AN N-GRAM BASED APPROACH TO THE AUTOMATIC CLASSIFICATION OF WEB PAGES BY GENRE

by

Jane E. Mason

Submitted in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

at

Dalhousie University
Halifax, Nova Scotia
December 2009

DALHOUSIE UNIVERSITY

FACULTY OF COMPUTER SCIENCE

The undersigned hereby certify that they have read and recommend to the Faculty of Computer Science for acceptance a thesis entitled "AN N-GRAM BASED APPROACH TO THE AUTOMATIC CLASSIFICATION OF WEB PAGES BY GENRE" by Jane E. Mason in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

Dated: December 10, 2009

External Examiner: ____________________
Stan Matwin

Research Supervisors: ____________________
Michael Shepherd

____
Evangelos E. Milios

Examining Committee: ____________________

____

DALHOUSIE UNIVERSITY

DATE: December 10, 2009

AUTHOR: Jane E. Mason

TITLE: AN *N*-GRAM BASED APPROACH TO THE AUTOMATIC CLASSIFICATION OF WEB PAGES BY GENRE

DEPARTMENT OR SCHOOL: Faculty of Computer Science

DEGREE: PhD CONVOCATION: May YEAR: 2010

Signature of Author

Table of Contents

List of Tables

List of Figures

Abstract

The extraordinary growth in both the size and popularity of the World Wide Web has generated a growing interest in the identification of Web page genres, and in the use of these genres to classify Web pages. Web page genre classification is a potentially powerful tool for filtering the results of online searches. Although most information retrieval searches are topic-based, users are typically looking for a specific type of information with regard to a particular query, and genre can provide a complementary dimension along which to categorize Web pages. Web page genre classification could also aid in the automated summarization and indexing of Web pages, and in improving the automatic extraction of metadata.

The hypothesis of this thesis is that a byte n-gram representation of a Web page can be used effectively to classify the Web page by its genre(s). The goal of this thesis was to develop an approach to the problem of Web page genre classification that is effective not only on balanced, single-label corpora, but also on unbalanced and multi-label corpora, which better represent a real world environment. This thesis research develops n-gram representations for Web pages and Web page genres, and based on these representations, a new approach to the classification of Web pages by genre is developed.

The research includes an exhaustive examination of the questions associated with developing the new classification model, including the length, number, and type of the n-grams with which each Web page and Web page genre is represented, the method of computing the distance (dissimilarity) between two n-gram representations, and the feature selection method with which to choose these n-grams. The effect of preprocessing the data is also studied. Techniques for setting genre thresholds in order to allow a Web page to belong to more than one genre, or to no genre at all are also investigated, and a comparison of the classification performance of the new classification model with that of the popular support vector machine approach is made. Experiments are also conducted on highly unbalanced corpora, both with and without the inclusion of noise Web pages.

Acknowledgements

The completion of a PhD thesis is not a journey to be undertaken alone, and I have been very fortunate to have had the best of support teams.

Words could never be enough to express my gratitude to my husband John, my daughter Michelle, and my son Gilles, for their unwavering patience, support, and encouragement. Thank you for believing in me, and for being with me every step of the way. Thank you, John, for the seemingly endless stream of presents and flowers that brightened the road. Life with you is better than I could ever have dreamed. Thank you, Michelle, for always being there to cheer me on, for helping me with rehearsing my presentations and proofreading my thesis, and for making sure that I remembered to take breaks and bake banana bread. You have kept me motivated and inspired. Thank you, Gilles, for all the hugs and phone calls, often when most needed and least expected. You have made me feel very special.

I would also like to thank my supervisors, Dr. Michael Shepherd and Dr. Evangelos Milios, for their support, guidance, encouragement, and understanding. My appreciation also goes to Dr. Jack Duffy and Dr. Vlado Kešelj for their helpful suggestions, expertise, and assistance. I would also like to thank my external examiner, Dr. Stan Matwin, for his insightful comments.

I have received funding during the course of my PhD from the Killam Trusts, the Natural Sciences and Engineering Research Council of Canada (NSERC), the Eliza Ritchie Doctoral Scholarship, and the Walter C. Sumner Foundation. I am very grateful for this support, without which this journey would not have been possible.

It is good to have an end to journey toward;
but it is the journey that matters, in the end.

–Ursula K. Le Guin

Chapter 1

Introduction

Genre has been said to be “the most powerful explanatory tool available to the literary critic” [97]. Genres are useful because they make communications more easily recognizable and understandable, with the most important role of genre being that of differentiating between documents [143] . Recent research has shown genre to be a potentially powerful tool for filtering the results of online searches. Although most information retrieval searches are topic-based, users are typically looking for a specific type of information with regard to a particular query, and genre can provide a complementary dimension along which to categorize documents.

Genre helps orient users as to the nature of a document, providing them with a means of interpreting the content, role, and function of the document [143] . This is of particular value when the documents are delivered through an undifferentiated digital medium such as the World Wide Web. Identifying the genre of a document provides information as to the document’s purpose and how well it fits the user’s information need; the relevant documents for two searches with the same topic and query keywords could be quite different, depending on the need of the user [31, 32]. For example, a medical doctor looking for information about lung cancer would likely be interested in research papers and clinical studies, whereas a patient with the disease could be more interested in resources and support groups. The relevance of a particular document to a given query depends on the information need of the user who issues the query, and information retrieval systems could be enhanced by providing users with the ability to filter documents according to their genre [43]. Categorizing Web pages by genre could allow a user to specify the genre which is of interest with regard to a particular query. Although this filtering could be provided by search engines, the genre classification could also be performed on the user end.

The exponential growth of the World Wide Web has resulted in both researchers and search engine companies looking for methods to improve the quality of search results. As early as 1998, researchers proposed using a richer representation of retrieval results in Web search interfaces, which would include Web genres. Based on a user study with 102 participants, Bretan et al. [20] developed a set of a dozen genres to represent the types of material that study participants reported finding and interacting with online. Bretan et al. suggested combining content-based clustering and genre-based classification to provide a two-dimensional front-end to online search engines, grouping documents along the dimensions of content and genre. Stubbe et al. [127] also made suggestions for a specialized search engine interface that allows a user to select a genre with which to filter the search results by either returning only Web pages from a particular genre, or by excluding a particular genre of Web pages from the results. Stubbe et al. also introduced feedback events, such as lingering time, that could be used to provide implicit feedback to help improve the genre search. In experiments performed by Dewdney et al. [37] the inclusion of genre information as part of a query resulted in a significant improvement in the precision of the documents retrieved, with only a small reduction in the recall.

A user study on the usefulness of Web genre classification found that 64% of the participants in the study thought that Web genre classification would be very useful, and that another 29% of the participants thought that it would sometimes be useful [84]. However, Web page genre classification is very difficult, partly because the growth in the World Wide Web has been matched by a growth in the number of new and evolving genres on the Web [118] .

In one Web genre user study with 184 participants, the participants were asked to report the purpose or task that they were performing when viewing a Web page. Out of 1234 Web pages, the researchers identified 116 different genres. However, when the participants were asked to assign labels from this 116 genre set to the Web pages they had viewed, only 1076 Web pages were successfully given genre labels, and of these, fewer than 50% matched the labels given to the Web pages by the researchers [100]. More recently, a group of seven Web genre researchers performed an experiment in which they each independently assigned genre labels to 50 randomly chosen Web pages. The resulting labels were so disparate that the group found it impossible to do any statistical analysis, or to measure inter-rater agreement. They concluded that, surprisingly, even researchers with expertise in Web genres cannot come up with similar genre labels for Web pages; assigning genre labels to Web pages is not an easy task [93]. This supports the conclusions of Kwasnik and Crowston [68] who found that it is often very difficult for people to identify genres by name, and that there are many nearly equivalent terms for the same genre.

Identifying a genre taxonomy is a subjective process, therefore there is likely to be disagreement about what constitutes a genre, or what documents belong to a particular genre [44]. However, it is likely that many search tasks could be satisfied with documents from a few groups of related genres, in which case distinguishing among documents in these few groups of genres could be almost as effective as being able to provide correct genre labels for every Web page [100]. For example, Sharoff [114] proposed a compact system of genres which could be used to describe the majority of texts on the Web, however Sharoff approached the problem from a functional viewpoint, suggesting a typology of genres which classify Web pages based on the function they fulfill. The suggested genres were DISCUSSION, INFORMATION, INSTRUCTION, PROPAGANDA, RECREATION, REGULATIONS, and REPORTING.

Although the most obvious use of Web page genre classification is as a method of filtering the results of online searches, there are many other potential applications of this work. Genre classification could aid in the automated summarization and indexing of documents; for example, documents from the NEWSPAPER ARTICLE genre can be effectively summarized by the first few sentences of the document, whereas such an approach would not work well for a document from the HOME PAGE or FAQ genres [32]. Web page genre classification could also be useful in the construction of catalogues that organize Web documents into hierarchical collections [115], in increasing the parsing accuracy in the field of natural language processing, in building more diversified and balanced corpora for the study of corpus linguistics, and in improving the automatic extraction of metadata in the field of information science [93].

Two prototype applications that incorporate Web page genre classification include X-Site [145] and WEGA [7]. X-Site is a prototype for an enterprise search system that incorporates the relationship between user's tasks and document genres to improve retrieval precision [145]. The system was developed for the software engineering domain, after a strong relationship was found to exist between the tasks performed by software engineers and the document genres they used. This genre-preference based on task was incorporated into the ranking algorithm of the X-Site system; when searching with the system, the user enters a query and then selects a work task from a drop-down menu. X-Site incorporates genre classification on the server-side; WEGA, (**WE**b **G**enre **A**nalysis), on the other hand, is a genre add-on for the Mozilla Firefox search engine, and as such, incorporates genre classification on the client-side. It provides genre labels for each list of search results returned by the search engine. Although still under development by a research team led by Benno Stein at the Bauhaus University, WEGA is currently available for download [7].

1.1 Research Objectives

The extraordinary growth in both the size and popularity of the World Wide Web has generated a growing interest in the identification of Web page genres, and in the use of these genres to classify Web pages. Web page genre classification is a potentially powerful tool for filtering the results of online searches, however because of the dynamic nature of the Web, the Web page genres are somewhat transient in nature; their numbers grow, their members evolve and merge, and the names by which they are recognized vary, making it difficult to determine the genre of a particular Web page. The study of Web page genres can be roughly broken down into two areas: investigating the taxonomy of Web genres, and classifying Web pages by their genre(s). This thesis focuses on the latter problem.

In general, classification is the task of automatically determining the class of an unseen object. Machine learning techniques have been widely used to automatically classify documents according to topic; typically, a machine learning algorithm is provided with a training set of documents, each labeled as belonging to one of a number of known topics, and based on these examples it learns to distinguish between documents of the given topics. Cavnar and Trenkle [22] refer to this process as "categorization by example," noting that a classifier must also be able to recognize when a particular document is not a member of any of the categories. In this thesis, machine learning techniques are used to classify Web pages according to their genre. This research does not investigate the practical application of the genre classification, but focuses on the improvement of Web page genre classification techniques.

The hypothesis of this thesis is that a byte n-gram representation of a Web page can be used effectively to classify the Web page by its genre(s). The goal of this thesis is to develop an approach to the problem of Web page genre classification that is effective not only on balanced, single-label corpora, but also on unbalanced and multi-label corpora, which better represent a real world environment. This thesis research develops n-gram representations for Web pages and Web page genres, and based on these representations, a new approach to the classification of Web pages by genre is developed. Although there are many different types of n-grams, the main focus of this work is on the use of byte and character n-grams, therefore unless a specific distinction is made, the use of the term *n-gram* refers to byte and character n-grams.

The general research problem is that of identifying Web pages of a specific genre out of sets of Web pages of a variety of genres. The proposed approach is based on calculating and comparing profiles of n-grams. Given a training set of genre-labeled Web pages, n-gram profiles are computed that represent each genre, and then n-gram profiles are computed for each Web page that is to be classified. A distance measure is used to compute the similarity between the Web page profiles and each of the genre profiles, and each Web page is labeled as belonging to the genre(s) to which this distance is smallest.

Specifically, the investigation of whether there is a simple, easily scalable method for Web page classification is addressed by the following research questions, in the context of a set of controlled experiments.

1. How should Web pages and Web page genres be represented?

   (a) Do Web pages of the same genre share a distribution of n-grams that is similar enough to allow n-gram representations of the Web pages to be used for Web page genre classification?

   (b) Does there exist a set of parameters for the proposed approach, with regard to n-gram type, length, and number, that will allow state-of-the-art classification of Web pages by their genres?

   (c) Can a threshold value be determined for each genre in order to allow a Web page to belong to more than one genre, or to no genre at all?

2. What classification method should be used?

    (a) Can an appropriate distance measure be found that computes the similarity between Web page and Web page genre profiles, leading to a classification method based on the nearest profile?

    (b) Does the proposed classification model give comparable performance to a support vector machine approach, using the same n-gram representations of the Web pages?

    (c) Is the proposed approach effective on highly unbalanced corpora?

    (d) Given noise Web pages that do not belong to any genre in a particular corpus, is the proposed approach effective when these noise pages are included in the corpus?

The studies investigating the feasibility of the use of the proposed n-gram model, to be discussed in Chapter 4, indicated that a simple byte n-gram approach to Web page representation for genre classification was a viable approach. The classification results were in the same range as those of other researchers on the same corpus, leading to the conclusion that Web pages of the same genre do share a distribution of n-grams that is similar enough to allow n-gram representations of the Web pages and Web page genres to be used for the task of Web page genre classification. An investigation of twelve distance measures found that there does exist a distance measure that allows the computation of the similarity between the Web page and Web page genre profiles, allowing the use of a classification method based on the nearest profile in which the Web pages in the training set are used to form a centroid n-gram profile for each genre. The development of this centroid classification model for Web page genre classification provides a classification model that is easily scalable; adding another genre to the model requires only the creation of a centroid from the training data for the new genre. Experiments indicated that the use of this model with an appropriate distance measure, in combination with appropriate parameters for n-gram type, length, and number, achieves state-of-the-art classification performance on a variety of corpora, including those that are highly unbalanced. Experiments with techniques for setting genre thresholds showed that a threshold value can be determined for each genre in order to allow a Web page to belong to more than one genre, or to no genre at all. The use of genre threshold values also allows the centroid classification model to effectively recognize noise Web pages that do not belong to any genre in a particular corpus. The new centroid classification model achieves classification performance comparable to that of a support vector machine, using a much less complex and more easily scalable approach.

1.2 Overview

This chapter has presented the motivation, hypothesis, and research questions for this thesis. Chapter 2 will give an overview of existing research relating to Web page genres and genre classification. Chapter 3 provides relevant background information, such as some of the limitations of existing Web page corpora, the characteristics of the Web page corpora chosen for the experiments to be discussed in this thesis, details about the cross-validation procedure and evaluation metrics, the statistical tests and terms to be used in the statistical analysis of the experimental results, and descriptions of the χ^2 statistic and Information Gain feature selection techniques. Chapter 4 discusses studies investigating the feasibility of the use of the proposed n-gram model, which helped establish the focus of the remainder of this thesis in terms of the open questions to be addressed and the tools with which to work. Chapter 5 discusses experiments which explore the effect of data preprocessing, as well as experiments which explore the use of the χ^2 statistic and Information Gain as feature selection measures. Chapter 6 investigates the use of thresholds in order to allow the multi-labeling of Web pages, and presents experiments conducted on highly unbalanced corpora. Finally, Chapter 7 gives a summary of the research contributions of this thesis, discusses the limitations of the research, and suggests directions for future work.

According to Swales [130], genres “not only provide maps of new territories, but also provide the means for their exploration.” The overall goal of this thesis is to provide insight into the development of a classification tool to aid in this exploration.

Chapter 2

Literature Review

This chapter gives an overview of existing research relating to Web page genres and genre classification. Section 2.1 explores the question of how the concept of genre is defined, while Section 2.2 discusses the emergence and evolution of Web page genres and reviews some of the early work in this area. Section 2.3 gives an overview of research on text and Web page genre classification, with a focus on the document and Web page representations used in genre classification. Section 2.4 discusses methods to minimize the size of the feature sets used to represent documents and Web pages in classification problems, while Section 2.5 looks at some of the real world challenges to consider during the design and testing of a Web page genre classifier. The chapter concludes with a brief summary, given in Section 2.6.

2.1 What is Genre?

The word *genre* comes from the Greek word *genus*, meaning *kind* or *sort* [97], however there is no definitive agreement among researchers as to what is meant by the term genre [43]. Finn and Kushmerick [43] describe genre as a vague concept that lacks clear boundaries, and note that for a given subject area, there is no fixed set of genre categories. Kwasnik and Crowston [67] observe that although genres are widely recognized and widely used, they are not so readily defined.

Working with text documents, Orlikowski and Yates [88] consider genre to be a distinctive type of communicative action, characterized by a socially recognized communicative purpose and common aspects of form. Kessler [60] provides a similar view of genre as any widely recognized class of texts defined by some common communicative purpose or other functional traits, while Finn et al. [44] describe genre as an abstraction based on a natural grouping of documents written in a similar style. Working with Web pages, Santini [106, 108] sees Web genres as communication artifacts that are linked to a society or community and characterized by particular conventions; the genres are assumed to raise certain expectations, while possibly exhibiting hybridism or individualization, and undergoing evolution. Kwasnik and Crowston [67] observe that most definitions of genre include the notions of document form, expected content, and intended communicative purpose, as well as the notion of social acceptance: a genre is only useful to the extent that it is known and recognized by users.

Subtle, vague, abstract, complex and *fuzzy* are all terms that have been used to describe the concept of genre, and yet despite this amorphous nature, genre facilitates the recognition of items that are similar, even in the midst of great diversity [116, 117]. Finn and Kushmerick [43] explain that genres depend on context, and therefore in practice, the usefulness of a particular genre class depends on how well it helps a user distinguish documents. In the context of the Web, where most searches are topic-based, useful genre classes are those that allow a user to distinguish between documents that have similar topics. Thus, a genre is useful in this context because it provides information about the type, rather than the topic, of the document. Although document genre may not be entirely orthogonal to document topic, genre provides a dimension with which to classify documents and Web pages that is complementary to that of topic. However, one of the challenges in genre classification is to identify a group of genres, or a genre taxonomy, that will be helpful to the user; this is a subjective process complicated by the fact that there is disagreement both about what constitutes a genre as well as about what criteria establishes membership in a particular genre [44]. Despite the lack of an agreed upon precise definition of genre, the results of a user study on Web genre usefulness showed that 64% of participants thought that genre classification was very useful, and that another 29% of participants thought that it would be sometimes useful [84]. However, further complicating the study of genre is the fact that genres are in a constant state of evolution. In the case of Web page genres, the dynamic nature of the Web has led to less stable genres compared to those in print technology [143].

This section has provided a glimpse of the fuzzy character and varying definitions of genre in general and Web page genres in particular, while noting that despite its rather vague nature, genre can be an important factor in categorizing texts or Web pages. Although the discussion about the concept of genre is interesting and useful, the focus of this thesis, and of Web page genre classification in general, is on the potential use of genre as a dimension, complementary to topic, with which to classify Web pages.

Section 2.2 looks at the emergence and evolution of genres on the Web, and reviews some of the early work in this area, while Section 2.3 explores the characteristics of Web page genres, and how Web pages can be represented during the classification process.

2.2 Web Genre Identification, Emergence, and Evolution

One of the areas of interest in Web page genre research is in the genres themselves. Because the Web is such a complex and dynamic medium, Web page genres are in a constant state of evolution. Researchers study not only the ways in which Web page genres emerge, evolve, multiply, or even disappear, but also the ways in which Web page genres could prove to be useful. However, Shepherd and Watters [118] suggest that because of the rapid speed at which the Web expands and evolves, the task of identifying Web page genres is akin to that of hitting a moving target.

Focusing on Web page genres, Roberts [95] demonstrated, through the use of a narrative analysis, that personal homepages represent a distinct genre, where genre is defined as texts that have a similar set of purposes, mode of transmission, and discourse properties. Roberts explained that given this definition of genre, Web page genres are best distinguished from one another by their discourse properties, and that narrative is the discourse property that separates personal homepages from other Web pages.

Crowston and Williams [33] conducted a survey of 100 English Web pages from educational, commercial, and government sites, from a total of 12 countries. The goal of the study was to identify the range of communicative genres that were in use on the Web. Crowston and Williams defined communicative genres as accepted types of communication sharing common form, content or purpose. They were surprised by the range of Web genres observed in the study, with 48 different genres being identified, at a very fine level of granularity. Approximately 80% of the Web pages surveyed belonged to genres or combinations of genres that were reproduced from other media, however about 11% appeared to be novel genres, while the other 9% were unknown genres, either because their purpose was unclear, or because they had mixed features. Crowston and Williams found that analyzing Web usage through genres was very effective, and encouraged designers of Websites to take users' expectations about familiar genres into consideration.

Shepherd and Watters [116] also examined the emergence of what they termed cybergenres. They noted that some genres seem to emerge spontaneously, such as HOTLISTS and HOMEPAGES, while others evolve from existing genres. Shepherd and Watters explain that genres evolve in response to various institutional changes and social pressures, and that in some cases, the changes to a genre are so extensive that it gradually emerges as a new genre. This agrees with the observation of Yates and Orlikowski [144] that communications in a new media will show the reproduction or adaptation of existing communicative genres, as well as the emergence of new genres. This observation is also supported by Crowston and Williams [33, 34], who emphasize that the very definition of genre is based on social acceptance, and therefore at some point, an emerging genre becomes generally recognized and named as a separate genre. They note, for example, that the FAQ (Frequently Asked Questions) has emerged as a distinct genre.

Eriksen and Ihlström [42] examined the evolution of the WEB NEWS genre, based on a longitudinal study of three Scandinavian news sites from 1996 to 1999. They concluded that by 1999, the WEB NEWS genre possessed characteristics that made it distinctive both from the print version of the NEWS genre, and from the 1996 WEB NEWS genre. Eriksen and Ihlström note that for Website designers, genre awareness is a tool with which to target audiences. For Web users, genre awareness is a tool with which to reduce the complexity of the World Wide Web by allowing the recognition of sites as belonging to distinctive genres; genre awareness frees users from the need to learn and recognize the purpose of each and every site visited.

Boese and Howe [17] examined the effects of Web page evolution on the task of classifying Web pages by genre. They hypothesized that knowing the genre of a Web page may help determine whether, and how much, the page is likely to change over time. Boese and Howe ran experiments using two versions of two different corpora; in each case, they tested with a 1999 and a 2005 version of the corpora. They concluded that although the World Wide Web changes very quickly, some Web page genres, at least during the six year period in question, remain quite stable. They also concluded that classifiers trained on old Web page corpora still work well on updated versions of the corpora.

Santini [103] explored the state of genre evolution through the users' perception of Web page genres, based on a study with 135 participants from an academic population. Santini found that the users' perception of Web page genres could be divided into three ranges, which could be interpreted in terms of genre evolution. The first range was that of high perception, which was associated with stable and acknowledged genres. The next range was that of medium perception, associated with emerging genres, while the final range of low perception was associated with highly ambiguous genres. Santini concluded that some of the more recent Web page genres, such as FAQ and PERSONAL HOMEPAGE are easily recognized by users, and that users can handle a certain degree of granularity, such as distinguishing a personal homepage from a corporate homepage. Santini et al. [110] suggest that the reason users can find many Web pages difficult to assign to a particular genre is that these Web pages may belong to emerging genres that are either not yet widely recognized, or still in a state of transition and formation.

