard deviations from the neural network when using the feature set from Appendix 1 and "prepca" with a threshold of 0.018.

Class	Recall	STD of Recall	Precision	STD of Precision
Personal	0.74	0.018	0.70	0.027
Corporate	0.59	0.012	0.69	0.018
Organization	0.60	0.030	0.61	0.050

Table 4.3: Three Genres: Precision, recall and their standard deviations from the neural network when using the feature set from Appendix 2.

4.3.2 Manually Selecting Features

In table 4.5 the results of the experiment are shown where the feature set from Appendix 2 is used and only personal and corporate home pages are examined. Organization home pages and noise are not included.

4.4 Adding Noise to the data set

In this section the classifier is tested with corporate, personal and organization home pages. Also, a set of 77 noise pages is added to the data set. These noise pages are made up of a range of genres that are not home pages. The results from this section do not show the precision, recall and standard deviations of the noise pages, but rather they demonstrate the effects that noise has on the precision and recall of the personal, corporate and organization home pages. Noise pages are considered to have no class.

Class	Recall	STD of Recall	Precision	STD of Precision
Personal	0.81	0.020	0.86	0.018
Corporate	0.84	0.017	0.85	0.020

Table 4.4: Two genres: Precision, recall and their standard deviations from the neural network when using feature set from Appendix 1 and "prepca" with a threshold of 0.018.

Class	Recall	STD of Recall	Precision	STD of Precision
Personal	0.77	0.039	0.79	0.028
Corporate	0.79	0.043	0.80	0.017

Table 4.5: Two Genres: Precision, recall and their standard deviations from the neural network when using the feature set from Appendix 2.

4.4.1 Noisy Data Using Principal Component Analysis

Table 4.6 shows the results of the experiment when non-home pages are added to the data set. This table shows the results when principal component analysis is used to preprocess the feature set from Appendix 1.

Class	Recall	STD of Recall	Precision	STD of Precision
Personal	0.66	0.012	0.77	0.003
Corporate	0.44	0.045	0.59	0.004
Organization	0.03	0.015	0.16	0.084

Table 4.6: Three Genres with Noise: Precision, recall and their standard deviations from the neural network when using feature set from Appendix 1 and "prepca" with a threshold of 0.018.

4.4.2 Noisy Data when Manually Selecting Features

Table 4.7 shows the results when non-home pages are added to the data set. The feature set from Appendix 2 is used in this experiment.

Class	Recall	STD of Recall	Precision	STD of Precision
Personal	0.63	0.037	0.63	0.022
Corporate	0.64	0.020	0.62	0.021
Organization	0.54	0.033	0.57	0.023

Table 4.7: Three Genres with Noise: Precision, recall and their standard deviations from the neural network when using the feature set from Appendix 2.

Chapter 5

Conclusion

5.1 Classifying the Three Genres

Using the set of features selected for this thesis, corporate and personal home pages were classified fairly well. The precision and recall for these two genres was always 68% or higher when no noise pages were added. Even when noise was added the precision and recall did not fall below 44%.

Organization home pages were not classified as well. When the feature set from Appendix 1 and Principal Component Analysis is used to reduce the feature set the recall is quite low, at 23%. These results are much worse when noise is introduced, with the precision and recall falling to 16% and 3% respectively. These values are improved by using the feature set from Appendix 2 where the features are manually extracted the precision and recall is always over 50%, even with noise added.

5.2 Principal Component Analysis vs Manually Assigned Features

This thesis demonstrates the advantages of using Principal Component Analysis (PCA) to preprocess the feature set to identify personal and corporate home pages. The two tailed t test was used for comparing results where we try to show that the absolute value of the test statistic t is greater than the critical value for t distributions ($|t| \geq t_{\alpha 2,\upsilon}$) where $\upsilon = 19$. It is shown that the precision of personal pages was significantly improved by preprocessing with PCA on the feature set from Appendix 1, over the manually selected feature set of Appendix 2, which is verified by a two tailed t test at $\alpha = 0.0005$. The precision of corporate pages was shown to be improved using the manual method of selecting features by the two tailed t test at a $\alpha = 0.1$. The recall of personal home pages, as well as the precision and recall of organization pages were found to be significantly improved using the Appendix 2

feature set verified by the two tailed t test at $\alpha = 0.0005$. The recall of corporate home pages was significantly improved using the feature set from Appendix 1 with PCA, verified by the two tailed t test at $\alpha = 0.0005$.

Preprocessing the feature set from Appendix 1 with PCA was shown to be significantly better for classifying corporate home pages in the experiment using only personal and corporate pages. It was also improved the results for personal home pages, however not as much. The recall of personal and corporate home pages was shown to be better by the two tailed t test at $\alpha = 0.005$. The precision of corporate and personal home pages was shown to be better by the two tailed t test at $\alpha = 0.0005$.