This section has given an overview of the research on Web page genre identification, emergence, and evolution. Researchers have demonstrated that Web genres, such as the HOMEPAGE and FAQ genres, are in fact distinct and recognizable genres, and that Web page genres continue to emerge and evolve. Although the topic of Web page genre evolution is not the focus of this thesis, it is relevant because it sheds light on some of the real world conditions that a Web page genre classifier could encounter. Ideally, a Web page genre classifier should be scalable enough to allow the addition of new genres, and adaptable enough to handle the drift that can occur in existing genres as they continue to evolve.

Any Web page genre classifier must have a method for representing Web pages, therefore Section 2.3 will give an overview of research on text and Web page genre classification, with a focus on the document and Web page representations used in genre classification.

2.3 Web Page Genre Characteristics and Representation

In order to classify Web pages by genre using a machine learning approach, it is necessary to identify features that effectively characterize each Web page and genre. As noted by Shepherd et al. [119], the goal is to select a set of features that will allow the classifier to distinguish one genre from another, and to assign the correct genre label to each Web page. The representations of Web pages used in the genre classification task tend to be based on those used in text classification, however the Web page representations may be augmented with information such as HTML tags, JavaScript code, and URL information.

Analyzing genre in academic and research settings, Swales [130] noted that members of a particular genre tend to exhibit similar patterns of content, style, structure, and intended audience. Analyzing genre in an online setting, Shepherd and Watters [116] proposed a similar characterization of cybergenres, using the properties content, form, and functionality, where functionality refers to capabilities available in the new medium, such as hyperlinks of video components. The properties of content, form, and functionality can be represented in a number of ways. Content can be represented by vectors of words, terms, or n-grams extracted from the text of the Web pages; these can be extracted on a statistical basis, such as frequency analysis, or they can be extracted on a syntactic basis, such as extracting all noun phrases. The form of a Web page can be represented by a variety of different features, including punctuation, part-of-speech statistics, the number of images or tables on the page, and the positioning of images or tables within the Web page. The functionality of a Web page can be represented by noting the presence or absence of executable code such as JavaScript and applets, or the number of hyperlinks on the page [119].

Section 2.3.1 gives a brief overview of text genre classification, with a focus on the representation of the text documents, while Section 2.3.2 reviews the research on Web page genre classification, again with an emphasis on the representation of the Web pages. Because the focus of this thesis is on n-gram representations of Web pages and Web page genres, Section 2.3.3 discusses research applications that make successful use of n-gram based representations, including those in the field of Web page genre classification.

2.3.1 Text Genre Classification

In an early study on categorizing texts into pre-determined genres, Karlgren and Cutting [58] focused on the properties of content and form, representing documents using linguistic features, such as third person pronoun occurrence. The computation of these features relies heavily on the use of a part-of-speech tagger to tag the documents as a preprocessing step. Karlgren and Cutting used discriminant analysis on the Brown corpus, which is a set of categorized English text samples. The results of these experiments indicated that classification error rates increased sharply as the number of categories was increased from two to fifteen. Karlgren and Cutting noted that distinguishing between different types of fiction proved to be particularly difficult. Despite their use of expensive part-of-speech tagging, Karlgren and Cutting found that the most discriminating features they used were word length and sentence length, and various derivatives of these two parameters.

In another early study on the automatic detection of genre in text, Kessler et al. [60] proposed decomposing genres into bundles of generic facets, which are simply attributes or properties that help distinguish one class of text from another. Working with a subset of the Brown corpus, Kessler et al. compared the use of what they call surface level cues with the structural level cues used by Karlgren and Cutting [58]. The former are easily computed facets such as counts of punctuation marks and lexical cues such as terms of address, whereas the latter are more computationally expensive facets such as part-of-speech tags. The results of these experiments indicated that there is only a small difference in the classification performance between using surface or structural facets to classify texts by genre. Kessler et al. concluded that in this case, the marginal advantage of using structural cues does not justify the additional computational expense their use incurs.

In other work on text genres, Stamatatos et al. [123] used an automated classification model based on discriminant analysis to classify a subset of the Wall Street Journal corpus into the four low-level genres EDITORIALS, LETTERS TO THE EDITOR, REPORTAGE, and SPOT NEWS. Although Stamatatos et al. used common word frequencies to select features, instead of extracting the most frequent words from the training corpus, they extracted the most frequent words from the British National corpus, which they used to represent the most frequent words in the entire written

English language. Stamatatos et al. achieved very high classification accuracy, and noted that including the frequencies of occurrence of the most frequent punctuation marks enhanced the classification performance of their model.

Also investigating the task of classifying text by genre, Dewdney et al. [37] compared the use of two different feature sets, one based on document content and one based on document form. The former consisted of the 323 word features, while the latter consisted of 89 linguistic features such as the sentence complexity, line spacing, tabulation, punctuation usage, and adjective occurrence. Dewdney et al. experimented with these feature sets using three different classifiers on a corpus of seven genres. Of the three classifiers, both the C4.5 decision tree and the support vector machine performed better using form features than content features, however the opposite was true for the Naïve Bayes classifier. Dewdney et al. then tested the classifiers with a feature set that combined the content and form attributes, and found that this combination of feature sets gave the best results for each classifier. Wastholm et al. [138] performed similar experiments on a corpus of 500 Swedish texts, divided into nine genres. Based on their experiments, Wastholm et al. concluded that text genre classification performance can be improved by using linguistic features instead of, or in addition to, word features.

Focusing on a small corpus of French scientific texts, Cleuziou and Poudat [28] also found that combining feature sets composed of both content and form-based features provided better classification results than using either feature set by itself. Cleuziou and Poudat classified a small corpus of 371 documents that belonged to one of the three genres ARTICLES, JOURNAL PRESENTATIONS, and REVIEWS. Cleuziou and Poudat represented the content of the documents by extracting singular and plural nouns, using Mutual Information as the feature selection measure. The form of the document was represented using features that included both part-of-speech and morphosyntactic information, and for this feature set Information Gain was used as a feature selection measure. Unfortunately, Cleuziou and Poudat do not explain their choice of feature selection measures, nor why they chose to use a different feature selection measure for form-based features than they used for content-based features.

Stamatatos et al. [124] used natural language processing (NLP) tools to classify a small corpus of Greek texts by genre and by author. The corpus was created by downloading Greek texts from the Internet, with the goal of covering as many genres as possible; the resulting corpus contains 250 documents distributed evenly across 10 genres. Stamatatos et al. used a combination of style markers and stylometric markers as features. The former include linguistic features such as the occurrence of noun phrases or verb phrases, while the latter included measures such as the number of words left unanalyzed after each pass of the NLP tool. The experiments were run using both discriminant analysis and multiple regression techniques, with discriminant analysis giving slightly better performance for both classification by genre and classification by author. The combination of stylistic and stylometric features used by Stamatatos et al. outperformed lexically based methods used on the same corpus. Stamatatos et al. observed that some genres, such as RECIPES and INTERVIEWS, were more homogeneous and thus easier to classify than others, such as PRESS EDITORIAL and CURRICULA VITAE.

Kim and Ross [63, 64, 65, 66] investigated the genre classification of documents represented in PDF, as a first step toward automating metadata extraction. Their motivation for this work was that identifying the genre of a document provides a means to limit the scope of the forms from which to extract other metadata; within a certain genre, specific metadata can be expected to appear in a specific place. For some classes of documents, genre-specific metadata extraction methods already exist, and so automatically identifying the document genre can be very helpful [63, 65]. For the task of genre classification, Kim and Ross proposed representing documents using five types of features: visual layout, style, topic, semantic patterns, and contextual elements. The visual layout features were extracted from the first page of a PDF file when it is treated as an image; Kim and Ross pointed out that some PDF files are textually inaccessible due to password protection, and that a visual layout classifier would be especially useful in this case. Kim and Ross hypothesized that genre classification could be viewed as a multi-dimensional combination of several independent classification tasks because the distinction between some genres may be largely structural, while the distinction between other genres may rely more on the visual representation or style. Their experiments investigated the effectiveness of some of these feature sets, both independently and in combination. Based on these experiments, Kim and Ross have concluded that the best feature type for detecting one particular genre may not be the best feature set for detecting other genres. Kim and Ross [66] noted that although their focus was on documents represented in PDF, their classification model is dependent on the PDF only to the extent that it uses PDF tools to convert the documents into image and text, and thus their classification model could be applied to other text documents or Web pages.

This section has given an overview of the research on the classification of text documents by genre, with a focus on the feature sets that are used to represent the documents for the classification. Choosing the best features with which to represent a document for classification is a challenging task. The goal is to choose a set of features that will allow the classifier to distinguish one genre from another, and to assign the correct genre label to each document. Using a combination of features that represent different properties of the genre, such as content and form, seems to produce better classification results than using features that represent only a single genre property. This observation is reinforced by some of the work to be discussed in Section 2.3.2. Section 2.3.2 will highlight the existing research on the classification of Web pages by genre, again with a focus on the feature sets and the combinations of the types of features that are used in the classification process.

2.3.2 Web Page Genre Classification

Early work on Web page classification included that by Chekuri et al. [23], who automatically classified Web pages into pre-specified categories, with the goal of increasing the precision of Web searches. Chekuri et al. used a term-frequency based classification process in which they began with a set of categories and a pre-classified training set of pages, and built a vector of term frequencies for each of the categories. A new Web page was classified by computing the word frequency vector for the page and comparing it with the vectors representing the various categories, using a distance measure; an ordered list of the document's most similar categories was then returned. Chekuri et al. concluded that the automatic classification of documents into categories can enhance Web searches, however they noted that the quality of the classification depended on the degree of separation between the categories; the fuzzier the categories, the poorer the classification results.

Craven et al. [30] explored the goal of creating a large knowledge base whose content would mirror that of the World Wide Web. They proposed classifying Web pages into a symbolic ontology and developing a system that could be trained to extract symbolic knowledge from hypertext, using a variety of machine learning methods. As part of this work, they constructed the WebKB corpus, which contains 8282 Web pages collected from Computer Science departments of various universities. The Web pages are labeled as belonging to the categories (STUDENT, FACULTY, STAFF, DEPARTMENT, COURSE, PROJECT, and OTHER). Using a bag-of-words approach and Naïve Bayes classifiers, Craven et al. compared classifiers trained using text from the whole Web page, text from only the title and HTML tags, and text from only the hyperlinks. They found that the best representation of a Web page depended upon the class of the Web page, however their attempts to combine the classifiers gave disappointing results.

Focusing on the problem of classifying Web pages by genre, Rehm [92] proposed developing a Web genre identification system for the academic domain, with a hierarchical genre structure. Rehm suggested that the characterization of Web genres proposed by Shepherd and Watters [116], using the properties of content, form, and functionality, should also be assigned to generalized classes of genre modules, from which the basic framework for a Web genre could be constructed. Rehm's proposed approach therefore used feature sets that combined attributes related to the properties of content, form, and functionality. These included content-based linguistic features, form-based graphics related data, and functionality-based hyperlink, HTML, and JavaScript information.

Asirvatham and Ravi [12] obtained encouraging results using feature sets that combined components based on content, form, and functionality to classify university Web pages into the three broad categories INFORMATION PAGES, RESEARCH PAGES, and PERSONAL HOME PAGES. The feature set used by Asirvatham and Ravi included counts of the characters and commonly used words (content-based features), the size, color, and placement of images (form-based features), and various hyperlink information (functionality-based features). Asirvatham and Ravi noted that although they did not include video and other multimedia information in their feature set, they would suggest this as an improvement for future implementations.

Shepherd et al. [119] also examined the use of feature sets which combined the attributes of content, form, and functionality to represent Web pages, and compared the classification results with those of feature sets which used only content and form to characterize the Web pages. They used a neural net classifier to distinguish HOMEPAGES from NON-HOMEPAGES and to classify those homepages as belonging to the PERSONAL HOMEPAGE, CORPORATE HOMEPAGE, or ORGANIZATION HOMEPAGE genre. Shepherd et al. found a significant improvement in identifying personal and corporate homepages when the functionality attribute was included. Although there was no significant improvement in identifying organization homepages when the functionality attribute was included, Shepherd et al. noted that this genre was significantly more difficult to classify than the other genres, and speculated that perhaps the style of the ORGANIZATIONAL HOMEPAGE genre is not as unique as that of the PERSONAL HOMEPAGE and CORPORATE HOMEPAGE genres.

Dong et al. [41] also investigated the use of feature sets with combinations of content, form, and functionality to represent Web pages for the task of genre classification. Their data set included four Web page genres, as well as noise pages that did not belong to any of the four genres. Based on their experiments, Dong et al. concluded that using combinations of features based on the properties of content, form, and functionality always gave better results than using feature sets based on only one of these properties. Interestingly, Dong et al. found that a smaller feature set was more effective in increasing precision than it was in increasing recall, whereas a larger feature set was more effective in increasing recall.

Jebari [52] and Jebari and Ounalli [53, 54] focused on the form and structure of Web pages in order to classify the pages by genre. They implemented two classifiers, one which uses URL information, and another which uses HTML tags. For the latter classifier, Jebari and Ounalli found that using the title, anchor, and headings tags gave the best results. Based on the training data, each classifier constructs a centroid feature vector for each genre, to which the feature vector for each Web page from the test set is compared using the cosine measure. Each classifier then returns a weight for each genre, indicating with what degree of confidence the Web page belongs to that genre. Jebari and Ounalli found that combining the results of the two classifiers gave better results than using either classifier on its own.

In more recent work on Web page genre classification, Levering et al. [71] compared three different sets of features: textual features that included part-of-speech statistics and word frequencies, a combination of textual and HTML features, such as link counts, URL lengths and JavaScript counts, and a combination of textual, HTML, and visual features, including image statistics and page dimensions. In terms of content-based, form-based, and functionality-based features, the textual feature set combines attributes from the Web page properties of content and form, while the other two feature sets combine attributes from the properties of content, form, and functionality. For each combination of features, Information Gain was used to select the top 100 features to be used for classification. Levering et al. ran their experiments on a corpus of Web pages of e-commerce stores, limited to the three genres HOMEPAGES, PRODUCT LISTS, and PRODUCT DESCRIPTIONS. Interestingly, Levering et al. also investigated the inclusion of noise Web pages in their corpus. The results of the experiments indicated that when noise pages were not present in the corpus, each feature set performed well, however the classification accuracy for each genre improved as HTML and then visual features were used in addition to the textual features. When noise was present in the corpus, the textual features alone performed very poorly. The combination of textual and HTML features gave the best classification performance for two of the three genres when noise was present, with visual features only improving the accuracy of the PRODUCT LIST genre in this case. Levering et al. concluded that the usefulness of visual features in a noisy environment appears to be very genre dependent. In terms of the properties of content, form, and functionality, the feature sets that combined features from all three attributes outperformed the feature set that contained only the content and form based features. It is interesting to observe that although the classification accuracy for the PRODUCT LIST and PRODUCT DESCRIPTION genres decreased for each feature set when noise pages were added to the corpus, the accuracy for the HOMEPAGE genre increased for the cases in which combinations of feature types were used in the presence of noise.

Rather than focusing on combining content, form, and functionality features to represent Web pages, Lee and Myaeng [69] presented an automatic genre classification method that was based on combining statistically selected term features from both subject-classified and genre-classified training data. Their hypothesis was that the use of features selected for one type of classification, such as genre-based classification, can be complemented with the use of features selected for another type of classification, such as subject-based classification. They compared two approaches, one based on document frequency ratios, and one based on term frequency ratios. Lee and Myaeng tested their classification method on both Korean and English document sets, each containing seven genres. They found that using the document frequency ratios gave the best results, but that some genres were difficult to classify using their method because the genres shared subject-related terms. In later work, Lee and Myaeng [70] investigated the question of whether using genre information would help in subject-based classification, but their results were not conclusive.

Also working on the Web genre classification problem, Finn and Kushmerick [43] compared the results of three different feature sets used to represent Web pages: a set of linguistic features, such as average sentence and word length, a bag-of-words feature vector, and a set of 36 part-of-speech statistics from tagged text. The former two feature sets contained content-based features, whereas the latter contained form-based features. In their Web page classification experiments, Finn and Kushmerick focused on the genre dimensions of subjectivity and sentiment. For the subjectivity dimension, documents from a particular genre were classified as being an objective report or a subjective opinion, whereas for the sentiment dimension, reviews were classified as being positive or negative. Finn and Kushmerick found that the performance of each feature set depended on the genre classification task (subjectivity or sentiment classification), but that using a classifier that combined the three feature sets gave the best results in most cases.

Lim et al. [72, 73] investigated the classification of Web pages by genre using a corpus of Korean Web pages containing 16 genres. In addition to characterizing Web pages using the content-based and form-based features extracted from lexical, token, and structural information, Lim et al. experimented with the use of information extracted from the URL and HTML tags of a Web page. They also investigated the relative usefulness of information found in different parts of the Web page, by dividing each page into three sections (title and meta content, anchor text, and body text). Lim et al. found that the best combination of feature sets included the combination of content and form, augmented with URL and HTML information; the features included the combination of the most frequently used words, punctuation, and chunks (multi-word expressions), token information, and information from the URL and the HTML tags. Lim et al. also concluded that including the body text of a Web page is essential for successful genre classification. In addition, they noted that for the Korean language, in which each word is composed of a content word and one or more function words, the function words have more discriminating power than the content words for identifying the genre of a Web page.

Crowston and Kwasnik [32] also explored Web page genre as a multidimensional phenomenon. They suggested the use of a facetted approach to Web page representation, where facets are basic attributes that allow the discrimination between genres. With this approach, rather than limiting the characterization of genre to the properties of content, form, and functionality, models would be built with facets that represent the Web page from many conceptual perspectives, including content, form, style, source, implied use, and the relationship of a document to other documents. Crowston and Kwasnik suggested that the use of a set of facets would allow for a genre representation that is more complex and flexible than with other approaches. They noted that such a model would be capable of accommodating new genres as they emerge, that the approach does not require complete knowledge about a genre, and that the approach allows the genre to be viewed from multiple perspectives. The drawbacks to the facetted approach are the difficulty in establishing appropriate facets, the lack of relationships among the facets, and the difficulty in visualizing the structure, in comparison to simple clusters or hierarchies.

In contrast to the facetted approach to genre classification, Clark and Watt [27] obtained encouraging results classifying XML documents by genre using only form-based features. Clark and Watt used a small set of 28 features that included part-of-speech information such as tense, prepositions, and modal verbs, and XML tag information such as the frequency of URL tags. They also included features such as the presence or absence of images and abstracts. Working on a small subset of the INEX corpus in which all of the documents belonged to the broad ARTICLE genre, Clark and Watt found that many articles from different genres were structurally identical and needed to be merged into a single genre; for example the THEME ARTICLE and FEATURE

ARTICLE genres were merged. Clark and Watt also found that the seven most effective features for this particular corpus were the number of tags per document, the average word length, the frequency of reference tags, the average paragraph size, the number of images, the size of the document, and the number of table tags. Although Clark and Watt achieved high classification accuracy using only form-based features, their particular document representation model is limited to XML and XHTML documents.

Similar to Clark and Watt, Santini [102, 110] focused on using small sets of carefully chosen features to represent documents for classification by genre. Santini introduced an inferential model for classifying Web pages by genre, based on a modified version of Bayes' theorem. This model used a small set of linguistic features to infer the text type of each Web page, where the text types were limited to DESCRIPTIVE NARRATIVE, EXPOSITORY INFORMATIONAL, ARGUMENTATIVE PERSUASIVE, and INSTRUCTIONAL. Once the text types were inferred, Santini used the two predominant text types, in combination with features such as layout and functionality tags, as input to a set of hand-crafted if-then rules which determined the genres of the Web page. The number of rules varied according to the particular genre and the computation for each genre was performed independently for each Web page, which allowed a page to be assigned to more than one genre, or to no genre at all. In a comparison of this inferential model with a support vector machine classifier and a Naïve Bayes classifier, Santini found that on a balanced corpus of seven Web genres, the inferential model's classification accuracy of 86% was close to that of the support vector machine's 89% accuracy, and well above the Naïve Bayes classifier's 67% accuracy. Santini concluded that the inferential model was effective, but that the parameters needed better tuning, and that the number of text types needed to be expanded.

In other work, Santini [101] compared the Web page classification performance of a support vector machine when three different feature sets were used. The first feature set was composed of the 50 most common English words, 24 part-of-speech tags, 8 punctuation marks, 28 HTML tags, a document length attribute, and a selection of genre specific words that were chosen through the manual analysis of various Web page genres. The second feature set also included the same punctuation symbols, genre-specific words, HTML tags, and document length attribute, but included part-of-speech trigrams instead of the 50 most common English words and 24 part-of-speech tags. The third feature set contained the genre-specific words and document length attribute, as well as 86 linguist facets and 6 HTML facets, where facets are groups of features. Experimental results indicated that although all three feature sets allowed good classification performance, the best classification accuracy was achieved using the first feature set, while the third feature set gave the poorest results.

Similar to Santini, Meyer zu Eissen and Stein [84] also performed Web page genre experiments to compare the classification performance of relatively small feature sets. Meyer zu Eissen and Stein combined genre-specific vocabulary and closed-class word sets with text statistics, part-of-speech information, and presentation-related features such as counts of tables and figures, and statistics related to HTML tags and URL information. They compared the classification results of two feature sets, one with 25 features that did not include part-of-speech features, and the other with 10 part-of-speech features in addition to the same 25 features used in the first feature set. The experiments were run using neural network learning and support vector machines on a corpus with 800 Web pages evenly distributed over eight genres. Meyer zu Eissen and Stein found that on average, the support vector machines achieved better classification performance than the neural networks, and that the larger feature set that included the part-of-speech features achieved better classification performance than the smaller set that did not include these features.

In more recent work, Stein and Meyer zu Eissen [126] showed that genre classification models that only used feature sets of genre-specific vocabulary could achieve state-of-the-art performance. Stein and Meyer zu Eissen found that feature sets of approximately 40 terms per genre were most effective. Their experiments were run on a corpus of 2000 Web pages evenly distributed over eight genres. Although high classification results were achieved using only the genre-specific vocabularies as features, Stein and Meyer zu Eissen noted that the results were improved when part-of-speech and HTML features were also included. Stein and Meyer zu Eissen's experiments, however, indicated that limiting the feature sets to genre-specific terms provided a less corpus-dependent and more generalizable classification model.

This section has provided highlights of the research on Web page genre classification, with a focus on the feature sets that are used to represent Web pages for the classification. As in text document classification, selecting the best features with which to represent a Web page for classification is a challenging task for which many approaches have been explored. Much of the research discussed in this section has shown that using a combination of features that represent different properties of the Web genre, such as content, form, and functionality, seems to produce better classification results than using features that represent only a single genre property.

Because the work in this thesis focuses on representing a Web page using n-grams, Section 2.3.3 will discuss some of the research applications in which n-gram representations are useful, including the classification of Web pages by genre.

2.3.3 n-gram Representations

Many research applications make successful use of n-gram-based representations. The use of n-grams has been common in language modeling since at least 1948 when Claude Shannon, considered the father of information theory, investigated the question of determining the likelihood of the next letter in a given sequence of characters [113]. Since that time, n-grams have been widely used in natural language processing and statistical analysis. In terms of document classification, the basic idea is to identify n-grams whose occurrence in a document gives strong evidence for or against identification of a text as belonging to a particular category [22]. Although there are many different types of n-grams, the main focus of the work discussed in this thesis is on the use of byte and character n-grams, therefore unless a specific distinction is made, the use of the term *n-gram* refers to byte and character n-grams.

Although the term n-gram typically refers to contiguous substrings of text, combinations of contiguous and non-contiguous substrings have also been used for the purpose of text classification. *String kernels* are substrings that are not necessarily contiguous, but that are assigned different weighting based on their degree of contiguity. Lodhi et al. [75], for example, test the performance of string kernels for the task of text categorization using a support vector machine. Using the Reuters data set, they compared the results of the use of a string kernel approach, a bag-of-words approach, and an n-gram based approach. Although the string kernel approach delivered state-of-the-art performance comparable to that of the bag-of-words approach, the use of contiguous n-grams outperformed the string kernel approach. The research for this thesis represents Web pages using fixed-length contiguous n-grams. These n-grams can be thought of as the contents of a fixed-size sliding window moved through the text.

Kešelj et al. [62] note that n-gram language models can be easily applied to any language, as well as to non-language sequences such as music and DNA. For example Nelson and Downie [87] use n-gram representations of music for a music database, as do Suyoto and Uitdenbogerd [129]. The field of bioinformatics makes extensive use of n-gram analysis. In this case the vocabulary is, for example, nucleotides and amino acids instead of words, bytes, or characters. In this field, n-gram representations may be used for such problems as protein sequence classification [26, 46], genome sequence classification and clustering [132], indexing and retrieving genomic sequences stored in large biological databases [51], identifying genomic and pathogenesis islands in bacterial genomes [86], or segmenting protein amino acid sequences [25]. In the latter case, the approach is to build n-gram language models of the three secondary protein structures (helices, extracellular loops, and intracellular loops), and then compare their performance in predicting the amino acid, to determine whether a boundary occurs at that position.

Cavnar and Trenkle [22] address the problem of text classification with a model that represents documents as profiles of variable-length character n-grams of lengths 1 to 5. These character n-grams have been selected based on their frequency; digits and all punctuation other than apostrophes are discarded, and word tokens are padded with blanks. A rank-order statistic which is referred to as the *out-of-place* measure is used to determine the similarity between a document and a category profile, with the document being assigned to the category to which it is most similar. In order to compute the out-of-place measure, the character n-grams in each document are ranked by their frequency, and the out-of-place measure determines how far out of place each character n-gram in one profile is from its ranked place in the other profile, with some maximum value being assigned if the character n-gram does not exist in the other profile. These absolute values are summed to obtain the distance between the two profiles. For language classification, this system gave its best performance with n-gram profile sizes of 400, achieving a classification rate of 99.8% when classifying Usenet newsgroup articles written in eight different languages. Very good results were also reported for the application of this technique to the classification of Usenet newsgroups articles by subject. Cavnar and Trenkle noted that their system required no semantic or content analysis, apart from the n-gram frequency profile, and concluded that their n-gram frequency method provided a highly effective, yet inexpensive way of classifying documents. A similar model that used fixed-length character n-grams was successfully implemented as an information retrieval system by combining the n-gram representations of documents with vector processing models [21].

In the area of document clustering, Miao et al. [85] found that in a comparison of n-gram based, term-based, and word-based clustering, using an n-gram representation gave the best clustering results. They also found that character n-gram based clustering gave better results, with lower dimensionality, than did byte n-gram based clustering, with a character tri-gram representation giving the best combination of clustering quality and practical dimensionality. Miao et al. [85] also noted that the use of their n-gram method was robust in the sense that it needed no language-dependent preprocessing, such as stopword removal.

Liu and Kešelj [74] used character n-grams to represent the contents of Web pages and combined them with user navigation profiles, also composed of a collection of character n-grams, in order to successfully classify Web user navigation patterns and to predict users' future requests. They pointed out that their n-gram representation had advantages that include robustness, independence from both language and topic, and applicability to different file formats.

Problems such as multi-lingual document classification [125], machine translation [76, 77], the evaluation of machine translation [39, 89], spam detection [61], the classification of Chinese text [139, 147], and the detection of spelling errors in Turkish text [13] have also been addressed using n-gram based representations.

Working on the problem of automated authorship attribution, which is the problem of identifying the author of a text whose authorship is in doubt, Kešelj et al. [62] used n-gram representations of documents to achieve state-of-the-art performance on three different languages: English, Greek, and Chinese. Document and author profiles were constructed using the L most frequent n-grams and their corresponding normalized frequencies; in the training set, documents by the same author were concatenated, and then the n-grams were extracted to form the author profile. The dissimilarity between each document in the test set and each author profile from the training set was compared using a distance measure, and the document was labeled as belonging to the author to which its profile was closest (most similar). For a pilot study of English texts, Kešelj et al. reported achieving 100% accuracy for several profile and n-gram sizes including, surprisingly, n-grams of length 1 with a profile size of only 20. They noted, however, that the worst performance for their model was for the Chinese language, possibly because Chinese characters are two bytes long, which means that 75% of the n-grams would not be sensible strings in Chinese.

Amasyali and Diri [11] also successfully used a character n-gram based approach to the authorship attribution problem, however their work was on Turkish texts, downloaded from Turkish daily newspapers. In addition to determining the author of the texts, this model was also used successfully for identifying the gender of the author, and for classifying the texts by genre. Using an n-gram frequency representation of the texts, Amasyali and Diri compared the performance of several classifiers, including the Naïve Bayes and support vector machine classification methods. They found that for each classifier and each problem, using a threshold value as a feature selection measure to reduce the number of n-grams for each text improved results. Amasyali and Diri also found that although the Naïve Bayes classifier gave the best results for identifying the author of the text, the support vector machine approach gave the best results for assigning the genre and author gender to the text. In these experiments, the use of bi-grams gave the best results overall.

Stamatatos [122] also used character n-gram representations of documents for the problem of author identification, in this case on two collections of Greek texts. Stamatatos commented that this n-gram based approach requires minimal text preprocessing, and that the n-gram extraction was language independent. Each text was represented as a vector of n-gram frequencies; for these experiments, the length of the n-gram was varied from 3 to 5, and the number of n-grams used was varied from 1000 to 10,000. In order to deal with the high dimensionality of the feature set, Stamatatos used feature space subspacing, in which the feature set is divided into smaller parts, each used to train a base classifier. The results of these base classifiers were then combined to predict the most likely class of the document. With this method, Stamatatos achieved classification results of 94% and 100% on two Greek corpora, for several of combinations of n-gram length and number.

Houvardas and Stamatatos [50], also working on the problem of authorship attribution, proposed a new feature selection technique for variable-length character n-grams in which each n-gram is compared with similar n-grams (either longer or shorter) in the feature set and the most important of them is kept. Houvardas and Stamatatos adapted this selection technique from an existing approach for extracting multiword terms (i.e., word n-grams of variable length) from texts [35, 36]. With this method, the factor affecting the importance of each n-gram is the frequency of its occurrence. Houvardas and Stamatatos noted that when using variable-length character n-grams, an aggressive feature selection method is necessary, due to the high dimensionality of the feature space. Once the feature selection was been made, they trained a support vector machine classifier on the reduced feature set, and then applied the support vector machine classifier to the test set. The experimental results indicated that the proposed feature selection technique resulted in higher classification accuracy than when Information Gain was used as a feature selection method. Houvardas and Stamatatos commented that character n-grams are able to capture complicated stylistic information on the lexical, syntactic, and structural levels, and that no tokenization was necessary when extracting the n-grams, making the approach language independent. They did note, however, that the optimal number of n-grams to use could depend on the language.

Kanaris and Stamatatos [56, 57] applied the feature selection method for variable-length character n-grams, presented by Houvardas and Stamatatos [50], to the problem of Web page genre identification. In their experiments, they tested two models: the first model used only feature sets of variable-length character n-grams from the textual content of each Web page, whereas the second model augmented the first model with structural information about the most frequent HTML tags. The HTML tag information was constructed by first creating a list of all HTML tags that appear three or more times in the entire data set. Each Web page was represented by a vector of the HTML tag frequencies; the ReliefF feature selection algorithm [96] was then applied to reduce the dimensionality of the vectors. As with Houvardas and Stamatatos [50], after the feature sets were reduced, classification was performed using a support vector machine classifier. Kanaris and Stamatatos reported that the accuracy of the Web page genre classification using their technique was higher than that previously reported by researchers on the same corpora.

This section has provided an overview of some of the research applications in which n-gram representations are used. Although the spirit of much of this work is similar to the research to be presented in this thesis, in the sense that various types of n-gram profiles are used to represent documents or other items for the purpose of classification, the essential details differ greatly. These differences include the number and type of n-grams used, the n-gram selection process, the representation of the classification categories, and the determination of similarity between two n-gram representations.

Regardless of the document representation that is used, it is necessary to consider the dimensionality of the feature space. Section 2.4 will discuss methods to minimize the size of the feature sets used to represent documents and Web pages in classification problems.

2.4 Dimensionality Reduction Using Feature Selection

With text classification tasks, once the representation for the document has been determined based on some combination of features or facets, it is often necessary to select a subset of the features in order to reduce the dimensionality of the search space, and thus reduce the computational complexity of the problem. This is also the case with the task of Web genre classification. The term *curse of dimensionality* was coined by Richard Bellman [15] to describe the problem that occurs when searching in high dimensional spaces. As the dimensionality of the input data space increases, it becomes exponentially more difficult (more computationally complex) to fit models for the parameter space. In practice, feature selection techniques may be necessary in order to select a subset of relevant features for building robust learning models, because most standard machine learning techniques cannot be directly applied when the dimensionality is very high. The appropriate selection of features is very important; as noted by Shepherd et al. [119], the success of Web page genre classification, as with other machine learning problems, is highly dependent on the feature set that is selected. The goal of feature selection techniques is to recognize and select the most discriminative features, and to reduce the feature set size by discarding the less informative features. There are several ways in which this feature selection can be performed.

One method for reducing the number of features used to represent a document or Web page is to rank the features according to a particular feature selection measure, and to select the top ranked features to represent the document or Web page. Kešelj et al. [62] note that this is not a trivial process because it involves setting thresholds to eliminate the less informative features. Yang and Pedersen [142] provide a comparative study of the traditional feature selection techniques used for dimensionality reduction in text classification. Yang and Pedersen evaluated five feature selection measures: term selection based on document frequency, Information Gain, Mutual Information, the χ^2 statistic, and Term Strength. In their experiments on the Reuters and OHSUMED text collections, Yang and Pedersen found that Information Gain and the χ^2 statistic were the most effective measures of feature selection for aggressive feature reduction in which up to 98% of the features were removed, but that performance for term selection based on document frequency was comparable to Information Gain and the χ^2 statistic for less aggressive feature reduction, when only up to 90% of term features were removed. Yang and Pedersen noted that Mutual Information had inferior performance compared to the other measures, and concluded that this was due to a bias favoring rare term features and a sensitivity to probability estimation errors.

Forman [45] also provides an extensive study of feature selection techniques for text classification. Forman compared twelve feature selection measures using a support vector machine classifier. The feature selection measures included in this comparison are Accuracy, Balanced Accuracy, Bi-Normal Separation (BNS), the χ^2 statistic, Document Frequency, Information Gain, Odds Ratio Numerator, Odds Ration, Power, Probability Ratio, and Random. Forman's experiments indicated that when feature sets were in the range of 20 to 50 features, Information Gain gave the best results, whereas when the feature sets were in the range from 500 to 1000 features, the BNS measure gave the best results. In terms of precision, Information Gain and the χ^2 statistic gave the best performance, however overall, the BNS measure gave the best performance. Probability Ratio, Document Frequency, and Random gave the worst results overall. Forman noted that only Information Gain and the BNS measure gave better performance than using all of the available features.

Focusing specifically on the classification of Web pages by genre, Dong et al. [40] evaluated three measures for feature selection. Dong et al. compared the performance of Information Gain, Mutual Information, and the χ^2 statistic for selecting features for the binary classification of Web page genres. They found that although all three selection measures were capable of detecting small sets of discriminating features, when feature sets were as small as five, only Information Gain and the χ^2 statistic were able to successfully select features that gave good performance.

The use of such feature selection measures is common in classification and clustering tasks. Miao et al. [85], working on the problem of document clustering, performed experiments with feature selection based on document frequency, and they found that this successfully reduced the dimensionality for character n-gram representations of documents. Liu and Kešelj [74] also employed document frequency to filter out the less important character n-grams when building profiles to aid in the classification of Web user navigation patterns and the prediction of users' future requests. Investigating the task of classifying documents by genre, Dewdney et al. [37] used the Information Gain algorithm over the whole corpus to rank the features, selecting the features with the highest Information Gain score. Zhang et al. [146] investigated the automatic classification of Web pages using Information Gain as a feature selection measure, however they then used principal component analysis to further reduce the vector space, before using a C4.5 decision tree classifier. Using the top 300 features ranked by Information Gain, they achieved about 80% precision and recall on a small corpus of 500 Web pages containing health information, education information, shopping information, and noise. Their features were selected entirely from the textual content of the Web page, after the removal of all HTML tags, images, and so forth; they also reduced the feature space by removing stopwords and performing stemming on the remaining words.

Another common approach to the issue of dimensionality is simply to build a model that uses only a relatively small number of carefully selected features to begin with, so that feature set reduction is unnecessary. Kennedy and Shepherd [59] conducted experiments to compare the effects of manual feature selection versus principal component analysis feature selection, for the task of Web page genre classification. Kennedy and Shepherd ran experiments to classify three subgenres of homepages, and found that in general, the feature sets that were selected manually provided better performance than the feature sets selected using principal component analysis. The problem with this method is in having enough prior knowledge or expertise to hand-pick appropriate features. For example, Stubbe et al. [128] used an intensive iterative process to create a hand-crafted collection of textual features for each genre. Interestingly, one of the features they used was the rate of errors, such as spelling mistakes, in a document, however this feature proved to be beneficial with the classification of some genres, such as PERSONS, but detrimental to others, such as REPORTAGE. Although Stubbe et al. achieved high precision on a corpus of 1280 English Web pages divided into 32 genres, this hands-on approach to creating a specialized classifier for each genre would require significant effort to make the transition to other genres or languages. Stubbe et al. also noted that the limits of their method were reached for documents that did not have a specific structure or vocabulary, and for genres with strong similarities.

Other researchers who have crafted small feature sets include Braslavski [19], who used nine carefully chosen features, such as average word length, adverb count, and smiley (happy face) count to compute a genre-related score for use in relevance ranking. In this work, Braslavski equated functional style with genre, and used a small collection of 305 Russian documents to compute genre-related scores. Each document in the corpus was already labeled as belonging to one of five functional styles: OFFICIAL, ACADEMIC, JOURNALISTIC, LITERARY, and EVERYDAY COMMUNICATION. The genre-related scores were then used to re-rank Russian Web page results from keyword searches. Brasklaski found that the genre-related scores, based on only nine document features, provided moderate improvements to the search results. Chen and Choi [24] analyzed hundreds of Web pages in order to choose a small set of 31 features with which to represent Web pages for genre classification; these features included information from the Web page content, URL, HTML tags, JavaScripts, and VB scripts. For every genre, the classification system estimated a weight for each of the 31 features, as well as a threshold for the genre. However, in order to achieve high accuracy with this feature set, Chen and Choi noted that the system required considerable manual fine tuning of the feature weights and the thresholds during the training phase, which would make it difficult to add new features or new genres.

Another less common but simple alternative to reducing the feature set size is to select features from only the first fraction of the document or Web page. This approach makes the assumption that the most important information and discriminating features are found near the beginning of the document. Shanks and Williams [112] were able to accurately classify text documents with this method of feature set reduction, while Wibowo and Williams [140] applied this approach to the hierarchical classification of Web pages. Kim and Ross [63, 64, 65, 66] also followed this approach for one of the classifiers they investigated for the task of classifying documents in PDF by genre; the visual layout features for the classifier were extracted from only the first page of a PDF file when it was treated as an image.

This section has discussed some approaches to dealing with the problem of the curse of dimensionality. The goal in dimensionality reduction is to recognize and select the most discriminative features, and to reduce the feature set size by discarding the less informative features. The selection of a relatively small number of features with which to represent a document or Web page can be a crucial step in the classification process. As described in this section, this challenge can be met in a number of different ways, including the use of feature selection measures such as Information Gain, the handcrafting of a small number of carefully selected features, or the selection of features from only the first fraction of a document or Web page. The approach in this thesis is to use a feature selection measure to select the top ranked features. The performance of three such measures, frequency, Information Gain, and the χ^2 statistic, is evaluated in Section 5.2.

In addition to genre evolution, Web page representation, and feature selection, there are several other conditions that should be considered when designing and testing a Web page genre classifier; these will be discussed in Section 2.5.

2.5 Real World Conditions

As with most young fields of research, much of the initial investigation of Web page genre classification has been carried out based on a number of simplifications that are not entirely consistent with real world conditions. This section discusses the limitations of some of the existing Web page genre classification research.

Most research on the classification of Web pages by genre has focused on labeling each Web page as belonging to a single genre, however, the difficulty of assigning a single genre label to a Web page has been acknowledged by researchers who have conducted surveys and user studies about Web page genre [33, 34, 84, 98, 99, 108]. For example, Shepherd and Watters [117] examined 96 Websites looking for patterns in Web design, and grouped the Websites into five genres at a high level of abstraction, namely HOMEPAGE, BROCHURE, RESOURCE, CATALOGUE and GAME. They found that the boundaries between genres were fuzzy, and that Websites may be composed of more than one genre. In more recent work, Rosso [99] explored the use of genre to improve the effectiveness of Web searching by conducting a series of user studies in which participants were asked to classify Web pages by genre. Rosso found that the two factors which seemed to hamper participant agreement on Web page genres were that some of the Web pages seemed to fit into multiple genres, and that some of the genres seemed to have fuzzy boundaries. Despite these issues, 90% of the Web pages were classified into a single genre by the majority of the participants, leading Rosso to conclude that although a multi-genre classification scheme would be superior, it is still possible that a single genre classification scheme would offer improvement in Web searching.

While building a corpus of genre-tagged Web pages, Kwasnik and Crowston [68] also found that Web pages can be composed of pieces of more than one genre, and that some genre boundaries are fuzzier than others. Vidulin et al. [137] constructed a Web page corpus of 20 genres selected with the intention of covering the whole Internet, however they found that the genres were not clearly delineated, and therefore chose a multi-label approach in which each Web page can be labeled as belonging to more than one genre. Stubbe et al. [128] also constructed a Web page corpus; this hierarchical corpus is composed of 1280 English Web pages divided into 32 fine-grained genres, which can be combined into 7 coarse-grained genres. Although Stubbe et al. attempted to include only unequivocal documents in the corpus, they found that approximately 22% of the Web pages contained material that belonged to more than one genre. Stubbe et al. concluded that, depending on the application, it could be better to allow multiple classification. They proposed a variant of multiple classification that used knowledge about genre interdependencies as a means of filtering the multi-class results; the filter had disqualification rules, specifying that if a Web page had been labeled as X, it could not also be labeled as Y.

Santini has proposed what she calls a zero-to-multi-genre classification scheme [105, 106, 107, 108]. Santini's user studies indicated that, at least from a user's perspective, a single-genre classification scheme is too narrow. Although some Web pages may fit neatly into a single genre, others do not fit into any genre, while still other Web pages may fit appropriately into several genres. When limiting users to assigning a single genre label to a Web page, Santini found not only that the users tended to disagree on the genre label, and but also that the users complained about the single-label limitation [108]. Santini suggested designing genre classification models that not only allow Web pages to have more than one genre label, but that also allow a Web page to have a zero-genre label. The latter would be the equivalent of the "I don't know" option in user studies, for the case in which a Web page does not match the conventions of any known genre; such Web pages could also be defined and labeled as noise. Labeling Web pages in this way is not only a more flexible approach from the users' perspective, but also better reflects the complex nature of the mixture of pages found on the World Wide Web.

In this thesis, a noise Web page is defined as any Web page in a particular corpus that does not belong to a recognized genre within that corpus; if the entire Web were to be considered as a corpus, noise Web pages would then be all of those Web pages that do not belong to any genre which the Web page genre classifier had been trained to recognize. One of the shortcomings of the existing research on Web page classification is that it has, for the most part, been carried out on corpora which do not contain any noise. There are, however, some exceptions. Shepherd et al. [119], for example, introduced noise pages in their classification of homepages. Shepherd et al. used a neural net classifier to distinguish homepages from non-homepages and to classify those homepages as belonging to the PERSONAL HOMEPAGE, CORPORATE

HOMEPAGE, or ORGANIZATION HOMEPAGE genre. They found that when noise pages were introduced, the performance of the classifier deteriorated. Levering et al. [71] added 798 noise Web pages to a three genre corpus containing 501 single-label Web pages. They found that textual features alone performed very poorly in the presence of noise, but that the addition of HTML features dramatically improved performance; for one genre, the classification accuracy using the combination of textual and HTML features was higher in the presence of noise pages than it was without them. Levering et al. also found that the usefulness of visual features, in combination with the textual and HTML features, was extremely genre dependent in the presence of noise. In other research, Dong et al. [40, 41] also incorporated noise pages in their corpus, however they did not comment on the effect of the noise pages on the classification accuracy.

Another shortcoming of the existing research on Web page genre classification is that much of the research has been carried out on relatively balanced corpora, in which each genre contains approximately the same number of Web pages. Although there is no way of determining the exact distribution of genres on the Web, many features of the World Wide Web seem to be governed by Zipf's law [8], and it is not unlikely that the distribution of Web page genres could also follow a Zipfian distribution in which there are relatively few genres of very large size, and many genres of very small size. Thus, in addition to being able to perform well when encountering Web pages which belong to multiple genres, or which belong to no recognized genre, a successful Web page genre classifier should also perform well when encountering a collection of genres that have extremely unbalanced distributions.

Santini [106] developed three feature sets that included different combinations and numbers of features that included HTML tags, part-of-speech tags, common word frequencies, certain punctuation symbols, genre-specific facets, and attributes such as the length of the Web page. The Web page genre classification was performed using a support vector machine classifier which implemented sequential minimal optimization for training the support vectors. This method gave high accuracy on the balanced corpora tested, but Santini found that the performance was disappointing when these classification models were tested on a random, unclassified Web page collection. She concluded that when real world conditions are simulated, automatically classifying Web pages by genre using machine learning is very difficult, particularly when a single-genre classification scheme is used [101, 104, 106].

In an exploration of the use of Web page genre classification under real world conditions, a research team led by Benno Stein at the Bauhaus University introduced WEGA (**WE**b **G**enre **A**nalysis), which is a genre add-on for the Mozilla Firefox search engine. Although the system is still under development, WEGA is available for download [7]. This add-on supplies genre labels for each list of search results returned by the Firefox search engine, using a multi-label scheme in which each Web page can be assigned to more than one genre. The labels provided by WEGA include the genres ARTICLE, DISCUSSION, DOWNLOAD, HELP, LINK LIST, PORTRAYAL (NON-PRIVATE), PORTRAYAL (PRIVATE), and SHOP, as well as non-genre labels such as NON CLASSIFIABLE, UNSUPPORTED LANGUAGE and OFFLINE. WEGA was trained using the well-known KI-04 and 7-Genre corpora, which are described in detail in Sections 3.1.1 and 3.1.2. In a preliminary evaluation using genre-annotated Web pages from other corpora, Santini and Rosso [111] found that WEGA's genre classification performance was below 50%; for approximately 56% of the Web pages tested, none of the genre labels assigned to the Web pages by WEGA matched any of the genre labels assigned to the Web pages by human annotators. Santini and Rosso concluded that this preliminary evaluation of WEGA indicates that there is a wide gap between the results of testing a genre classifier on specific corpora in the laboratory and testing it under real-world conditions.