Using the Matlab "prepca" functions reduces the dimensionality of your input. As a result you can start with as many features as you want and then let the "prepca" process the feature set to decide what is input into the neural network. When using "prepca" the threshold for the variance of the features was 0.018. This method is shown to be better for classifying personal and corporate home pages, especially when no organization pages are used. Manually selecting features, as in Appendix 2, and not preprocessing with PCA gave better results for organization home pages. It also improved the personal recall and corporate precision by a small amount, 3% and 1% respectively.

5.3 Noisy Data

When noise was added the results for the precision and recall of every genre was reduced significantly when using the features from Appendix 1 with principal component analysis. This is not surprising, however the results for organization home pages were by far the worst. Precision and recall for all three genres were shown to be significantly worse by the two tailed t test with $\alpha = 0.0005$.

When using the features from Appendix 2 the results were not quite as bad. The precision and recall of personal and corporate home pages was significantly worse using the two tailed t test with $\alpha = 0.0005$. The precision of organization home pages was worse, verified by the two tailed t test with $\alpha = 0.025$, while the recall is improved without noise, verified by the two tailed t test at $\alpha = 0.0005$.

As expected noisy data does give worse results, however these results are not much worse when using the feature set from Appendix 2. In fact the precision and recall of every genre is $\geq 54\%$.

When using noisy data the recall and precision of corporate and organization home pages were significantly improved using the feature set from Appendix 2, verified by the two tailed t test with $\alpha = 0.0005$. Personal home pages are improved by using principal component analysis on the feature set from Appendix 1. The precision of personal home pages was improved using principal component analysis verified by the two tailed t test with $\alpha = 0.0005$, while the recall was improved and verified by the two tailed t test with $\alpha = 0.025$.

5.3.1 Problems with classifying Organization Home Pages

Organization home pages are much harder to classify than corporate or personal home pages. Organization home pages represent any home page that is not personal or corporate. As a result there are many different topics and themes that exist in these pages. For example, organization home pages could be home pages for religious, society, non-profit organizations and sports. These page contain a much more diverse set of themes than personal or corporate home pages would.

The fact that organization home pages were very diverse would have caused a lot of trouble when noisy data were added because the noise web pages came from a large range of topics and genes of web pages. As a result it is likely that the neural network could not tell the difference between the two and misclassified many of them as noise.

Another problem for organization home pages is that depending on the organization, an organization home page may look very closely like a personal home page, or a corporate home page. There is not much middle ground. Often personal home pages are made by a single person with little, skill or experience in web page design, where as corporate home pages tend to be much more slick and professional looking. Depending on the money that an organization has their home page could look like either of these.

5.4 Summary and Future Work

This thesis has shown the effectiveness of the proposed feature set for classifying different genres of home pages. The method of manually selecting features (Appendix 2) has shown the potential to be more resistant to noise from other genres. More research is needed to confirm this. Other future work could include attempting to divide the organization genre of web pages into several smaller genres and testing the effectiveness of this method of genre classification on it. Another topic to research is to test how well this method of genre classification works on genres of pages that are not home pages. For example this method could be applied to identifying hierarchies of information resources, such as electronic new articles and brochures.

Other future work could include using Latent Semantic Indexing to select terms to be used as features. Also identifying genres through a clustering technique rather than manually identifying genres could be an interesting avenue of research.

Bibliography

[1] Ricardo Baeza-Yates and Berthier Riveiro-Neto. *Modern Information Retrieval*. Addison Wesley, 1999.

[2] Ivan Bretan, Johan Dewe, Anders Hallberg, Niklas Wolkert, and Jussi Karlgren. Web-specific genre visualization. In *WebNet '98*, Orlando, Florida, 1998.

[3] Kevin Crowston and Barbara H. Kwasnik. A framework for creating a facetted classification for genres: Addressing issues of multidimensionality. In *Proceedings of the 37th Hawaii International Conference on System Sciences*, 2004.

[4] Kevin Crowston and Marie Williams. Reproduced and emergent genres of communication on the world-wide web. In *Proceedings of the 30th Hawaii International Conference on System Sciences*, page 30. IEEE Computer Society, 1997.

[5] Kevin Crowston and Marie Williams. The effects of linking on genres of web documents. In *Proceedings of the 32nd Hawaii International Conference on System Sciences*, 1999.

[6] Ingrid de Saint-Georges. Click here if you want to know who i am. deixis in personal homepages. *Thirty-First Annual Hawaii International Conference on System Sciences*, VOL. II, 1998.