This section has pointed out some of the challenging real world conditions that a Web page genre classifier is likely to encounter, such as Web pages that belong to more than one genre, noise Web pages that do not belong to any recognized genre, and the unbalanced nature of the Web. The research conducted for this thesis addresses these issues by developing a Web page genre classifier that performs well on a multi-label corpus, as well as on noisy and unbalanced corpora; experiments with these corpora are discussed in Chapter 6.

2.6 Summary

This chapter has discussed the major issues and existing research in the field of Web page genre classification. Although researchers differ on precisely how to define the concept of Web page genres, they seem to agree on the potential of Web page genre as a means of classification. The subject of how Web page genres emerge, evolve, and multiply has been a topic of much interest, engendering a number of user studies and surveys. These studies provide rich insight into the topic of genre evolution on the Web. Although the focus of this thesis is on the automatic classification of Web pages by genre, rather than on the evolution and development of Web page genres, it is vital to recognize the importance of building a genre classification model that is flexible enough to accommodate the perpetual growth and development of Web page genres.

An important step in classifying Web pages by genre is the identification of the attributes and features with which to represent the Web pages; this chapter has given an overview of the document and Web page representations used in genre classification. Because this thesis focuses on representing a Web page using byte and character n-grams, particular attention has been paid to existing classification models that use n-gram based representations.

Regardless of the type of features that are chosen to represent a Web page, the dimensionality of this feature space must be kept relatively small in order to apply most machine learning algorithms, thus dimensionality reduction is also an issue that must be considered when classifying Web pages by genre. A popular approach to feature set reduction involves the use of feature selection measures such as Information Gain and the χ^2 statistic, but alternatives, such as handpicking a very limited feature set or using only the first fragment of a Web page so that further reduction is unnecessary, were also discussed. The approach in this thesis is to use a feature selection measure to select the top ranked features. The performance of three such measures, frequency, Information Gain, and the χ^2 statistic, is evaluated in Section 5.2.

This chapter has also discussed some of the challenging real word conditions that should be considered when designing and testing a Web page genre classifier. The research conducted for this thesis addresses these issues by developing a Web page genre classifier that performs well under a variety of conditions, such as on a multi-label corpus, and on noisy and unbalanced corpora.

Chapter 3

Background Information

This chapter provides background information that is integral to the remainder of the thesis. Section 3.1 discusses some of the limitations of existing Web page corpora, and describes the characteristics of the Web page corpora chosen for the experiments to be discussed in this thesis. Section 3.2 provides details about the cross-validation procedure and evaluation metrics, while Section 3.3 reviews the statistical tests and terms to be used in the statistical analysis of the experimental results. Descriptions of the χ^2 statistic and Information Gain feature selection techniques are found in Section 3.4, while Section 3.5 gives definitions of word, character, and byte n-grams.

3.1 Web Page Genre Corpora

Web page genre classification is a relatively new field of research, and a Web page genre benchmark corpus does not yet exist. Recently Rehm et al. [93] presented plans for an international and interdisciplinary collaboration to construct a reference corpus of Web page genres, in which the Web pages will be annotated with multi-level genre tags. They note that the creation of such a corpus is a very challenging task; it raises issues such as what level(s) of abstraction, or granularity, should be used for the genres, what genre labels would be most recognizable to Web users, and what number of genres should be included. A number of experiments and user studies have investigated these questions; for example Santini's experiments [101] suggest that classifier performance is enhanced when a corpus is built with a consistent level of abstraction, or granularity. Roussinov et al. [100] investigated which genres best fit users' needs, and although they identify 116 different genres, they suggest that a large number of search tasks could be satisfied with as few as five groups of related genres. A user study conducted by Meyer zu Eissen and Stein [84] investigated which genres are considered useful by search engine users, and resulted in a list of eight suggested genres. Rosso [99] also explores the question of which genres should be used, and notes that the number of genres cannot be both exhaustive and useful for searching; based on user studies, Rosso has developed a list of 18 useful and recognizable Web page genres. Despite the insight provided by these user studies, there remain many open questions that must be resolved before a reference corpus of Web page genres can be constructed. Both Rosso [99] and Rehm et al. [93] also note that because there is no way of determining the distribution of genres on the Web, there is no accurate method for determining the appropriate distribution of genres within a Web page genre corpus.

Currently, the best option for Web page genre researchers is to cross-test their classifiers over several existing Web genre collections, which is the method used in the research conducted for this thesis. The Web page corpora for this research have been carefully selected to fulfill two goals. The first goal is to provide a solid basis for evaluating the experimental results by comparing them with those of other researchers, therefore established corpora for which published results are either available or anticipated are used. The second goal is to test the new genre classification models on Web page collections that have a variety of characteristics. The balance, granularity, and labeling scheme are all expected to have an appreciable impact on the performance of the genre classifier. A classifier that works well on a perfectly balanced, evenly grained, single-label corpus may show very different performance on, for example, an unbalanced, multi-label corpus.

Four Web page collections, created and made available by other researchers, have been used in the research for this thesis. These are the 7-Genre corpus, the KI-04 corpus, the 20-Genre corpus, and the Syracuse corpus. In addition to these collections, experiments are also run on a combination of the 7-Genre and KI-04 corpora, which will be referred to as the 15-Genre corpus. Detailed descriptions of these collections are found in Sections 3.1.1–3.1.5. Although some genre researchers, such as Jebari and Ounalli [53, 54], perform Web page genre experiments on the WebKB corpus [6], the corpus was not chosen for the research conducted for this thesis. The Web pages from the WebKB corpus, which were collected from Computer Science departments of various universities, are labeled as belonging to particular categories, but these categories (STUDENT, FACULTY, STAFF, DEPARTMENT, COURSE, PROJECT, and OTHER) are not necessarily Web page genres. For this reason, the WebKB corpus was excluded from this research in favor of Web page corpora that were specifically constructed for the purpose of Web page genre classification.

Although some researchers, such as Gupta et al. [49], Mehler et al. [83], and Symonenko [131], investigate genres at the Website level, the unit of analysis for each of the corpora used in this thesis is the individual Web page. Each collection contains Web pages labeled by genre, however the collections differ somewhat in the level of abstraction, or granularity, of the genres which they include. Genres can range from being quite narrow to being very broad; for example, they can range from the relatively narrow subgenre PRESS RELEASE to the broader genre NEWS STORIES, to the even broader supergenre JOURNALISTIC. Santini [101, 106] discusses genre granularity in terms of a three tiered hierarchy with subordinate, basic, and superordinate levels. She observes that the basic level (the genre level) is the level at which concepts are most easily recognized, learned, and remembered. The 7-Genre and KI-04 collections have a consistent, basic level of abstraction, but the 20-Genre and Syracuse collections vary in their levels of abstraction. For example, the Syracuse corpus has a lower level of abstraction than the 20-Genre corpus; the former includes genres such as ENCYCLOPEDIA ENTRY, RECIPE, and TUTORIAL AND HOW-TO, whereas the latter considers these to be subordinate to (and therefore included in) the genre INFORMATIVE. Table 3.1 gives a comparison of some of the characteristics of the Web page collections chosen for the research in this thesis, while Sections 3.1.1–3.1.5 describe the collections in more detail.

3.1.1 7-Genre Corpus

The 7-Genre, or 7-Web-genre collection, was constructed by genre researcher Marina Santini; the Web pages were downloaded for the collection in early 2005. The collection is described by Santini in [101] and [106], and is available online [2]. This corpus contains 1400 English Web pages, and is evenly balanced with 200 Web pages in each of seven genres. These genres are BLOG, ESHOP, FAQ, ONLINE NEWSPAPER FRONT PAGE, LISTING, PERSONAL HOMEPAGE, and SEARCH PAGE. The granularity of the collection is consistent, with the exception of the LISTING genre, which can be decomposed into the subgenres CHECKLIST, HOTLIST, SITEMAP, and TABLE. The PERSONAL HOMEPAGE genre, although it does not have specifically labeled subgenres,

Corpus Name	Number of Web Pages	Number of Genres	Web Page Balance	Labeling Style	Level of Abstraction
7-Genre	1400	7	perfectly balanced	single	medium
KI-04	1205	8	slightly unbalanced	single	medium
15-Genre	1611	15	highly unbalanced	single	low/medium mixture
20-Genre	1539	20	unbalanced	multi	medium/high mixture
Syracuse	1985	24	highly unbalanced	single	low/medium/high mixture

Table 3.1: Web page corpus characteristics.

includes a variety of types of personal homepages, such as academic and administrative personal homepages [106]. Each Web page in the corpus is labeled with one, and only one, genre label. The experiments conducted for this thesis with this corpus use 10-fold cross-validation, in order to provide robustness against overfitting and give additional strength to the statistical analysis.

3.1.2 KI-04 Corpus

The KI-04 corpus was constructed by Sven Meyer zu Eissen, using eight genres suggested by participants in a user study on the usefulness of Web page genres [84]; the Web pages were downloaded for the collection in January, 2004. The original corpus includes 1295 Web pages, but 90 of these are empty pages. Following the lead of Jebari and Ounalli [53, 54], Kanaris and Stamatatos [56], and Santini [101, 106], only the 1205 non-empty pages are used in the research for this thesis. Both versions of the collection are available online [2]. Meyer zu Eissen and Stein used a perfectly balanced 800 page subset of this collection in their work [84]. The 1205 page collection is somewhat less balanced, with the number of Web pages in each genre ranging from 126 to 205. Table 3.2 lists the eight genres and the number of Web pages in each genre. Note that each Web page in this corpus is labeled with one, and only one, genre label. As with the 7-Genre corpus, 10-fold cross-validation is used in the experiments with this corpus, in order to provide robustness against overfitting and give additional strength to the statistical analysis.

Genre	Number of Web Pages
PERSONAL HOMEPAGE	126
ARTICLE	127
DISCUSSION	127
HELP	139
DOWNLOAD	151
NON-PERSONAL HOMEPAGE	163
SHOP	167
LINK COLLECTION	205

Table 3.2: Genre densities for the KI-04 corpus.

The following description of the genres in the KI-04 corpus is based on the corpus description given by Meyer zu Eissen and Stein [84]. Note that some genre names have been updated: PERSONAL HOMEPAGE was originally PORTRAYAL (PRIVATE), NON-PERSONAL HOMEPAGE was previously PORTRAYAL (NON-PRIVATE), and LINK COLLECTION was originally referred to as LINK LIST.

- PERSONAL HOMEPAGE: typical private homepages with informal content.
- ARTICLE: documents with long passages of text, such as research articles, reviews, technical reports, or book chapters.
- DISCUSSION: all pages that provide forums, mailing lists, or discussion boards.
- HELP: all pages that provide assistance, such as Q&A or FAQ pages.

- DOWNLOAD: pages on which freeware, shareware, demo versions of programs, etc., can be downloaded.
- NON-PERSONAL HOMEPAGE: Web appearances of companies, universities, and other public institutions.
- SHOP: all kinds of pages whose main purpose is product information or sale.
- LINK COLLECTION: documents that consist mainly of lists of links.

3.1.3 15-Genre Corpus

There is no way of determining the exact distribution of genres on the Web, and therefore no accurate method for determining the appropriate distribution of genres within Web page genre corpora [93, 99]. However, many features of the World Wide Web seem to be governed by Zipf's law [8]. Zipf's law can be expressed as a power law, meaning that the probability of attaining a certain size x is proportional to $x^{-\tau}$, where τ is greater than or equal to 1. Zipf's law models the occurrence of distinct objects in particular sorts of collections, for example word frequencies in text [148] and city sizes [149]. According to Zipf's law, the probability of the occurrence of words or other items starts high and tapers off, thus only a few items occur very often while many others occur very rarely. If the Zipf curve is plotted on a log-log scale, it appears as a straight line with a slope of -1.

Adamic and Huberman [8] note that many features of the Web appear to follow Zipf's law [8]; for example only a few sites consist of millions of pages, but millions of sites contain only a handful of pages. Web access statistics and Internet traffic characteristics also have Zipfian distributions [8, 9]. Based on the number of Zipfian distributions associated with the Web, it is not unreasonable to speculate that the distribution of Web page genres could also follow a Zipfian distribution in which there are relatively few genres of very large size, and many genres of very small size. With this in mind, the 15-Genre corpus was created for this thesis by combining the 7-Genre and KI-04 collections as follows. First, the ESHOP and SHOP genres were merged, as were the three HOMEPAGE genres. Next, the LISTING genre was subdivided into its four subgenres of equal size: CHECKLIST, HOTLIST, SITEMAP, and TABLE. This resulted in a total of 15 genres ranging in size from 489 to 50; these genres, other than the largest, were then reduced in size to give the corpus a Zipfian distribution. The genre sizes for this Zipfian distribution were chosen using the formula $f = k/r$ as a guide, where f is the frequency, (number of Web pages per genre) k is the constant (in this case 489), and r is the rank of the genre. Figure 3.1 gives a log-log scale plot of the resulting distribution and Table 3.3 lists the 15 genres and the number of Web pages used in each genre. Because of the small size of some of the genres, 3-fold, rather than 10-fold, cross-validation is used for the experiments with this corpus.

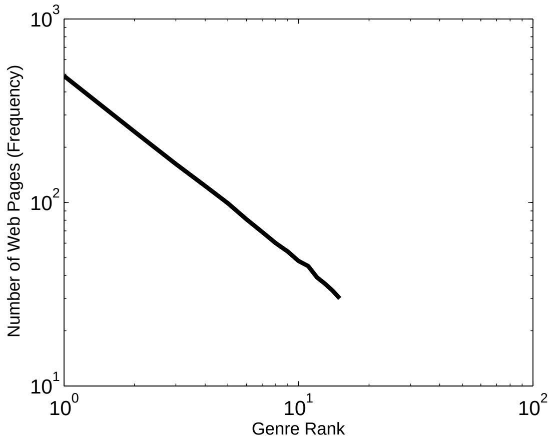

Figure 3.1: Log-log scale plot of the Web page distributions for the 15-Genre corpus. The Web page frequency per genre is on the y-axis. The rank of the genre, from 1 to 15, is on the x-axis, where 1 is the rank of the largest genre and 15 is the rank of the smallest genre.

3.1.4 20-Genre Corpus

The 20-Genre corpus was constructed by Mitja Luštrek and Andrej Bratko at the Jožef Stefan Institute, and is available online [1]. The creation of the corpus is described by Vidulin et al. [137]. The genres were chosen with the intention of covering the whole Internet, but the creators of the collection found that the Web page genres were not clearly delineated, and that many Web pages belong to more than one genre, either because all of the Web page spans multiple genres, or because the Web page has sections belonging to different genres. Luštrek and Bratko therefore chose a multi-label approach. This collection contains 1539 English Web pages, each with one or more genre labels. Of the 1539 Web pages, 1059 have one genre label, 438 have two genre labels, 39 have three labels, and 3 have four labels. This gives a total of 2064

Genre	Number of Web Pages
TABLE	30
SITEMAP	33
HOTLIST	36
CHECKLIST	39
DISCUSSION	45
ARTICLE	48
HELP	54
DOWNLOAD	60
SEARCH PAGE	69
ONLINE NEWSPAPER FRONT PAGE	81
FAQ	99
BLOG	123
LINK COLLECTION	162
SHOPPING	243
HOMEPAGE	489

Table 3.3: Genre densities for the 15-Genre corpus.

labels. The number of Web pages in each genre ranges from 55 to 227. Table 3.4 lists the 20 genres and the number of Web pages that belong to each genre; Web pages with more than one label are included in each genre for which they have been given a label. Because of the small size of some of the genres, 3-fold, rather than 10-fold, cross-validation is used for the experiments with this corpus.

The following description of the genres in the 20-Genre corpus is condensed from the documentation provided with the collection [1] and from the genre descriptions given by Vidulin et al. [137]. Note that the ADULT genre was originally referred to as PORNOGRAPHIC by Vidulin et al. [137], but was modified to ADULT in the corpus documentation and in other publications [135, 136].

Genre	Number of Web Pages
OFFICIAL	55
SHOPPING	66
PROSE FICTION	67
ADULT	68
FAQ	70
POETRY	72
ENTERTAINMENT	76
SCIENTIFIC	76
BLOG	77
GATEWAY	77
ERROR MESSAGE	79
COMMUNITY	82
USER INPUT	84
CHILDREN'S	105
PERSONAL	113
COMMERCIAL/PROMOTIONAL	121
CONTENT DELIVERY	138
JOURNALISTIC	186
INFORMATIVE	225
INDEX	227

Table 3.4: Genre densities for the 20-Genre corpus.

- OFFICIAL: legal materials, official reports, and rules.
- SHOPPING: Web pages selling goods or services online.
- PROSE FICTION: short stories, narratives, etc.
- ADULT: pornographic pictures, videos, and stories.

- FAQ: frequently asked questions (and answers).
- POETRY: poems and lyrics.
- ENTERTAINMENT: jokes, puzzles, horoscopes, games, etc.
- SCIENTIFIC: books, papers, theses, and lecture notes, for a specialized audience.
- BLOG: Web logs, journals, and pages containing time-stamped updates.
- GATEWAY: introductory pages, redirection pages, and login pages.
- ERROR MESSAGE: HTTP and non-HTTP error pages.
- COMMUNITY: dedicated multi-party correspondence such as forums.
- USER INPUT: forms and surveys soliciting input.
- CHILDREN'S: Web pages with content specifically suited for children.
- PERSONAL: homepages of one or more persons acting as individuals.
- COMMERCIAL/PROMOTIONAL: presentations of institutions, products, press releases, etc.
- CONTENT DELIVERY: download pages, image and movie galleries, and games.
- JOURNALISTIC: news reports, editorials, interviews, reviews.
- INFORMATIVE: encyclopedic materials, recipes, user manuals, how-tos, lecture notes for a wide audience, informative books, and biographies.
- INDEX: link collections and tables of contents; an index page is a page with lots of links to other Web pages, and little else.

3.1.5 Syracuse Corpus

The Syracuse corpus was assembled by a team of researchers led by Barbara Kwasnik and Kevin Crowston at Syracuse University, as part of a study to determine how providing genre metadata can help with accessing information sources in a digital environment [68]. Kwasnik et al. note that in constructing the corpus, much of the rich genre information they collected was either too difficult to represent, or had to be pared away [68]. The resulting corpus is roughly organized as a shallow hierarchy, with a level of abstraction, or granularity, that varies from broad supergenres such as ADVERTISING to narrow subgenres such as BIOGRAPHICAL TIMELINE. The collection contains a total of 2748 labeled Web pages, each of which has one, and only one, genre label. There are 118 different genres represented, with the number of Web pages in each genre ranging from 1 to 350. Of these, only 24 of the genres contain 30 or more Web pages each. Because genres with fewer than 30 Web pages do not allow for an adequate representation of the genre, particularly when k-fold cross-validation is used, the experiments conducted for this thesis use a subset of the Syracuse corpus consisting of the 24 genres that contain at least 30 Web pages; the remainder of the labeled Web pages from the corpus are reserved for use as noise Web pages. Within the 24 genre subset there are 1985 Web pages, with the number of Web pages per genre ranging from 30 to 350. Table 3.5 gives a list of the 24 largest genres and the number of Web pages in each of these genres, and Figure 3.2 shows a log-log scale plot of these genre densities; the plot has a Zipf-like distribution in which there are a few genres with many Web pages, and many genres with very few Web pages. Because of the small size of some of the genres, 3-fold cross-validation, rather than 10-fold cross-validation, is used in the experiments with this corpus.

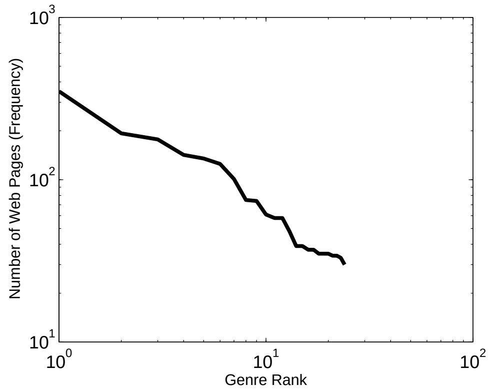

Figure 3.2: Log-log scale plot of the Web page distributions for the Syracuse corpus. The Web page frequency per genre is on the y-axis. The rank of the genre, from 1 to 24, is on the x-axis, where 1 is the rank of the largest genre and 24 is the rank of the smallest genre.

Genre	Number of Web Pages
INDEX TO MISCELLANEOUS RESOURCES	30
TABLE OF CONTENTS	33
BIBLIOGRAPHIC RECORD	34
BIOGRAPHY	34
DEFINITION/DESCRIPTION	35
ENCYCLOPEDIA ENTRY	35
TUTORIAL AND HOW-TO	35
OTHER BIOGRAPHY	37
LESSON PLAN	37
DIRECTORY OF COMPANIES	39
PRESS RELEASE	39
ABOUT A PROGRAM	48
ABOUT AN ORGANIZATION	58
FACTS-AND-FIGURES PAGE	58
DISCUSSION FORUM	61
COMPANY/ORGANIZATION HOMEPAGE	74
ADVERTISING	75
MAGAZINE ARTICLE	101
RECIPE	125
BLOG	135
DIRECTORY OF RESOURCES/LINKS	142
OTHER ARTICLE	177
NEWS STORY	193
PRODUCT/SERVICE DESCRIPTION PAGE	350

Table 3.5: Genre densities for the Syracuse corpus.

3.2 Cross-Validation and Evaluation Metrics

3.2.1 *k*-fold Cross-Validation

All of the experiments conducted for this thesis are run using k-fold cross-validation. With this method, the Web pages in a particular corpus are partitioned randomly into k groups of equal size and distribution. For each of the k cross-validation iterations, one of the partitions is used as the test (validation) set, and the other $k-1$ partitions make up the training set; over the k iterations, each of the k partitions is used exactly once as the test set. The results for all of the iterations are then averaged to give the final results. The advantage of this method is that it produces a relatively unbiased estimate of the accuracy of the classifier. This estimate can be highly variable, however the variance in the accuracy for each iteration decreases as k increases. A disadvantage of the k-fold cross-validation method is that the sample size must be large enough that division of the data into subsets is meaningful (i.e., each subset has the same distribution of data). Although the k-fold cross-validation method is computationally expensive because the training algorithm has to be run from scratch k times, it allows the simulation of k experiments, thus increasing the strength of the statistical analysis.

3.2.2 Precision, Recall, and F1-measure

A variety of evaluation metrics may be used to evaluate the performance of classification algorithms; different measures evaluate different characteristics of the classifier. Precision, for example, is a measure of the purity or exactness of a classification. In the case of Web page classification by genre, precision reports the proportion of Web pages classified as belonging to particular genre that actually do belong to the genre, without taking into account the number of Web pages from that genre that have been incorrectly classified. The recall metric, on the other hand, is a measure of the completeness of a classification. With respect to Web page classification by genre, recall reports the proportion of the Web pages of a particular genre that were correctly classified, without regard to the number of Web pages that have been incorrectly assigned to that genre. The F1-measure, also known as the balanced F-score or F-measure, is the harmonic mean of precision and recall.

In a problem such as Web page classification, precision and recall can be defined in terms of true positives, false positives, and false negatives. For each genre, a true positive is a Web page that the classifier correctly labels as belonging to that genre, whereas a false positive is a Web page that the classifier incorrectly labels as belonging to that particular genre. A false negative is a Web page that the classifier fails to recognize as belonging to the particular genre, and incorrectly assigns to another genre. A true negative is a Web page correctly recognized as not belonging to a particular genre. Table 3.6 shows a confusion matrix for a binary classification problem.

	Predicted Class	
True Class	Positive	Negative
Positive	True Positive	False Negative
Negative	False Positive	True Negative

Table 3.6: Confusion matrix for a binary classification problem.

For evaluating the binary classification case, such as whether or not a document is relevant to an information retrieval query, precision, P, and recall, R, are defined as follows.

$$P = \frac{Tp}{Tp + Fp} \quad \text{and} \quad R = \frac{Tp}{Tp + Fn}, \tag{3.1}$$

where Tp is the number of true positives, Fp is the number of false positives, and Fn is the number of false negatives. For the situation in which there are a number of different classes, or in this case, genres, precision and recall can be micro-averaged or macro-averaged. Micro-precision, P_{mi}, and micro-recall, R_{mi}, are defined as follows.

$$P_{mi} = \frac{\sum\limits_{i=1}^{|G|} Tp}{\sum\limits_{i=1}^{|G|} (Tp + Fp)} \quad \text{and} \quad R_{mi} = \frac{\sum\limits_{i=1}^{|G|} Tp}{\sum\limits_{i=1}^{|G|} (Tp + Fn)}, \tag{3.2}$$

where $|G|$ is the number of genres. Macro-precision, P_{ma}, and macro-recall, R_{ma}, are defined as follows.

$$P_{ma} = \frac{1}{|G|}\left(\sum_{i=1}^{|G|}\frac{Tp}{(Tp+Fp)}\right) \quad \text{and} \quad R_{ma} = \frac{1}{|G|}\left(\sum_{i=1}^{|G|}\frac{Tp}{(Tp+Fn)}\right). \tag{3.3}$$

Micro-precision and micro-recall give equal weight to each Web page, whereas macro-precision and macro-recall give equal weight to each genre. The F1-measure, the harmonic mean of precision and recall, is defined as follows.

$$\mathit{F1} = \frac{(2 \times \mathit{precision} \times \mathit{recall})}{(\mathit{precision} + \mathit{recall})}. \tag{3.4}$$

For the case in which every Web page belongs to one, and only one genre, the micro-precision, micro-recall, and micro-F1-measures will all be equal; if it is also the case that every genre contains exactly the same number of Web pages, then the macro-recall will also be equal to the micro-precision, micro-recall, and micro-F1-measures.

3.2.3 Classification Accuracy

In a classification problem, accuracy is the proportion of true results; in the case of Web page genre classification, this is the proportion of correctly classified Web pages. Accuracy can be defined in terms of true positives, false positives, false negatives, and true negatives. For each genre, the classification accuracy, *Acc* can be defined as follows.

$$\mathit{Acc} = \frac{(Tp+Tn)}{(Tp+Tn+Fn+Fp)}, \tag{3.5}$$

where Tp is the number of true positives, Tn is the number of true negatives, Fn is the number of false negatives, and Fp is the number of false positives. As with precision and recall, the classification accuracy can be micro-averaged or macro-averaged. The micro-averaged classification accuracy, Acc_{mi}, and the macro-averaged classification accuracy, Acc_{ma}, are defined as follows.

$$Acc_{mi} = \frac{\sum_{i=1}^{|G|} (Tp + Tn)}{\sum_{i=1}^{|G|} (Tp + Tn + Fn + Fp)}, \tag{3.6}$$

and

$$Acc_{ma} = \frac{1}{|G|} \left(\sum_{i=1}^{|G|} \frac{(Tp + Tn)}{(Tp + Tn + Fn + Fp)} \right), \tag{3.7}$$

where $|G|$ is the number of genres.

3.3 Statistical Tests

In order to interpret the experimental results of the research conducted for this thesis, the data are subjected to statistical analysis. The statistical tests and terms that will be used in the discussion of these results are described in Sections 3.3.1 and 3.3.2.

The statistical tests conducted as part of this thesis were run using the Statistical Package for the Social Sciences (SPSS) software, which has since been re-branded as the Predictive Analytics SoftWare (PASW) [3]. A brief description of the steps taken in running these tests using the SPSS software is given in Appendix A.

3.3.1 Analysis of Variance

The statistical test known as analysis of variance (ANOVA) is based on the concept of analyzing the variance that appears in groups of data by testing differences in the means for statistical significance [18]. The total observed variance is partitioned according to the factors assumed to be responsible for producing that variation in the data. This is accomplished by partitioning the total variance into the component that is due to true random error, and the components that are due to differences between means. These latter variance components, or sources of variance, are then tested for statistical significance. If the difference is significant, the null hypothesis of no differences between the means is rejected, and the alternative hypothesis, that the means are different from each other, is accepted. ANOVA also allows the detection of interaction effects between variables. When a significant effect is reported, it is typically expressed in terms of a p-value. The p-value refers to the probability (ranging from zero to one) that the observed results (or results more extreme) could have occurred by chance, if in reality the null hypothesis is true. A small p-value leads to the conclusion that the null hypothesis of no differences between the means is rejected; a p-value of less than 0.05 is considered to indicate statistical significance, and is reported in the form $p < 0.05$. When the p-value indicates statistical significance, the smaller this value is, the more confidence can be given to the corresponding interpretation of the results [18].

The effect size is the degree to which changing an independent variable (such as n-gram length) affects the value of a dependent variable (such as classification accuracy) [18]. The partial Eta squared (η^2) indicates the proportion of the total variability that is attributable to a particular factor, or independent variable, when controlling for other factors.

3.3.2 Scheffé Post Hoc Test

A statistically significant effect in ANOVA is often succeeded with a follow-up, or post hoc test. This can be done in order to assess which groups are different from which other groups, or to test various other hypotheses. Post hoc tests such as Scheffé's test most commonly compare every group mean with every other group mean. For the statistical analysis in this thesis, Scheffé's test is used because it is considered to be very conservative, meaning that it is more difficult to achieve statistical significance with this test than with less conservative tests, such as the Tukey or Duncan post hoc tests [18].

3.4 Theoretically Based Feature Selection Measures

As discussed in Section 2.4, it is often necessary to use feature selection techniques to select a subset of relevant features with which to represent documents for classification, because most standard machine learning techniques cannot be directly applied when the dimensionality is very high [10]. Two theoretically based feature selection measures that will be investigated as part of the research for this thesis include the χ^2 statistic and Information Gain. The χ^2 statistic is a statistically based measure, while Information Gain is based on information theory.

3.4.1 χ^2 Statistic

The χ^2 statistic is a statistical measure of the dependence between a term and a category, or for the purposes of this thesis, between an n-gram, m, and a genre, g. There are two types of χ^2 tests: the test for goodness of fit, and the test for independence. In this thesis, as in Yang and Pedersen [142] and Dong et al. [40], the χ^2 test for independence is used.

Table 3.7 gives the notation for the contingency table presenting the joint distribution of the two variables, where A is the number of times n-gram m and genre g co-occur, B is the number of times g occurs without m, C is the number of times m occurs without g, D is the number of times neither g nor m occurs, and N is the total number of documents. For the work to be discussed in this thesis, the number of occurrences of n-gram m is defined as the number of Web pages containing m, and the number of occurrences of genre g is defined as number of Web pages belonging to genre g.

	n-gram m	$\neg$ n-gram m	Total
Genre g	A	B	A + B
$\neg$ Genre g	C	D	C + D
Total	A + C	B + D	A + B + C + D = N

Table 3.7: Contingency table for the χ^2 statistic.

The χ^2 statistic is defined as follows.

$$\begin{aligned}\chi^2(m,g) &= \frac{(A+B+C+D)\times(AD-CB)^2}{(A+C)\times(B+D)\times(A+B)\times(C+D)} \\ &= \frac{N\times(AD-CB)^2}{(A+C)\times(B+D)\times(A+B)\times(C+D)} \qquad (3.8)\end{aligned}$$

The χ^2 statistic has a value of zero if m and g are independent. The computation of χ^2 statistics has a quadratic complexity, based on the vocabulary size and the number of genres [142].

3.4.2 Information Gain

Information Gain is a measure, based on information theory, that estimates the information gained for the task of predicting a category, by knowing whether a particular term is present or absent in a document. The computation includes the estimation of conditional probabilities and has a quadratic complexity, based on the vocabulary size and the number of genres [142].

Let g_i define a genre from the set of genres G; then Information Gain, IG, for an n-gram m is defined as follows.

$$\begin{aligned}IG(m) = &- \sum_{i=1}^{|G|} P(g_i)\log P(g_i) \\ &+ P(m)\sum_{i=1}^{|G|} P(g_i|m)\log P(g_i|m) \\ &+ P(\bar{m})\sum_{i=1}^{|G|} P(g_i|\bar{m})\log P(g_i|\bar{m}), \qquad (3.9)\end{aligned}$$

where the following definitions hold.

$|G|$ = number of genres,

$$P\left(g_i\right) = \frac{\text{number of documents in } g_i}{\text{total number of documents}},$$

$$P\left(m\right) = \frac{\text{number of documents containing } m}{\text{total number of documents}},$$

$$P\left(\bar{m}\right) = 1 - \frac{\text{number of documents containing } m}{\text{total number of documents}},$$

$$P\left(m|g_i\right) = \frac{\text{number of documents in } g_i \text{ containing } m}{\text{number of documents in } g_i},$$

$$P\left(\bar{m}|g_i\right) = \frac{\text{number of documents in } g_i \text{ not containing } m}{\text{number of documents in } g_i},$$

$$P\left(g_i|m\right) = \frac{\text{number of documents in } g_i \text{ containing } m}{\text{number of documents containing } m}, \quad \text{and}$$

$$P\left(g_i|\bar{m}\right) = \frac{P\left(\bar{m}|g_i\right)P\left(g_i\right)}{P\left(\bar{m}\right)}.$$

3.5 Word, Character, and Byte n-grams

Although the use of n-grams is common in many areas of research, researchers are not always clear in defining what type of n-gram is being used, or in making the distinction between character and byte n-grams. The main focus of the work in this thesis is on the use of byte and character n-grams, therefore unless a specific distinction is made, the use of the term *n-gram* refers to byte and character n-grams. For this thesis, the following definitions hold.

An n-gram is an n-word, n-character or n-byte substring of a longer string, and the term is used for contiguous substrings only, meaning that the string is subdivided into a set of overlapping n-grams. More formally, given a sequence of tokens $S = (s_1, s_2, ..., s_{N+(n-1)})$ over the token alphabet A, where N and n are positive integers, an n-gram sequence S is any n-long subsequence of consecutive tokens. Note that there are N such n-grams in S, and there are $(|A|^n)$ unique n-grams over the alphabet A where $|A|$ is the size of A [132], therefore the complete set of 5-grams is, for example, much larger than the complete set of 3-grams. This exponential explosion is one of the drawbacks of n-gram use.

Byte n-grams are raw character n-grams in which no bytes are ignored, including the whitespace characters. Character n-grams use letters only, and all letters are converted to uppercase. A non-letter character, (a digit for example) is replaced by a space, and two or more contiguous spaces are replaced by a single space. An underscore character is used to represent the space. Word n-grams are are simply sequences of n contiguous words. Table 3.8 gives an example of the byte and character n-grams for a short string.

String	Use 3-grams!
Byte 3-grams	Use se_ e_3 _3- 3-g -gr gra ram ams ms!
Character 3-grams	USE SE_ E_G _GR GRA RAM AMS

Table 3.8: Example of byte and character n-grams.

When all of the n-grams used are the same length, they are referred to as fixed-length n-grams; otherwise, they are referred to as variable-length n-grams.

Byte and character n-grams can provide a low-level representation of documents and Web pages that has the advantage of being relatively tolerant of textual errors. Cavnar and Trenkle [22] explain that one of the main benefits of the use of n-gram based document representations is that, because every string is decomposed into small parts, any errors that are present tend to affect only a limited number of those parts. If all of the n-grams that are common to two strings are counted, the resulting measure of their similarity is resistant to a wide variety of textual errors, such as spelling errors. Cavnar and Trenkle also note that the use of n-gram representations essentially provides the effect of word stemming, in which words are reduced to their grammatical roots, at no additional cost. This is because n-grams for related forms of a word will automatically have much in common when viewed as sets of n-grams [22]. Cavnar and Trenkle further note that word stemmers such as Porter's stemmer [91] are language dependent, whereas n-gram extraction is language independent.

Tomovic et al. [132] give the following summary of the benefits of n-gram representation.

- robustness: relatively insensitive to spelling variations/errors.
- completeness: token alphabet known in advance.
- domain independence: language and topic independent.
- efficiency: one pass processing.
- simplicity: no linguistic knowledge is required.

3.6 Summary

This chapter has provided background information that is relevant to the remainder of the thesis. Section 3.1 discussed some of the limitations of existing Web page corpora, and described the characteristics of the Web page corpora chosen for the experiments to be discussed in this thesis. Section 3.2 provided details about the cross-validation procedure and evaluation metrics, while Section 3.3 reviewed the statistical tests and terms to be used in the statistical analysis of the experimental results. Section 3.4 gave descriptions of the χ^2 statistic and Information Gain feature selection techniques, and Section 3.5 defined word, character, and byte n-grams.

Chapter 4 will discuss studies which investigate the feasibility of using n-gram Web page and Web page genre representations for the task of Web page genre classification.

Chapter 4

n-gram Use for Web Page Genre Classification

This chapter discusses studies which investigate the feasibility of using *n*-gram representations of Web pages for the task of Web page genre classification. The hypothesis of this thesis is that a byte *n*-gram representation of a Web page can be used effectively to classify the Web page by its genre(s), and the goal of this thesis is to develop a simple, easily scalable method for Web page classification using an *n*-gram based approach. This chapter explores the two general research questions of how Web pages and Web page genres should be represented, and what classification method should be used. Specifically, the studies discussed in this chapter answer the questions of whether Web pages of the same genre share a distribution of n-grams that is similar enough to allow n-gram representations of the Web pages to be used for Web page genre classification, whether there exists a set of parameters for the proposed approach, with regard to *n*-gram type, length, and number, that will allow state-of-the-art classification of Web pages by their genre, and whether an appropriate distance measure be found that computes the similarity between Web page and Web page genre profiles, leading to a classification method based on the nearest profile.

The idea is to generate a set of *n*-gram frequency profiles from the training data to represent each of the genres. When a Web page from the test data is classified, the classification model first computes its *n*-gram frequency profile. It then compares this profile to the profiles that have been generated from the training set, using an easily calculated distance measure. The model classifies the Web page as belonging to the genre of the profile from the training set to which its profile is closest.

A number of different distance measures are also explored, in order to determine whether the specific distance measure used has an appreciable impact on the classification performance of the model, and if so, which distance measure (of those tested) achieves the best performance. Additionally, a comparison between the use of byte and character *n*-grams is made, in order to determine whether the type of *n*-gram used in the Web page representations has an impact on the performance of the classification model, and if so, which of these two n-gram types allows the model to achieve the highest classification performance. A comparison is also made between a byte n-gram Web page representation and a bag-of-words representation.

For these experiments, each Web page is preprocessed to remove all HTML tags and JavaScript code, however no stemming of terms is performed, which means that the preprocessing steps are not language dependent. After the preprocessing, the remaining textual content of each Web page is used to form an n-gram representation of the Web page. This n-gram representation is a profile consisting of the L most frequent fixed-length byte or character n-grams and their normalized frequencies within the document; the n-gram frequencies are normalized by dividing the frequencies by the total number of n-grams, of the given length, in the document. These n-gram profiles are produced using the `Perl` package `Text:Ngrams` [5]. The experiments were run on the 7-Genre corpus, which is a perfectly balanced corpus with 1400 single-label Web pages and seven genres.

Sections 4.1 and 4.2 describe two of the classification approaches that were tested, and a comparison of the results of these approaches is given in Section 4.3; see Mason et al. [79] for a more detailed discussion of this work. Section 4.4 explores the use of character n-grams, and compares the results with those of the experiments in which byte n-grams were used. Section 4.5 compares the results of a bag-of-words approach with those of the experiments in which byte n-gram Web page representations were used. Section 4.6 gives an overview of the distance measures that were investigated, while Section 4.7 provides a summary of the experiments and discusses the conclusions that were drawn.

4.1 k-Nearest Neighbor Model

The k-Nearest Neighbor (k-NN) classification algorithm is one of the simplest and best known machine learning algorithms, therefore initial experiments were carried out using this classification model. With this approach, a document is classified based on the majority vote of its neighbors, with the document being assigned to the most common class amongst its k nearest neighbors, where k is a positive integer [10]. A distance measure is used to determine the distance between the document and its neighbors. For the experiments to be discussed in this section, results are compared for the traditional Euclidean and Manhattan distance measures, as well as for a distance measure based on the arithmetic mean. This measure, which will be referred to as the Kešelj distance measure, was suggested by Kešelj et al. in their paper on the use of n-gram profiles for authorship attribution [62]. These distance measures, d_e, d_m, and d_k respectively, are defined as follows.

$$d_e(P_1, P_2) = \sqrt{\sum_{m \in (\mathrm{P}_1 \cup \mathrm{P}_2)} (f_1(m) - f_2(m))^2}, \tag{4.1}$$

$$d_m(P_1, P_2) = \sum_{m \in (\mathrm{P}_1 \cup \mathrm{P}_2)} |f_1(m) - f_2(m)|, \qquad \text{and} \tag{4.2}$$

$$\begin{aligned} d_k(P_1, P_2) &= \sum_{m \in (\mathrm{P}_1 \cup \mathrm{P}_2)} \left(\frac{f_1(m) - f_2(m)}{\frac{f_1(m) + f_2(m)}{2}} \right)^2 \\ &= \sum_{m \in (\mathrm{P}_1 \cup \mathrm{P}_2)} \left(\frac{2 \cdot (f_1(m) - f_2(m))}{f_1(m) + f_2(m)} \right)^2, \end{aligned} \tag{4.3}$$

where $f_1(m)$ and $f_2(m)$ are the frequencies of n-gram m in the profiles P_1 and P_2 respectively. These measures always return a non-negative number, and for two identical profiles, give a distance of zero.

The choice of k is important to the k-NN classification algorithm. Choosing too small a value of k can lead to a large variance in the predictions, whereas choosing too large a value of k may lead to underfitting. Preliminary experiments using $k = 1$, $k = 3$, and $k = 5$ showed that the best classification results were always achieved when $k = 1$, therefore the results given for the k-NN model are for the case in which $k = 1$. It is somewhat surprising that $k = 1$ gave the best performance in the preliminary experiments, however this could be at least partly related to the problem of sparsity. Because the number of unique n-grams is very large, the number of n-grams that any two Web pages will have in common may be very small. As noted by Grcar et al. [48], distance measures require some overlap in the features; the fewer the number of features in common, the lower the reliability of the distance measures. If there is a high level of sparsity, the k-NN classifier is unable to form reliable neighborhoods [48], and thus in this case the k-NN classifier may give better performance when k=1.

Two of the questions of interest in this study are whether the n-gram length influences the performance of the classification model, and whether the size of the Web page profile influences the classification performance. Therefore, in these experiments, the n-gram length is varied from 2 to 6 in increments of 1, and for each n-gram length, the Web page profile size is varied from 500 to 2500 in increments of 500. Tables 4.1 and 4.2 give the results of these experiments in terms of the precision, recall, and F1-measure, averaged over Web page profile sizes of 500 to 2500 and n-gram lengths of 2 to 6 respectively. In each case, the precision, recall, and F1-measure are macro-averages.

Over n-gram lengths from 2 to 6 and Web page profile sizes from 500 to 2500, the Euclidean distance measure outperforms the Manhattan distance measure, but the Kešelj distance measure provides significantly better performance than either the Manhattan or Euclidean distance measures for the precision and F1-measure ($p < 0.05$ in each case). Although the mean recall for the Euclidean distance measure is higher than that of the Kešelj distance measure, the latter has a better overall performance. Unlike the Euclidean or Manhattan distance measures, the Kešelj distance measure "normalizes" the differences in the n-gram frequencies by dividing by the average frequency for a given n-gram. As noted by Tomovic et al. [132], the use of n-grams of lengths greater than 2 can result in the problem of sparsity, in which there can be a large variance in the frequency of n-grams. By normalizing the difference in the n-gram frequencies, the Kešelj distance measure prevents more frequent n-grams from being given too much emphasis, and this may be why it allows better performance than either the Euclidean or Manhattan distance measures. Based on these results, the Kešelj distance measure is selected for use in subsequent experiments. For each distance measure, the general trend in the results is that as the n-gram length increases, the precision, recall, and F1-measure values decrease; the same trend is seen for the Web page profile sizes.

	Manhattan Distance			Euclidean Distance			Kešelj Distance		
n-gram Length	Prec.	Recall	F1	Prec.	Recall	F1	Prec.	Recall	F1
2	0.779	0.642	0.704	0.682	0.610	0.644	0.828	0.824	0.826
3	0.677	0.480	0.562	0.660	0.578	0.616	0.790	0.689	0.741
4	0.568	0.444	0.498	0.605	0.519	0.559	0.654	0.451	0.539
5	0.533	0.408	0.462	0.586	0.483	0.530	0.543	0.334	0.419
6	0.497	0.375	0.427	0.595	0.468	0.524	0.464	0.291	0.368
Average	0.611	0.470	0.531	0.626	0.532	0.575	0.656	0.518	0.579

Table 4.1: Mean classification results for the k-NN model on the 7-Genre corpus, averaged over Web page profile sizes of 500 to 2500. Standard error ≤ 0.011.

	Manhattan Distance			Euclidean Distance			Kešelj Distance		
Web Page Profile Size	Prec.	Recall	F1	Prec.	Recall	F1	Prec.	Recall	F1
500	0.624	0.487	0.547	0.626	0.542	0.581	0.653	0.523	0.581
1000	0.615	0.478	0.534	0.629	0.536	0.579	0.657	0.512	0.576
1500	0.610	0.464	0.527	0.628	0.531	0.575	0.657	0.515	0.577
2000	0.604	0.469	0.522	0.625	0.526	0.571	0.658	0.520	0.581
2500	0.601	0.459	0.520	0.622	0.523	0.568	0.654	0.521	0.580
Average	0.611	0.470	0.531	0.626	0.532	0.575	0.656	0.518	0.579

Table 4.2: Mean classification results for the k-NN model on the 7-Genre corpus, averaged over n-gram lengths of 2 to 6. Standard error ≤ 0.011.

4.2 Centroid Model

One of the questions being explored in these experiments is what classification method should be used, in order to develop a simple, easily scalable method for Web page classification. The straightforward k-NN model described in Section 4.1 provides a basis for another classification model. This second model, henceforth referred to as the centroid model, differs from the previous model in its treatment of the Web pages in the training set.