[7] Aidan Finn and Nicholas Kushmerick. Learning to classify documents according to genre. In *IJCAI-03 Workshop on Computational Approaches to Style Analysis and Synthesis*, 2003.

[8] Susan C. Herring, Lois Ann Scheidt, Sabrina Bonus, and Elijah Wright. Recognizing text genres with simple metrics using discriminant analysis. In *Proceedings of the Thirty-Seventh Annual Hawaii International Conference on System Sciences-Volume 2*, 2004.

[9] Carina Ihlstrom and Marim Akesson. Genre characteristics - a front page analysis of 85 swedish online newspapers. In *Proceedings of the Thirty-Seventh Annual Hawaii International Conference on System Sciences-Volume 2*, 2004.

[10] Carina Ihlstrom and Lars Bo Eriksen. Evolution of the web news genre -the slow move beyond the print metaphor. *Proceedings of the 33rd Hawaii International Conference on System Sciences*, 2000.

[11] The MathWorks Inc. help prepca. *MATLAB Version 6.5.1*, 2003.

[12] The MathWorks Inc. help prestd. *MATLAB Version 6.5.1*, 2003.

[13] Jussi Karlgren, Ivan Bretan, Johan Dewe, Anders Hallberg, and Niklas Wolkert. Iterative information retrieval using fast clustering and usage-specic genres. In *In Proc Eighth DELOS Workshop on User Interfaces in Digital Libraries*, pages 85–92, Stockholm, Sweden, 1998.

[14] Jussi Karlgren and Douglass Cutting. Recognizing text genres with simple metrics using discriminant analysis. In *Proceedings of the 15th. International Conference on Computational Linguistics (*Coling 94*)*, volume II, pages 1071 – 1075, Kyoto, Japan, 1994.

[15] Brett Kessler, Geoffrey Nunberg, and Hinrich Schütze. Automatic detection of text genre. In Philip R. Cohen and Wolfgang Wahlster, editors, *Proceedings 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*, pages 32–38, Somerset, New Jersey, 1997. Association for Computational Linguistics.

[16] Yong-Bae Lee and Sung Hyon Myaeng. Text genre classification with genre-revealing and subject-revealing features. In *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 145–150. ACM Press, 2002.

[17] Yong-Bae Lee and Sung Hyon Myaeng. Automatic identification of text genres and their roles in subject-based categorization. In *Proceedings of the Thirty-Seventh Annual Hawaii International Conference on System Sciences-Volume 2*, 2004.

[18] Andreas Rauber and Alexander Múller-Kógler. Integrating automatic genre analysis into digital libraries. In *ACM/IEEE Joint Conference on Digital Libraries*, pages 1–10, 2001.

[19] Georg Rehm. Towards automatic web genre identification. In *Proceedings of the 35th Annual Hawaii International Conference on System Sciences (HICSS'02)-Volume 4*, page 101. IEEE Computer Society, 2002.

[20] Dmitri Roussinov, Kevin Crowston, Mike Nilan, Barbara Kwasnik, Jin Cai, and Xiaoyong Liu. Genre based navigation on the web. *Proceedings of the 33rd Hawaii International Conference on System Sciences*, 2001.

[21] Warren Sarle. How many hidden layers should i use? http://www.faqs.org/faqs/ai-faq/neural-nets/part3/section-9.html, 2004.

[22] Michael Shepherd and Carolyn Watters. Evolution of cybergenre. *Proceedings of the Thirty-First Annual Hawaii International Conference on Systems Sciences.*, 1998.

[23] Michael Shepherd and Carolyn Watters. The functionality attribute of cybergenres. In *Proceedings of the Thirty-Second Annual Hawaii International Conference on System Sciences-Volume 2*, page 2007. IEEE Computer Society, 1999.

Appendix A

List of All Features

- 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.
- 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?
- Number of Meta tags used.
- Is JavaScript included from an external file?
- Is JavaScript written into the HTML?
- Does the page contain any phone numbers?
- Does the page have its own domain, or is it in a sub-directory within a domain?
- Are there any forms?
- Is the first tag a Script tag?
- Number of form inputs

- Size of file in bytes.
- Number of words in the page.
- List of most common words appearing in between 16% and 40% of all documents.

Appendix B

List of Manually Selected Features

- 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 to locations within the same page.
- Proportion of links to other pages on other sites.
- Number of images.
- Number of Meta tags used.
- Does the page contain any phone numbers?
- Does the page have its own domain, or is it in a sub-directory within a domain?
- Is the first tag a Script tag?
- Number of form inputs.
- Size of file in bytes.
- Number of words in the page.
- List of terms found in table 4.1.