In the centroid model, the Web pages in the training set are used to form an n-gram profile for each genre. As with the test set, an n-gram profile of the L most frequent n-grams is constructed for each Web page in the training set. For this model, however, these Web page profiles are then combined based on the genre of the Web pages, creating a centroid profile for each genre. Thus, the n-gram centroid profile for each genre initially contains the combination of the L most frequent n-grams from each Web page in the training set (that belongs to that particular genre). The corresponding n-gram frequencies are averages of the frequencies for all of the Web pages of that genre in the training set. Because the L most frequent n-grams are unlikely to be identical in every Web page profile, combining the Web page profiles to form a centroid genre profile typically results in genre profiles much larger than L. When all of the centroid genre profiles have been constructed, the n-grams in each profile are sorted by frequency, and the centroid genre profiles are then truncated to the size of the smallest centroid genre profile. The size of these centroid genre profiles varies greatly depending on the length of the n-gram that is used. As noted by Miao et al. [85], the longer the n-gram length, the more unique n-grams there are, and the fewer n-grams documents will have in common. Table B.1 in Appendix B gives the size ranges for the centroid genre profiles, as well as the range in the number of unique n-grams used, for each n-gram length from 2 to 6. Table B.1 shows, not unexpectedly, that both the centroid genre profile size and the total number of unique n-grams used in the profiles increases as the n-gram length and/or Web page profile size is increased.

Once the centroid genre profiles have been constructed, the n-gram profile for each Web page in the test set is then compared to each genre profile from the training set. Each test Web page is assigned the label of the genre of the genre profile to which it is closest (most similar), according to the Kešelj distance measure in Equation 4.3. Figure 4.1 gives a visual representation of the centroid model, and Section 4.3 gives a comparison of the results for this centroid model with those of the k-NN approach.

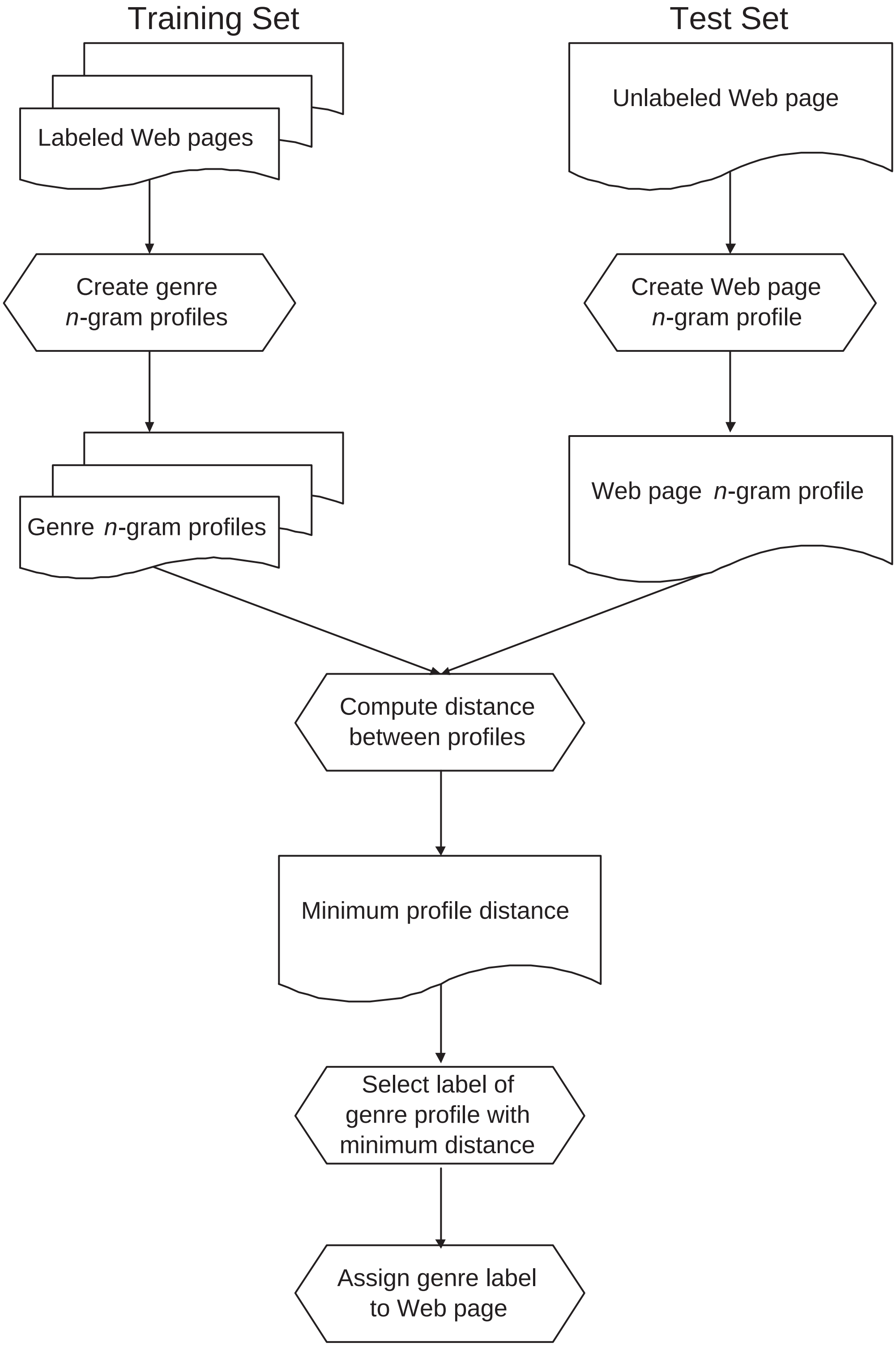
Training Set
Test Set
Labeled Web pages
Unlabeled Web page
Create genre *n*-gram profiles
Create Web page *n*-gram profile
Genre *n*-gram profiles
Web page *n*-gram profile
Compute distance between profiles
Minimum profile distance
Select label of genre profile with minimum distance
Assign genre label to Web page

Figure 4.1: A visual representation of the centroid classification model.

4.3 A Comparison of the k-NN and Centroid Models

The experiments reported in this section were carried out in order to compare the basic k-NN model with the slightly more complex centroid model, working toward the goal of determining what classification method should be used for Web genre classification. Each model uses the Kešelj distance measure given in Equation 4.3 to determine the dissimilarity between two n-gram profiles. As in Section 4.1, the results given for the k-NN model are for the case in which $k = 1$. For each model, 25 trials were performed, each with a different combination of n-gram length and Web page profile size. The n-gram length is varied from 2 to 6 in increments of 1, and for each n-gram length, the Web page profile size is varied from 500 to 2500 in increments of 500. In each case, the precision, recall, and F1-measure are macro-averages.

4.3.1 Effect of n-gram Length

One of the questions of interest to this study is whether the length of the n-gram used in the Web page profiles influences the classification performance. Table 4.3 gives the mean classification results for the k-NN and centroid models on the 7-Genre corpus, averaged over Web page profile sizes of 500 to 2500, for each n-gram length from 2 to 6.

The general observable trend is that with the k-NN model, the precision, recall, and F1-measure values decrease as the n-gram length is increased, however with the centroid model, the precision, recall, and F1-measure values increase as the n-gram length is increased from 2 to 6. This difference in trends for the two models may relate to the increasing sparsity of n-grams as the n-gram length increases; as the length of the n-gram increases, the number of unique n-grams of that length also increases, and this increase becomes exponential for large n. Because the k-NN classifier is comparing the similarity of a Web page profile of limited size to that of other Web page profiles of the same limited size, increasing the n-gram length, and thus the number of unique n-grams, has a detrimental effect on the classification performance. The centroid model, on the other hand, compares each Web page profile to centroid genre profiles that are much larger than the Web page profiles, and this seems to make the performance of the centroid classifier less sensitive to n-gram sparsity.

ANOVA and Scheffé post hoc testing indicate that the effect of n-gram length on the precision, recall, and F1-measure for each model is statistically significant at $p < 0.01$. As explained in Section 3.3, the p-value refers to the probability that the observed results could have occurred by chance if the means of two groups are equal. A small p-value leads to the conclusion that the means are different, with a p-value of less than 0.05 considered to indicate statistical significance.

In these experiments, the partial η^2 for the n-gram length for the precision, recall and F1-measure for each model is greater than 0.20. The partial η^2, computed during the ANOVA, is the proportion of the total variability that is attributable to a factor, therefore these results indicate that the total variability in the precision, recall, and F1-measure for each model is moderately influenced by the n-gram length.

These results suggest that future studies with the centroid model should include experiments with n-grams of greater length. As an anticipatory note however, studies in which the χ^2 statistic (rather than frequency) is used as a feature selection measure exhibit the opposite behaviour with regard to n-gram length, with n-grams of length 2 providing the best overall performance for the centroid model. These studies will be discussed in Chapters 5 and 6.

	k-NN Model			Centroid Model		
n-gram Length	Precision	Recall	F1	Precision	Recall	F1
2	0.828	0.824	0.826	0.863	0.841	0.852
3	0.790	0.689	0.741	0.904	0.893	0.898
4	0.654	0.451	0.539	0.926	0.921	0.923
5	0.543	0.334	0.419	0.931	0.926	0.928
6	0.464	0.291	0.368	0.933	0.928	0.931
Average	0.656	0.518	0.579	0.911	0.902	0.906

Table 4.3: Mean classification results for the k-NN model and the centroid model on the 7-Genre corpus, averaged over Web page profile sizes of 500 to 2500. Standard error ≤ 0.009.

4.3.2 Effect of Web Page Profile Size

Table 4.4 gives the mean classification results for the k-NN and centroid models on the 7-Genre corpus, averaged over n-gram lengths of 2 to 6 for Web page profile sizes from 500 to 2500. Table 4.4 shows that for the centroid classification model, the precision, recall, and F1-measure values decrease as the Web page profile size is increased.

One of the questions of interest to this study is whether the number of n-grams used to represent each Web page (the Web page profile size) influences the classification performance of the k-NN and centroid models. The general observable trend in the results of these experiments is that the precision, recall, and F1-measure values fluctuate slightly with the k-NN model as the Web page profile size is increased, whereas with the centroid model, these values decrease as the Web page profile size is increased.

ANOVA and Scheffé post hoc testing indicate that the effect of the Web page profile size is not statistically significant on the precision of the k-NN classification model. The Web page profile size is statistically significant on the precision for the centroid classification model and on the recall and F1-measure for both models ($p < 0.05$). The p-value refers to the probability that the observed results could have occurred by chance if the means of two groups are equal. A small p-value leads to the conclusion that the means are different, with a p-value of less than 0.05 considered to indicate statistical significance.

The partial η^2 for the Web page profile size for the precision in the centroid classification model and the recall and F1-measures in both models is less than 0.05 in each case. As discussed in Section 3.3.1, the partial η^2 is computed during the ANOVA, and is the proportion of total variability attributable to a factor. These results therefore indicate that the Web page profile size accounted for less than 5% of the overall variance of the dependent variables. Thus, increasing the Web page profile size beyond a base number of 500 does not appreciably affect the performance of either the k-NN classification model or the centroid classification model. It is possible, however, that decreasing Web page profile size below 500 could have an impact on classification performance; experiments incorporating smaller Web page profile sizes will be discussed in Chapters 5 and 6.

	k-NN Model			Centroid Model		
Web Page Profile Size	Precision	Recall	F1	Precision	Recall	F1
500	0.653	0.523	0.581	0.917	0.906	0.912
1000	0.657	0.512	0.576	0.912	0.905	0.909
1500	0.657	0.515	0.577	0.911	0.904	0.908
2000	0.658	0.520	0.581	0.908	0.899	0.903
2500	0.654	0.521	0.580	0.907	0.895	0.901
Average	0.656	0.518	0.579	0.911	0.902	0.906

Table 4.4: Mean classification results for the k-NN model and the centroid model on the 7-Genre corpus, averaged over n-gram lengths of 2 to 6. Standard error ≤ 0.009.

4.3.3 Effect of Genre

As already discussed in Section 3.1.1, the 7-Genre corpus is evenly balanced with 200 Web pages in each of seven genres. Table 4.5 lists the proportion of actual and assigned genres for the case in which the n-gram length is 5 and the Web page profile size is 500, using the centroid model.

The rows in Table 4.5 represent the target genres to which the Web pages actually belong, and the diagonal entries, in bold, give the proportion of Web pages of each genre that were correctly classified; if the classifier had performed perfectly, all diagonal entries would be 1.000, and all off-diagonal entries would be empty. Looking across the row of a particular genre shows the genres across which the Web pages belonging to that genre were distributed by the centroid classification model. Looking down a column of a particular genre shows the proportion of Web pages from other genres that were misclassified as belonging to that genre.

Actual Genre	Assigned Genre						
	BLOG	ESHOP	FAQ	FRONT PAGE	LISTING	PERSONAL HOMEPAGE	SEARCH PAGE
BLOG	**0.995**				0.005		
ESHOP	0.015	**0.925**			0.035	0.005	0.020
FAQ			**0.990**		0.010		
FRONT PAGE				**1.000**			
LISTING	0.035	0.010	0.020	0.015	**0.845**	0.020	0.055
PERS. HOMEPAGE	0.115				0.025	**0.845**	0.015
SEARCH PAGE	0.020	0.015			0.030	0.005	**0.930**

Table 4.5: Misclassification table for the centroid model on the 7-Genre corpus, for the case in which the n-gram length is 5 and the Web page profile size is 500.

Table 4.5 shows that in this case, the ONLINE NEWSPAPER FRONT PAGE genre was the easiest genre to classify. Every Web page belonging to this genre was correctly classified, and only 1.5% of the Web pages were misclassified as belonging to this genre. This indicates that the classifier is able to differentiate very well between this genre and the other genres. The same is true for the FAQ genre, although in this case a few pages belonging to the genre were misclassified as belonging to the LISTING genre. The LISTING and PERSONAL HOMEPAGE genres have the highest proportion of misclassified pages. This is particularly interesting, because these are the only two genres in this corpus that are documented as containing subgenres. As discussed in Section 3.1.1, the LISTING genre can be decomposed into the subgenres CHECKLIST, HOTLIST, SITEMAP, and TABLE. The PERSONAL HOMEPAGE genre, although it does not have specifically labeled subgenres, includes a variety of types of personal homepages, such as academic and administrative personal homepages [106]. Table 4.5 suggests that the presence of these subgenres hampers the classifier's ability to correctly identify the genres.

ANOVA and Scheffé post hoc testing both indicate that the effect of genre on the variability in the precision, recall, and F1-measure for both the k-NN and centroid classification models is significant at $p < 0.01$. The p-value refers to the probability that the observed results could have occurred by chance if the means of two groups are equal. A small p-value leads to the conclusion that the means are different, with a p-value of less than 0.05 considered to indicate statistical significance.

The partial η^2 for genre, for both the precision and recall in each of the classification models, is greater than 0.50. The partial η^2 is computed during the ANOVA, and is the proportion of total variability attributable to a factor, when controlling for other factors. This result therefore indicates that genre accounted for more than than half of the overall variance of the dependent variables. This finding suggests that genres can be successfully differentiated to a high degree, which supports the findings of Dong et al. [41].

4.3.4 Best Results

The best classification results in this study are for the centroid model using byte n-grams of length 5 and a Web page profile size of 500; this combination achieves a best precision of 0.937 with a corresponding recall and classification accuracy of 0.933.

Table 4.6 gives a comparison of the best results obtained in these experiments with those of three other researchers on the same corpus. Santini [106] uses a support vector machine classifier using three different feature sets that include, for example, HTML tags, part-of-speech tags, common word frequencies, and genre-specific facets. Kanaris and Stamatatos [56] also use a support vector machine classifier, but their feature set includes a combination of variable-length character n-grams and structural information from HMTL tags. Dong et al. [41] use a Naïve Bayes classifier on a subset of the corpus containing the genres ESHOP, FAQ, ONLINE NEWSPAPER FRONT PAGE, and PERSONAL HOMEPAGE. Their best accuracy is obtained using a feature set that combines the attributes of content, form, and functionality. When using only content information (word features), Dong et al. achieved a precision of only 0.905.

Researchers	Accuracy
Santini [106]	0.906
Kanaris and Stamatatos [56]	0.965
Dong et al. [41]	0.965
Mason	0.933

Table 4.6: Best classification accuracy results for the 7-Genre corpus. The last line gives the best results for the experiments discussed in this section, achieved using the centroid model with byte n-grams of length 5 and a Web page profile size of 500.

4.3.5 Overall Results

The major contribution of these experiments is to show that an n-gram approach to Web page representation for genre classification is feasible, and that the new centroid model, which gives classification results in the same range as those of other researchers, warrants further investigation. The centroid model achieved significantly higher classification results than the k-NN model, in terms of the precision, recall, and F1-measure ($p < 0.05$ in each case). Based on these results, subsequent work will be based on the centroid model.

These experiments also show that increasing the n-gram length beyond a base size of 2 has a significant impact on the precision, recall, and F1-measure for each model, and should be further explored. These results also reveal that increasing the Web page profile size beyond a base number of 500 does not appreciably affect the performance of either the k-NN model or the centroid model; it is possible, however, that using Web page profile sizes smaller than 500 could have a greater impact on the classification performance.

4.4 A Comparison of Character and Byte n-grams

This study compares the classification results for the centroid model using byte n-grams with those of the same model using character n-grams. Byte n-grams are raw character n-grams in which no bytes are ignored, including the whitespace characters, therefore byte n-grams capture some of the structure of a document. As discussed in Section 3.5, character n-grams use letters only, and ignore digits, punctuation, and most whitespace. These experiments were run using the 7-Genre corpus, with n-gram lengths ranging from 2 to 6 in increments of 1, and a Web page profile size of 500.

Table 4.7 reports the mean precision, recall, and F1-measure values for classification using character and byte n-grams respectively; in each case, the precision, recall, and F1-measure results are macro-averages. For each of the five n-gram lengths, the performance of the centroid model using byte n-gram representations of the Web pages is as good or better than the performance of the same model using character n-gram representations of the Web pages.

Overall, the use of byte n-grams with the centroid model is significantly better than the use of character n-grams ($p < 0.05$), in terms of the average precision, recall, and F1-measure. Based on these results, subsequent work will make use of byte n-grams rather than character n-grams.

	Character n-grams			Byte n-grams		
n-gram Length	Precision	Recall	F1	Precision	Recall	F1
2	0.855	0.825	0.840	0.867	0.832	0.849
3	0.870	0.832	0.850	0.917	0.908	0.912
4	0.929	0.924	0.927	0.930	0.929	0.929
5	0.932	0.929	0.930	0.937	0.933	0.935
6	0.934	0.931	0.932	0.934	0.939	0.932
Average	0.904	0.888	0.896	0.917	0.906	0.912

Table 4.7: Classification results for character and byte n-grams, using the centroid model on the 7-Genre corpus, with a Web page profile size of 500. Standard error ≤ 0.007.

4.5 A Comparison of Word and Byte n-grams

Although the main focus of this research is on using byte n-gram representations of Web pages in order to classify the Web pages by genre, this study explores the use of words to represent Web pages. This is essentially a bag-of-words approach, however the words can be considered to be word n-grams of length 1. As with the byte n-gram Web page representations, the L most frequent words in each Web page are used to form a profile to represent that Web page, and the Web page profiles from the training set are combined based on their genre to create a centroid profile for each genre.

The experiments were run using the 7-Genre corpus. The experiments with words were run both with the same level of data preprocessing that was used for the byte n-gram approach, namely with the HTML tags and JavaScript code removed from the Web pages, and also with an additional preprocessing step to remove stopwords. Stopwords are extremely common words, such as articles, prepositions, and conjunctions, that occur in documents so frequently that they do not help differentiate between documents [14]. Their removal is a common preprocessing step when word representations of documents are used, both in natural language processing and in information retrieval. In these experiments, this preprocessing step made use of a standard stopword list of 571 words, constructed by Salton and Buckley [4].

The classification results for word Web page representations are compared with the average results for byte n-grams of lengths 2 to 6, for Web page profiles of size 500 to 2500.

Table 4.8 reports the mean precision, recall, and F1-measure values for classification using word n-grams, both without stopword removal and with stopword removal, and for classification using byte n-grams. In each case, the precision, recall, and F1-measure results are macro-averages.

Although the removal of stopwords benefitted the classification performance of the word Web page representations, the byte n-gram approach still gave better results. For each of the five Web page profile sizes, the performance of the centroid model using byte n-gram representations of the Web pages is significantly better than the performance of the same model using word 1-gram representations of the Web pages ($p < 0.05$). The p-value refers to the probability that the observed results could have occurred by chance if the means of two groups are equal.

Overall, the use of byte n-grams with the centroid model is significantly better than the use of word 1-grams ($p < 0.05$), in terms of the average precision, recall, and F1-measure values. Based on these results, subsequent work will make use of byte n-grams rather than words.

Note that although stopword removal improved the classification performance for word 1-gram Web page representations, it does not improve the overall classification performance when character or byte n-grams are used. See Tables B.2 and B.3 in Appendix B for a comparison of classification results for character and byte n-gram representations, with and without stopword removal.

	Word 1-grams No Stopword Removal			Word 1-grams With Stopword Removal			Byte n-grams		
Web Page Profile Size	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1
500	0.809	0.797	0.803	0.906	0.898	0.902	0.917	0.906	0.912
1000	0.901	0.887	0.894	0.899	0.886	0.892	0.912	0.905	0.909
1500	0.900	0.885	0.892	0.909	0.897	0.903	0.911	0.904	0.908
2000	0.897	0.882	0.889	0.900	0.887	0.893	0.908	0.899	0.903
2500	0.897	0.881	0.889	0.909	0.899	0.904	0.907	0.895	0.901
Average	0.881	0.866	0.873	0.905	0.893	0.899	0.911	0.902	0.906

Table 4.8: Classification results for word and byte n-grams, using the centroid model on the 7-Genre corpus. The results for byte n-grams are averaged over n-gram lengths of 2 to 6; the results for word n-grams are for length 1. Standard error ≤ 0.009.

4.6 A Comparison of Distance Measures

The experiments with the centroid model discussed in Sections 4.3 to 4.5 determined the dissimilarity between a Web page profile and a genre profile using the Kešelj distance measure defined in Equation 4.3, however other distance measures were also tested for use in the centroid model. As suggested by Tomović et al. [132], the distance measure should meet the following conditions.

- $d\left(P_1, P_1\right) = 0$
- $d\left(P_1, P_2\right) = d\left(P_2, P_1\right)$
- if P_1 and P_2 are similar, $d\left(P_1, P_2\right)$ should be small.
- if P_1 and P_2 are dissimilar, $d\left(P_1, P_2\right)$ should be large.

Tomović et al. [132] note that the last two conditions are informal, as the terms *similar* and *dissimilar* are not strictly defined.

The distance measures examined in this study include those based on the arithmetic mean, the geometric mean, the quadratic mean, the Euclidean distance, and the Manhattan distance, and they were selected based on those examined by Miao et al. [85], Liu and Kešelj [74] and Tomović et al. [132]. The experiments were run using the 7-Genre corpus, with n-gram lengths ranging from 2 to 6 in increments of 1, and a Web page profile size of 500.

Table 4.9 lists the distance measures and reports the mean precision, recall, and F1-measure values, averaged over the five n-gram lengths; in each case the metrics are macro-averages. The results for the Kešelj distance measure (Equation 4.3) are given in the first line of Table 4.9; none of the distance measures tested in these experiments out-performed this distance measure in terms of the precision, recall, and F1-measure. The results in Table 4.9 emphasize the importance of using an appropriate distance measure in the centroid classification model.

It is interesting to note that two of the three distance measures based on the arithmetic mean outperform those based on the quadratic mean, which in turn outperform all of the measures based on the geometric mean. It is also interesting that in these experiments with the centroid classification model, the Manhattan distance measure outperforms the Euclidean distance measure, whereas in the experiments with the basic k-NN model in Section 4.1, the Euclidean distance measure outperformed the Manhattan distance measure.

Based on the results of these distance measure experiments with the centroid model, subsequent work will make use of the Kešelj distance measure that is given in Equation 4.3 and again in the first line of Table 4.9.

Distance Measure	Mean	Mean	Mean
	Precision	Recall	F1
$d\left(P_1, P_2\right) = \sum_{m \in (\mathrm{P}_1 \cup \mathrm{P}_2)} \left(\frac{2(f_1(m)-f_2(m))}{f_1(m)+f_2(m)}\right)^2$	0.917	0.906	0.912
$d\left(P_1, P_2\right) = \sum_{m \in (\mathrm{P}_1 \cup \mathrm{P}_2)} \frac{2\lvert f_1(m)-f_2(m)\rvert}{f_1(m)+f_2(m)}$	0.912	0.903	0.908
$d\left(P_1, P_2\right) = \sum_{m \in (\mathrm{P}_1 \cup \mathrm{P}_2)} \left(\frac{\sqrt{2}(f_1(m)-f_2(m))}{\sqrt{f_1(m)^2+f_2(m)^2}}\right)^2$	0.912	0.902	0.907
$d\left(P_1, P_2\right) = \sum_{m \in (\mathrm{P}_1 \cup \mathrm{P}_2)} \frac{\sqrt{2}\lvert f_1(m)-f_2(m)\rvert}{\sqrt{f_1(m)^2+f_2(m)^2}}$	0.905	0.896	0.901
$d\left(P_1, P_2\right) = \sum_{m \in (\mathrm{P}_1 \cup \mathrm{P}_2)} \frac{2(f_1(m)-f_2(m))^2}{(f_1(m)+f_2(m))}$	0.462	0.458	0.460
$d\left(P_1, P_2\right) = 1 - \frac{2\sum_{m \in (\mathrm{P}_1 \cup \mathrm{P}_2)} f_1(m) f_2(m)}{\sum_{m \in (\mathrm{P}_1 \cup \mathrm{P}_2)} f_1(m)^2 + \sum_{m \in \mathrm{profile}} f_2(m)^2}$	0.382	0.421	0.401
$d\left(P_1, P_2\right) = \sum_{m \in (\mathrm{P}_1 \cup \mathrm{P}_2)} \frac{\lvert f_1(m)-f_2(m)\rvert}{\sqrt{f_1(m) f_2(m)}+1}$	0.373	0.433	0.400
$d\left(P_1, P_2\right) = \sum_{m \in (\mathrm{P}_1 \cup \mathrm{P}_2)} \lvert f_1\left(m\right) - f_2\left(m\right) \rvert$	0.363	0.430	0.393
$d\left(P_1, P_2\right) = \sum_{m \in (\mathrm{P}_1 \cup \mathrm{P}_2)} \left(\frac{f_1(m)-f_2(m)}{\sqrt{f_1(m) f_2(m)}+1}\right)^2$	0.335	0.376	0.354
$d\left(P_1, P_2\right) = \sum_{m \in (\mathrm{P}_1 \cup \mathrm{P}_2)} \left(\frac{f_1(m)-f_2(m)}{\sqrt{f_1(m) f_2(m)}+10}\right)^2$	0.329	0.373	0.350
$d\left(P_1, P_2\right) = \sqrt{\sum_{m \in (\mathrm{P}_1 \cup \mathrm{P}_2)} \left(f_1\left(m\right) - f_2\left(m\right)\right)^2}$	0.328	0.372	0.348
$d\left(P_1, P_2\right) = 1 - \frac{\sum_{m \in (\mathrm{P}_1 \cup \mathrm{P}_2)} f_1(m) f_2(m)}{\sqrt{\left(\sum_{m \in (\mathrm{P}_1 \cup \mathrm{P}_2)} f_1(m)^2\right)\left(\sum_{m \in (\mathrm{P}_1 \cup \mathrm{P}_2)} f_2(m)^2\right)}}$	0.156	0.223	0.184

Table 4.9: Mean classification results for distance measures using the centroid model on the 7-Genre corpus, with a Web page profile size of 500. The results are averaged over n-gram lengths of 2 to 6.

4.7 Conclusions

This chapter has discussed studies which investigated the feasibility of using n-gram representations of Web pages for the task of Web page genre classification. These studies helped establish the focus for the remainder of the research for this thesis, in terms of the open questions to be addressed and the tools with which to work.

The major contribution of the research studies discussed in this chapter is to show that an n-gram approach to Web page representation for genre classification is feasible, and that the new centroid model, which gives classification results in the same range as those of other researchers, warrants further investigation. The results support the hypothesis that Web pages of the same genre share a distribution of n-grams that is similar enough to allow n-gram representations of the Web pages and Web page genres to be used to classify the Web pages by their genres. The fact that the n-gram representation of the Web pages achieved good performance on both the k-NN and centroid classification models suggests that this is a robust method of representing Web pages for classification, and that it should be tested on other classification models, such as the support vector machine classifier.

The results of these experiments also reveal that increasing the Web page profile size beyond a base number of 500 does not appreciably affect the performance of either the k-NN model or the centroid model. The general observable trend in the results of these experiments is that the classification performance of the k-NN model fluctuates slightly as the Web page profile size is increased, whereas the classification performance of the centroid model decreases slightly as the Web page profile size is increased. However, based on ANOVA and Scheffé post hoc testing, Web page profile size accounted for less than 5% of the overall variance of the dependent variables. It is possible, however, that using Web page profile sizes smaller than 500 could have a greater impact on the classification performance, therefore future experiments will include an investigation of smaller Web page profile sizes.

In the experiments discussed in this chapter, frequency was used as the feature selection method for choosing the n-grams with which to represent each Web page, however it is not unreasonable to hypothesize that a more theoretically sound feature selection measure could be more effective. Future work will include an investigation of the χ^2 statistic and Information Gain as feature selection measures.

The experiments discussed in this chapter show that increasing the n-gram length beyond a base size of 2 has a significant impact on the precision, recall, and F1-measure values for each model, and should be further explored. It is very interesting that the effect of n-gram length differed between the k-NN and centroid models, with longer n-grams resulting in better performance for the centroid model, but worse performance for the k-NN model. These results suggest that a larger range of n-gram lengths should be investigated with the centroid model, and that a wide range of n-gram lengths should be explored each time a new classification model is introduced.

Although the centroid model achieved high classification performance with word, character and byte n-gram representations of Web pages, the use of byte n-grams gave the best overall performance. Unlike character n-grams, byte n-grams do not ignore any bytes in the text, and therefore capture some of the structure of the Web pages. For these experiments, the HTML tags and JavaScript code were removed from the Web pages as a preprocessing step, however the success of the byte n-gram approach suggests that this may not be necessary. Future work will include experiments to determine whether removing the HTML tags and/or JavaScript code has an significant effect on classification performance.

The results discussed in this chapter also indicate that the Kešelj distance measure given in Equation 4.3 gives the best performance when compared with several other distance measures, therefore subsequent work with the centroid classification model will make use of this distance measure.

Chapter 5 will discuss experiments which further the investigation of the best set of parameters for the proposed approach, as well as experiments using the χ^2 statistic and Information Gain as feature selection measures. The effect of data preprocessing is also explored, and the experiments are expanded to include the KI-04 corpus.

Chapter 5

Classification Experiments: Balanced Corpora

Chapter 4 discussed studies which investigated the feasibility of using n-gram representations of Web pages for Web page genre classification. The results of these studies indicated that an n-gram approach to Web page representation for genre classification is feasible, and that the new centroid model, which gives classification results in the same range as those of other researchers, should be the subject of further study. This chapter discusses studies which further the investigation of the best set of parameters for the proposed approach, including n-gram length and Web page profile size. The effect of data preprocessing is also explored, as is the use of the χ^2 statistic and Information Gain as feature selection measures. The experiments in these studies broaden the research scope to include both the 7-Genre and KI-04 corpora; as described in Section 3.1, these are balanced corpora with a similar level of abstraction.

Section 5.1 discusses a set of experiments investigating data preprocessing, while Section 5.2 describes the exploration of feature selection measures. Section 5.3 revisits the question of data preprocessing, providing further insight based on the results of the feature selection experiments. The chapter concludes with a summary of the studies and their results, given in Section 5.4.

5.1 Data Preprocessing

This research study investigates the effects of typical data preprocessing steps when using an n-gram approach to Web page representation: are preprocessing steps beneficial or detrimental to the classification accuracy achieved by the centroid classification model? In order to investigate the effect of some common preprocessing steps, experiments are run with three different levels of preprocessing on the 7-Genre and KI-04 corpora. These three levels are no preprocessing, removing only the JavaScript code, and removing both the HTML tags and JavaScript code.

On each corpus, 25 trials were performed with each level of preprocessing, each with a different combination of n-gram length and Web page profile size. The n-gram length was varied from 2 to 6 in increments of 1, and for each n-gram length, the Web page profile size was varied from 500 to 2500 in increments of 500. Although the study discussed in Section 4.3 indicated that increasing the Web page profile size beyond a base number of 500 does not appreciably affect the performance of the centroid classification model, this may be dependent on the level of preprocessing, therefore a range of Web page profile sizes was investigated in this study.

5.1.1 Effect of n-gram Length

Tables 5.1 and 5.2 give the mean classification results for the 7-Genre corpus and the KI-04 corpus respectively; the precision, recall, and F1-measure values are averaged over Web page profile sizes of 500 to 2500, for each n-gram length from 2 to 6. Tables 5.1 and 5.2 indicate that of the three levels of preprocessing that were examined, removing both the HTML tags and JavaScript code gives the best classification performance, while doing no preprocessing gives the worst performance. Although there were exceptions, the general trend with regard to n-gram length is that as the n-gram length increases, the precision, recall, and F1-measure values also increase.

For both the 7-Genre and the KI-04 corpora, for each of the three levels of preprocessing, the effect of the n-gram length was significant at $p < 0.01$, for the precision, recall, and F1-measure results. The p-value refers to the probability that the observed results could have occurred by chance if the means of two groups are equal, with a small p-value leading to the conclusion that the means are different.

The partial η^2, computed during the ANOVA, is the proportion of total variability attributable to a factor. For the 7-Genre corpus, the partial η^2 was at least 0.45 in each case, indicating that the total variability in the precision, recall, and F1-measure values is quite heavily influenced by the n-gram length. The influence of the n-gram length was less pronounced on the KI-04 corpus, however the partial η^2 was greater than 0.25 in each case, therefore n-gram length had a moderate influence. In all cases for each corpus, the n-gram length of 2 was significantly worse ($p < 0.01$) than n-gram lengths from 3 to 6; although an n-gram length of 3 was not as good as n-gram lengths from 4 to 6, this difference was significant only for the 7-Genre corpus. In each case, there was no significant difference in n-gram lengths of 4, 5, and 6.

7-Genre Corpus	No Preprocessing			JavaScript Removed			HTML Tags and JavaScript Removed		
n-gram Length	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1
2	0.862	0.852	0.857	0.871	0.862	0.867	0.863	0.841	0.852
3	0.905	0.899	0.902	0.910	0.900	0.905	0.904	0.893	0.898
4	0.919	0.914	0.916	0.923	0.916	0.920	0.926	0.921	0.923
5	0.918	0.910	0.914	0.923	0.916	0.920	0.931	0.926	0.928
6	0.913	0.903	0.908	0.916	0.906	0.911	0.933	0.928	0.931
Average	0.903	0.896	0.900	0.909	0.900	0.904	0.911	0.902	0.906

Table 5.1: Mean classification results for three levels of preprocessing, using the centroid model on the 7-Genre corpus, averaged over Web page profile sizes of 500 to 2500. Standard error ≤ 0.003.

KI-04 Corpus	No Preprocessing			JavaScript Removed			HTML Tags and JavaScript Removed		
n-gram Length	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1
2	0.750	0.717	0.733	0.758	0.724	0.739	0.804	0.778	0.790
3	0.950	0.945	0.947	0.954	0.946	0.950	0.950	0.943	0.947
4	0.962	0.957	0.960	0.962	0.957	0.959	0.961	0.951	0.956
5	0.964	0.959	0.962	0.964	0.959	0.961	0.962	0.951	0.956
6	0.970	0.965	0.967	0.967	0.960	0.963	0.962	0.954	0.958
Average	0.919	0.909	0.914	0.921	0.909	0.915	0.928	0.915	0.921

Table 5.2: Mean classification results for three levels of preprocessing, using the centroid model on the KI-04 corpus, averaged over Web page profile sizes of 500 to 2500. Standard error ≤ 0.017

5.1.2 Effect of Web Page Profile Size

Tables 5.3 and 5.4 give the mean classification results for the 7-Genre corpus and the KI-04 corpus respectively; the precision, recall, and F1-measure values in each case are averaged over n-gram lengths of 2 to 6, for each Web page profile size from 500 to 2500. As with Tables 5.1 and 5.2, Tables 5.3 and 5.4 indicate that of the three levels of preprocessing that were examined, removing both the HTML tags and JavaScript code gives the best classification performance, while doing no preprocessing gives the worst performance.

For the KI-04 corpus, the Web page profile size had no significant effect on the precision, recall, and F1-measure values for any of the levels of preprocessing. This could be partly due to a ceiling effect, because these values are all relatively high.

When no preprocessing was done on the 7-Genre corpus, the general trend was for the precision, recall, and F1-measure values to increase as the Web page profile size increased, however, the differences were not statistically significant for Web page profile sizes of 1000 to 2500. When only the JavaScript code was removed from the 7-Genre corpus, the precision, recall, and F1-measure values also tended to increase as the Web page profile size increased, but these differences were not significant for Web page profile sizes of 1500 to 2500. In contrast, for the case in which both the HTML tags and JavaScript code were removed from the 7-Genre corpus as preprocessing steps, the precision, recall, and F1-measure values tended to decrease as the Web page profile size increased.

For the 7-Genre corpus, the effect of the Web page profile size was significant at $p < 0.01$, for the precision, recall, and F1-measure not only for the case in which no preprocessing was done, but also for the case in which the JavaScript code was removed. However, for the case in which both HTML tags and JavaScript code were removed from the corpus, the effect of the Web page profile size was only significant for the recall ($p < 0.05$). For the latter case, the partial η^2 computed during the ANOVA was less than 0.05, indicating that very little of the variability in the recall could be attributed to the Web page profile size. However, for the former two cases, in which no preprocessing was done and in which only the JavaScript code was removed, the partial η^2 was at least 0.20, indicating that the total variability in the precision, recall, and F1-measure was moderately influenced by the Web page profile size.

7-Genre Corpus	No Preprocessing			JavaScript Removed			HTML Tags and JavaScript Removed		
Web Page Profile Size	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1
500	0.882	0.875	0.878	0.885	0.877	0.881	0.917	0.906	0.912
1000	0.902	0.893	0.897	0.907	0.898	0.902	0.912	0.905	0.909
1500	0.908	0.899	0.903	0.913	0.904	0.908	0.911	0.904	0.908
2000	0.912	0.904	0.908	0.917	0.908	0.912	0.908	0.899	0.903
2500	0.914	0.907	0.910	0.922	0.914	0.918	0.907	0.895	0.901
Average	0.903	0.896	0.900	0.909	0.900	0.904	0.911	0.902	0.906

Table 5.3: Mean classification results for three levels of preprocessing, using the centroid model on the 7-Genre corpus, averaged over n-gram lengths of 2 to 6. Standard error ≤ 0.003

KI-04 Corpus	No Preprocessing			JavaScript Removed			HTML Tags and JavaScript Removed		
Web Page Profile Size	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1
500	0.912	0.894	0.902	0.916	0.894	0.904	0.921	0.909	0.914
1000	0.908	0.891	0.899	0.912	0.893	0.902	0.922	0.910	0.916
1500	0.919	0.911	0.915	0.918	0.909	0.914	0.930	0.917	0.923
2000	0.926	0.921	0.924	0.927	0.921	0.924	0.932	0.920	0.926
2500	0.930	0.926	0.928	0.932	0.927	0.930	0.933	0.921	0.927
Average	0.919	0.909	0.914	0.921	0.909	0.915	0.928	0.915	0.921

Table 5.4: Mean classification results for three levels of preprocessing, using the centroid model on the KI-04 corpus, averaged over n-gram lengths of 2 to 6. Standard error ≤ 0.017

5.1.3 Data Preprocessing: Conclusions

This set of experiments investigated the effects of common data preprocessing steps when using an n-gram approach to Web page representation with the centroid classification model. Experiments were run with three different levels of preprocessing on the 7-Genre and KI-04 corpora; these three levels were no preprocessing, removing only the JavaScript code, and removing both the HTML tags and JavaScript code.

Tables 5.1–5.4 show that for both the 7-Genre and KI-04 corpora, the classification performance of the centroid model increases with each level of preprocessing: doing no preprocessing gives poorer results than does removing the JavaScript code from the Web pages, and removing both the HTML tags and the JavaScript code further improves the classification performance of the model. However, based on both ANOVA and Scheffé post hoc testing, these differences are not statistically significant on the KI-04 corpus, and on the 7-Genre corpus, there is no statistically significant difference between removing only the JavaScript code, and removing both the HTML tags and the JavaScript code. Both of these levels of preprocessing, however, are significantly better than doing no preprocessing, with $p < 0.01$ for the precision, recall, and F1-measure.

The lack of a statistically significant difference between preprocessing levels for the KI-04 corpus may be due to a ceiling effect; the classification performance of the centroid model is very good for each preprocessing level. In addition, fewer Web pages in the KI-04 corpus contained JavaScript code than in the 7-Genre corpus; in the 7-Genre corpus, 78% of the Web pages contain JavaScript code, whereas in the KI-04 corpus, only 53% of the Web pages contain JavaScript code.

It is interesting to note that the only experiments in which the precision, recall, and F1-measure values tended to decrease as the Web page profile size increased were for the case in which both the HTML tags and JavaScript code were removed as preprocessing steps, with the 7-Genre corpus. This indicates that smaller Web page profiles can be advantageous when the feature set has been reduced by preprocessing, and leads to the question of whether reducing the feature set by other means could also be effective. Section 5.2 investigates theoretically based measures for feature selection, while Section 5.3 revisits the question of data preprocessing.

5.2 Feature Selection Measures

Although the centroid classification model achieves high classification performance on the 7-Genre and KI-04 corpora when n-gram frequency is used as a feature selection measure, it is reasonable to hypothesize that a more theoretically sound feature selection measure could be even more effective. Thus, this study compares the classification results for the centroid model when n-gram frequency, Information Gain, and the χ^2 statistic are used as feature selection measures. As discussed in Section 3.4, Information Gain is based on information theory, while the χ^2 statistic is statistically based. The basic idea behind using a feature selection measure is to better identify the n-grams whose presence in a Web page gives strong evidence either for or against the identification of a Web page as belonging to a particular genre [22].

On each corpus, 162 trials were performed with each feature selection measure, each with a different combination of n-gram length and Web page profile size. Based on the results of the experiments in Section 5.1, the n-gram length was varied from 2 to 10 in increments of 1, and for each n-gram length, the Web page profile size was varied from 5 to 50 in increments of 5 and from 50 to 500 in increments of 50; HTML tags and JavaScript code were removed from the Web pages as preprocessing steps.

5.2.1 Effect of n-gram Length

Tables 5.5 and 5.6 give the mean classification results for the 7-Genre corpus and the KI-04 corpus respectively; the precision, recall, and F1-measure in each case is averaged over Web page profile sizes of 5 to 500, for each n-gram length from 2 to 10. Tables 5.5 and 5.6 indicate that of the feature selection measures that were examined, using the χ^2 statistic gives the best classification performance, while using n-gram frequency gives the worst performance. The general observable trend is that for the n-gram frequency and Information Gain feature selection measures, the classification performance tends to increase and then gradually decrease as the n-gram length is increased from 2 to 10, whereas for the χ^2 statistic, the performance is more stable. The n-gram length that gave the best results depended on the corpus, the feature selection measure, and in some cases, the evaluation measure.

For the 7-Genre corpus, when n-gram frequency was used as the feature selection measure, Scheffé post hoc tests indicated that an n-gram length of 6 gave the best results in terms of the precision, recall, and F1-measure, but that there was no statistically significant difference between the n-gram lengths of 4, 5, and 6. When n-gram frequency was used as the feature selection measure with the KI-04 corpus, an n-gram length of 7 gave the best results in terms of precision, recall, and the F1-measure, but there was no statistically significant difference between any of the n-gram lengths from 5 to 10. An n-gram length of 2 gave the worst performance on each corpus when n-gram frequency was used as the frequency selection measure.

When Information Gain was used as the feature selection measure with the 7-Genre corpus, an n-gram length of 8 gave the best results, however Scheffé post hoc tests showed that there was no significant difference in n-gram lengths from 5 to 10 for the precision, from 6 to 9 for the recall, and from 6 to 10 for the F1-measure. On the KI-04 corpus, the Information Gain measure performed best with an n-gram length of 6 for the precision and 7 for the recall and F1-measure, however there was no statistically significant difference between n-gram lengths of from 5 to 10 for the precision, recall, or F1-measure. An n-gram length of 2 gave the worst performance on each corpus when Information Gain was used as the frequency selection measure.

Although an n-gram length of 2 gave the worst performance for the n-gram frequency and Information Gain feature selection measures on both corpora, it gave the best performance for the χ^2 statistic on the 7-Genre corpus in terms of the precision, and on the KI-04 corpus in terms of the precision, recall, and F1-measure. Although an n-gram length of 6 gave the best results in terms of the recall and F1-measure on the 7-Genre corpus, there was no statistically significant difference between n-gram lengths of 2, 5, and 6. On the KI-04 corpus, there was no significant difference between n-gram lengths of 2, 6, 7, 8, 9, and 10 when the χ^2 statistic was used.

For each corpus, the effect of n-gram length was most pronounced for the Information Gain feature selection measure, and least pronounced for χ^2 statistic. For both the 7-Genre and the KI-04 corpora, for each of the three feature selection measures tested in these experiments, the effect of the n-gram length was significant at $p < 0.01$, for the precision, recall, and F1-measure. The p-value refers to the probability that the observed results could have occurred by chance if the means of two groups are equal; a small p-value leads to the conclusion that the means are different.

7-Genre	Frequency			Information Gain			χ^2 Statistic		
n-gram Length	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1
2	0.765	0.737	0.750	0.798	0.743	0.769	0.974	0.968	0.971
3	0.819	0.811	0.815	0.897	0.852	0.873	0.956	0.947	0.951
4	0.837	0.833	0.835	0.922	0.860	0.888	0.966	0.961	0.963
5	0.838	0.835	0.837	0.948	0.897	0.920	0.973	0.969	0.971
6	0.839	0.836	0.837	0.950	0.901	0.924	0.973	0.970	0.972
7	0.827	0.823	0.825	0.951	0.902	0.925	0.968	0.964	0.966
8	0.809	0.807	0.808	0.952	0.907	0.928	0.962	0.957	0.959
9	0.799	0.795	0.797	0.951	0.904	0.927	0.956	0.950	0.953
10	0.780	0.775	0.777	0.949	0.896	0.921	0.950	0.946	0.948
Average	0.812	0.806	0.809	0.924	0.874	0.897	0.964	0.959	0.962

Table 5.5: Mean classification results for three feature selection measures, using the centroid model on the 7-Genre corpus, averaged over Web page profile sizes of 5 to 500. Standard error ≤ 0.002.

KI-04	Frequency			Information Gain			χ^2 Statistic		
n-gram Length	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1
2	0.725	0.704	0.714	0.799	0.669	0.726	0.984	0.981	0.982
3	0.845	0.835	0.839	0.913	0.847	0.876	0.969	0.959	0.964
4	0.893	0.889	0.891	0.958	0.934	0.945	0.969	0.964	0.966
5	0.920	0.918	0.919	0.982	0.974	0.978	0.974	0.971	0.973
6	0.931	0.929	0.930	0.986	0.981	0.983	0.977	0.973	0.975
7	0.935	0.934	0.935	0.987	0.982	0.984	0.979	0.975	0.977
8	0.935	0.934	0.934	0.986	0.981	0.983	0.978	0.973	0.975
9	0.934	0.932	0.933	0.987	0.981	0.984	0.979	0.975	0.977
10	0.930	0.928	0.929	0.987	0.982	0.984	0.980	0.978	0.979
Average	0.894	0.889	0.892	0.954	0.926	0.938	0.976	0.972	0.974

Table 5.6: Mean classification results for three feature selection measures, using the centroid model on the KI-04 corpus, averaged over Web page profile sizes of 5 to 500. Standard error ≤ 0.005.

5.2.2 Effect of Web Page Profile Size

Tables 5.7 and 5.8 give the mean classification results for the 7-Genre corpus and the KI-04 corpus respectively; the precision, recall, and F1-measure values in each case are averaged over n-gram lengths of 2 to 10, for each Web page profile size from 5 to 500. As with Tables 5.5 and 5.6, Tables 5.7 and 5.8 indicate that all three of the three feature selection measures that were examined in these experiments give high classification performance, however using the χ^2 statistic gives the best performance, while using n-gram frequency gives the worst performance. The general observable trend in Tables 5.7 and 5.8 is that when n-gram frequency is used as a feature selection measure, the classification accuracy tends to increase as the Web page profile size is increased from 5 to 500, whereas for Information Gain and the χ^2 statistic, the classification accuracy tends to increase and then gradually decrease as the Web page profile size is increased from 5 to 500. This is an indication that the n-grams selected when Information Gain or the χ^2 statistic is used as a feature selection measure are better able to discriminate between Web page genres than those n-grams selected based on n-gram frequency, however it is important to note that unlike the use of Information Gain or the χ^2 statistic, the use of n-gram frequency for a particular genre does not require knowledge about the other genres in the corpus. As was the case with n-gram length, the best Web page profile size varied, depending on the corpus, the feature selection measure, and in some cases, the evaluation measure.

For both the 7-Genre and KI-04 corpora, when n-gram frequency was used as the feature selection measure, the precision, recall, and F1-measure values all increased as the Web page profile size increased from 5 to 500, with a Web page profile size of 500 giving the best results in terms of the precision, recall, and F1-measure, and a Web page profile size of 5 giving the worst results. On the 7-Genre corpus, Scheffé post hoc tests indicated that there was no statistically significant difference between Web page profile sizes from 250 to 500, and on the KI-04 corpus there was no significant difference in using Web page profile sizes from 30 to 500. These results are interesting, because using a smaller Web page profile size could have the advantage of being less computationally intensive.

When Information Gain was used as the feature selection measure with the 7-Genre corpus, a Web page profile size of 150 gave the best results in terms of the precision, whereas a Web page profile size of 200 gave the best results in terms of the recall and F1-measure. However, Scheffé post hoc tests indicated that there was no significant difference in Web page profile sizes of 100 to 350 for any of three evaluation measures, and in terms of the recall, there was also no significant difference when a Web page profile size of 400 was used. On the KI-04 corpus, the best results were obtained with a Web page profile size of 50, however there was no significant difference in the precision with Web page profile sizes of 10 to 200, and no significant difference in the recall with Web page profile sizes of 5 to 250. There was no significant difference with Web page profile sizes of 10 to 250 on the F1-measure values when Information Gain was used as the feature selection measure on the KI-04 corpus. It is interesting that with Information Gain, as with n-gram frequency, using larger Web page profile sizes is not necessarily beneficial.

On the 7-Genre corpus, the use of the χ^2 statistic as a feature selection measure was most successful with a Web page profile size of 30, although there was no significant difference on the precision when Web page profile sizes of 10 to 150 were used, and there was no significant difference on the recall and the F1-measure for Web page profile sizes of 15 to 50. On the KI-04 corpus, the use of the χ^2 statistic as a feature selection measure was most successful with a Web page profile size of 40, although there was no statistically significant difference when Web page profile sizes of 10 to 200 were used. As with n-gram frequency and Information Gain, larger Web page profile sizes do not necessarily mean significantly better classification results.

For both the 7-Genre and the KI-04 corpora, for each of the three feature selection measures tested in these experiments, the effect of the Web page profile size was significant for the precision, recall, and F1-measure, at $p < 0.01$. The p-value refers to the probability that the observed results could have occurred by chance if the means of two groups are equal; a small p-value leads to the conclusion that the means are different. The effect of the Web page profile size was greater on the 7-Genre corpus than on the KI-04 corpus, however this could be partly due to a ceiling effect with the KI-04 corpus, because the precision, recall, and F1-measure values are all relatively high on this corpus. For each corpus, the effect of the Web page profile size was most pronounced when n-gram frequency was used as the feature selection measure.

7-Genre	Frequency			Information Gain			χ^2 Statistic		
Web Page Profile Size	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1
5	0.564	0.556	0.560	0.888	0.719	0.793	0.952	0.932	0.942
10	0.637	0.632	0.634	0.901	0.789	0.841	0.971	0.967	0.969
15	0.675	0.670	0.672	0.907	0.816	0.859	0.979	0.977	0.978
20	0.720	0.714	0.717	0.910	0.833	0.870	0.981	0.980	0.980
25	0.741	0.737	0.739	0.914	0.846	0.878	0.981	0.980	0.980
30	0.763	0.758	0.761	0.918	0.857	0.887	0.981	0.980	0.981
35	0.777	0.771	0.774	0.919	0.864	0.891	0.981	0.980	0.981
40	0.790	0.784	0.787	0.923	0.872	0.896	0.982	0.981	0.981
45	0.805	0.799	0.802	0.926	0.880	0.902	0.982	0.981	0.981
50	0.810	0.803	0.806	0.927	0.885	0.905	0.981	0.980	0.981
100	0.865	0.858	0.861	0.939	0.918	0.928	0.976	0.974	0.975
150	0.888	0.880	0.884	0.943	0.929	0.936	0.967	0.965	0.966
200	0.899	0.892	0.895	0.941	0.931	0.936	0.959	0.955	0.957
250	0.908	0.901	0.904	0.941	0.928	0.934	0.954	0.948	0.951
300	0.913	0.906	0.910	0.939	0.923	0.931	0.948	0.940	0.944
350	0.918	0.910	0.914	0.936	0.917	0.926	0.942	0.933	0.938
400	0.920	0.913	0.916	0.937	0.908	0.922	0.936	0.925	0.930
450	0.921	0.912	0.916	0.929	0.895	0.910	0.935	0.925	0.930
500	0.923	0.915	0.919	0.923	0.885	0.902	0.932	0.920	0.926
Average	0.812	0.806	0.809	0.924	0.874	0.897	0.964	0.959	0.962

Table 5.7: Mean classification results for three feature selection measures, using the centroid model on the 7-Genre corpus, averaged over n-gram lengths of 2 to 10. Standard error ≤ 0.002.

KI-04	Frequency			Information Gain			χ^2 Statistic		
Web Page Profile Size	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1
5	0.712	0.697	0.704	0.943	0.910	0.925	0.961	0.950	0.955
10	0.787	0.778	0.782	0.959	0.940	0.948	0.977	0.975	0.976
15	0.829	0.821	0.824	0.965	0.947	0.955	0.983	0.983	0.983
20	0.851	0.842	0.846	0.967	0.950	0.958	0.985	0.985	0.985
25	0.866	0.859	0.862	0.967	0.950	0.958	0.987	0.987	0.987
30	0.877	0.871	0.874	0.968	0.950	0.958	0.988	0.987	0.988
35	0.884	0.879	0.882	0.968	0.950	0.958	0.988	0.987	0.987
40	0.891	0.886	0.888	0.969	0.952	0.960	0.988	0.988	0.988
45	0.897	0.892	0.894	0.969	0.951	0.960	0.988	0.988	0.988
50	0.900	0.896	0.898	0.970	0.952	0.960	0.988	0.988	0.988
100	0.926	0.922	0.924	0.962	0.943	0.951	0.984	0.983	0.984
150	0.935	0.934	0.934	0.956	0.932	0.943	0.978	0.975	0.977
200	0.944	0.943	0.943	0.951	0.924	0.936	0.973	0.968	0.970
250	0.946	0.945	0.945	0.944	0.914	0.928	0.969	0.962	0.965
300	0.948	0.946	0.947	0.938	0.901	0.917	0.967	0.960	0.963
350	0.948	0.947	0.947	0.935	0.891	0.910	0.965	0.957	0.961
400	0.948	0.946	0.947	0.933	0.884	0.905	0.964	0.954	0.959
450	0.949	0.947	0.948	0.932	0.887	0.901	0.962	0.950	0.956
500	0.950	0.947	0.949	0.931	0.871	0.896	0.959	0.945	0.952
Average	0.894	0.889	0.892	0.954	0.926	0.938	0.976	0.972	0.974

Table 5.8: Mean classification results for three feature selection measures, using the centroid model on the KI-04 corpus, averaged over n-gram lengths of 2 to 10. Standard error ≤ 0.008.

5.2.3 Feature Selection Measures: Conclusions

These experiments investigated the effects of three different feature selection measures when using an n-gram approach to Web page representation with the centroid classification model. Experiments were run using n-gram frequency, Information Gain, and the χ^2 statistic as feature selection measures on the 7-Genre and KI-04 corpora.

Tables 5.5–5.8 show that although the centroid model achieves high classification performance using each of the three feature selection measures, for both the 7-Genre and KI-04 corpora the centroid model achieves the best classification performance when the χ^2 statistic is used as a feature selection measure. An ANOVA test shows that this performance is significantly better than using either n-gram frequency or Information Gain as feature selection measures ($p < 0.01$).

Information Gain and the χ^2 statistic have several characteristics in common: each favors common terms (n-grams), each makes use of category (genre) information, and each is two-sided, making use of both the presence and absence of terms (n-grams). Both Yang and Pedersen [142] and Forman [45] found a strong correlation between classification results using Information Gain and the χ^2 statistic as feature selection measures, however Forman's studies indicated that the χ^2 statistic is more likely to select terms based on the positive documents containing the term than is Information Gain. It is possible that this distinction could account for the difference in performance between the two measures shown in Tables 5.5–5.8. Although Information Gain and the χ^2 statistic each outperform n-gram frequency as a feature selection measure in these experiments, it is important to consider that unlike the use of Information Gain or the χ^2 statistic, the use of n-gram frequency for a particular genre does not require knowledge about the other genres in the corpus.

It is interesting to note that the χ^2 statistic feature selection measure not only results in the best performance for the centroid classifier, but also allows the use of a small n-gram length of 2, and a relatively small Web page profile size of less than 50; the use of short n-gram lengths and small Web page profiles may have the advantage of being less computationally intensive. These experiments also indicated that if n-gram frequency is used as the feature selection measure, then Web page profile sizes as small as 250 would give good results on both of the corpora tested.

Although the results of the preprocessing experiments in Section 5.1 indicated that removing the HTML tags and JavaScript code from the Web pages resulted in the best classification performance with the centroid model, the experiments were performed using n-gram frequency as the feature selection measure. Using the χ^2 statistic as a feature selection measure not only gives superior classification performance, but also allows the successful use of shorter n-grams and smaller Web page profile sizes. Because using the χ^2 statistic rather than n-gram frequency to select n-grams seems to result in the selection of more discriminating n-grams, the preprocessing steps may be unnecessary. Section 5.3 therefore revisits the question of whether preprocessing steps are beneficial to the classification accuracy achieved by the centroid classification model when the χ^2 statistic is used as a feature selection measure.

5.3 Data Preprocessing Revisited

The study discussed in this section revisits the question of whether preprocessing steps are beneficial to the classification accuracy achieved by the centroid classification model. Although the results of the preprocessing experiments in Section 5.1 indicated that removing the HTML tags and JavaScript code from the Web pages resulted in the best classification performance with the centroid model, the experiments were performed using n-gram frequency as the feature selection measure. The results of the experiments in the study discussed in Section 5.2 indicate that using the χ^2 statistic as a feature selection measure produces better classification results with the centroid model, and allows the use of shorter n-grams and smaller Web page profiles.

Because using the χ^2 statistic rather than n-gram frequency to select n-grams seems to result in the selection of more discriminating n-grams, the use of the χ^2 statistic may make preprocessing the Web pages unnecessary. This study therefore revisits the question of whether preprocessing steps are beneficial to the classification accuracy achieved by the centroid classification model. In these experiments, the χ^2 statistic was used as the feature selection measure, with n-gram lengths ranging from 2 to 6 in increments of 1, and Web page profile sizes ranging from 5 to 50 in increments of 5. The experiments were run with three different levels of preprocessing on the 7-Genre and KI-04 corpora; the preprocessing levels were no preprocessing, removing only the JavaScript code, and removing both the HTML tags and JavaScript code.

5.3.1 Effect of n-gram Length

Tables 5.9 and 5.10 give the mean classification results for the 7-Genre corpus and the KI-04 corpus respectively; the precision, recall, and F1-measure in each case is averaged over Web page profile sizes of 5 to 50, for each n-gram length from 2 to 6. Tables 5.9 and 5.10 indicate that of the three levels of preprocessing that were examined, removing both the HTML tags and JavaScript code gives the worst classification performance, while there is no significant difference between removing only the JavaScript code and doing no preprocessing at all. This is in contrast to the results of the preprocessing study discussed in Section 5.1, and is an indication of the superior ability of the χ^2 statistic to select n-grams with high discriminatory powers.

For each level of preprocessing on the 7-Genre corpus, Scheffé post hoc tests indicated that an n-gram length of 2 gave the best results in terms of the precision, recall, and F-1 measure, but that there was no statistically significant difference between the n-gram lengths of 2, 3, and 4. The general trend was that for each level of preprocessing, the classification performance increased slightly as the n-gram length decreased from 6 to 2.

When both the HTML tags and JavaScript code were removed as preprocessing steps with the KI-04 corpus, Scheffé post hoc tests indicated that an n-gram length of 2 gave significantly better classification results in terms of the precision, recall, and F-1 measure ($p < 0.01$). An n-gram length of 2 also gave the best classification results on this corpus when only the JavaScript code was removed as a preprocessing step, however in this case there was no statistically significant difference between n-gram lengths of 2, 3, and 4. For the case in which no preprocessing was performed with this corpus, an n-gram length of 2 gave the best performance in terms of the recall and F1-measure, while an n-gram length of 3 gave the best precision. For this level of preprocessing there was, however, no statistically significant difference in any of the n-gram lengths from 2 to 5. As with the 7-Genre corpus, the general trend was that for each level of preprocessing, the classification accuracy increased slightly as the n-gram length decreased from 6 to 2, indicating that the use of the χ^2 statistic as a feature selection measure allows the centroid model to achieve high classification accuracy even when relatively short n-grams are used.

For both the 7-Genre and the KI-04 corpora, for each of the three levels of preprocessing tested in these experiments, the effect of the n-gram length was significant for the precision, recall, and F1-measure at $p < 0.01$. The p-value refers to the probability that the observed results could have occurred by chance if the means of two groups are equal; a small p-value leads to the conclusion that the means are different. For the 7-Genre corpus, the partial η^2 for the n-gram length was less than 0.15 in each case, while for the KI-04 corpus it was less than 0.22 in each case. The partial η^2, computed during the ANOVA, is the proportion of the total variability that is attributable to a factor, therefore these results indicate that for each corpus, the total variability in the precision, recall and F1-measure for each level of preprocessing is only moderately influenced by the n-gram length.

7-Genre Corpus	No Preprocessing			JavaScript Removed			HTML Tags and JavaScript Removed		
n-gram Length	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1
2	0.999	0.999	0.999	0.999	0.999	0.999	0.989	0.988	0.988
3	0.997	0.997	0.997	0.998	0.998	0.998	0.987	0.986	0.987
4	0.997	0.996	0.997	0.998	0.998	0.998	0.985	0.983	0.984
5	0.996	0.995	0.996	0.996	0.995	0.995	0.984	0.981	0.982
6	0.994	0.992	0.993	0.995	0.993	0.994	0.980	0.976	0.978
Average	0.997	0.996	0.996	0.997	0.996	0.997	0.985	0.983	0.984

Table 5.9: Mean classification results for three levels of preprocessing, using the centroid model on the 7-Genre corpus, averaged over Web page profile sizes of 5 to 50. The χ^2 statistic is used as a feature selection measure. Standard error ≤ 0.001.

KI-04 Corpus	No Preprocessing			JavaScript Removed			HTML Tags and JavaScript Removed		
n-gram Length	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1
2	0.996	0.996	0.996	0.997	0.997	0.997	0.994	0.994	0.994
3	0.996	0.996	0.996	0.996	0.996	0.996	0.988	0.988	0.988
4	0.995	0.995	0.995	0.996	0.995	0.995	0.986	0.985	0.986
5	0.994	0.994	0.994	0.993	0.993	0.993	0.985	0.984	0.985
6	0.987	0.987	0.987	0.989	0.988	0.989	0.982	0.980	0.981
Average	0.994	0.993	0.994	0.994	0.994	0.994	0.987	0.986	0.987

Table 5.10: Mean classification results for three levels of preprocessing, using the centroid model on the KI-04 corpus, averaged over Web page profile sizes of 5 to 50. The χ^2 statistic is used as a feature selection measure. Standard error ≤ 0.001.

5.3.2 Effect of Web Page Profile Size

Tables 5.11 and 5.12 give the mean classification results for the 7-Genre corpus and the KI-04 corpus respectively; the precision, recall, and F1-measure in each case is averaged over n-gram length from 2 to 6 for each Web page profile sizes of 5 to 50. As with Tables 5.9 and 5.10, Tables 5.11 and 5.12 indicate that of the three levels of preprocessing that were examined, removing both the HTML tags and JavaScript code gives the worst classification performance, while there is no significant difference between removing only the JavaScript code and doing no preprocessing at all.

On the 7-Genre corpus, Web page profile sizes of 35 gave the best classification results in terms of the precision, recall, and F1-measure, both when no preprocessing was performed and when only the JavaScript code was removed. When both the HTML tags and JavaScript code were removed as preprocessing steps, using a Web page profile size of 25 gave the best classification results. However, Scheffé post hoc tests indicated that there were no statistically significant differences in using Web page profile sizes from 15 to 50 for any of the three levels of preprocessing.

The results on the KI-04 corpus were similar. When no preprocessing was performed, using Web page profiles of size 45 gave the best results. When only JavaScript code was removed, using Web page profile sizes of 35 gave the best results, and when both HTML tags and JavaScript code were removed, using Web page profile sizes of 25 gave the best classification results. However, as with the 7-Genre corpus, Scheffé post hoc tests indicated that there was no statistically significant difference when the Web page profile size ranged from 15 to 50 for any of the three levels of preprocessing.

For both the 7-Genre and the KI-04 corpora, for each of the three preprocessing levels tested in these experiments, the effect of the Web page profile size was significant on the precision, recall, and F1-measure, at $p < 0.01$. The effect of the Web page profile size was somewhat greater on the 7-Genre corpus than on the KI-04 corpus. The partial η^2, computed during the ANOVA, is greater than 0.44 for the 7-Genre corpus and greater than 0.30 for the KI-04 corpus. This effect, however, seems to be largely due to the relatively poor performance of the centroid model when Web page profile sizes of 5 were used. For both the 7-Genre and KI-04 corpora, using a Web page profile size of 5 resulted in significantly worse precision, recall, and F1-measure values for each level of preprocessing ($p < 0.01$).

7-Genre Corpus	No Preprocessing			JavaScript Removed			HTML Tags and JavaScript Removed		
Web Page Profile Size	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1
5	0.980	0.974	0.977	0.982	0.977	0.979	0.961	0.950	0.956
10	0.994	0.993	0.993	0.994	0.994	0.994	0.982	0.980	0.981
15	0.998	0.998	0.998	0.998	0.998	0.998	0.988	0.987	0.987
20	0.999	0.999	0.999	0.999	0.999	0.999	0.990	0.989	0.989
25	0.999	0.999	0.999	0.999	0.999	0.999	0.990	0.989	0.990
30	0.999	0.999	0.999	0.999	0.999	0.999	0.989	0.988	0.988
35	0.999	0.999	0.999	1.000	1.000	1.000	0.988	0.987	0.988
40	0.999	0.999	0.999	0.999	0.999	0.999	0.989	0.988	0.988
45	0.999	0.999	0.999	0.999	0.999	0.999	0.987	0.986	0.987
50	0.999	0.999	0.999	0.999	0.999	0.999	0.985	0.984	0.985
Average	0.997	0.996	0.996	0.997	0.996	0.997	0.985	0.983	0.984

Table 5.11: Mean classification results for three levels of preprocessing, using the centroid model on the 7-Genre corpus, averaged over n-gram lengths of 2 to 6. The χ^2 statistic is used as a feature selection measure. Standard error ≤ 0.001.

KI-04 Corpus	No Preprocessing			JavaScript Removed			HTML Tags and JavaScript Removed		
Web Page Profile Size	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1
5	0.982	0.981	0.982	0.982	0.980	0.981	0.968	0.963	0.966
10	0.990	0.989	0.990	0.990	0.989	0.989	0.983	0.982	0.982
15	0.992	0.992	0.992	0.993	0.993	0.993	0.989	0.989	0.989
20	0.995	0.995	0.995	0.996	0.996	0.996	0.990	0.990	0.990
25	0.996	0.996	0.996	0.996	0.995	0.996	0.991	0.990	0.991
30	0.996	0.996	0.996	0.997	0.997	0.997	0.990	0.990	0.990
35	0.997	0.997	0.997	0.997	0.997	0.997	0.990	0.990	0.990
40	0.997	0.997	0.997	0.997	0.997	0.997	0.990	0.990	0.990
45	0.997	0.997	0.997	0.997	0.997	0.997	0.990	0.990	0.990
50	0.996	0.997	0.997	0.997	0.997	0.997	0.990	0.990	0.990
Average	0.994	0.993	0.994	0.994	0.994	0.994	0.987	0.986	0.987

Table 5.12: Mean classification results for three levels of preprocessing, using the centroid model on the KI-04 corpus, averaged over n-gram lengths of 2 to 6. The χ^2 statistic is used as a feature selection measure. Standard error ≤ 0.001.

5.3.3 Data Preprocessing Revisited: Conclusions

This set of experiments investigated the effects of common data preprocessing steps when the χ^2 statistic was used as the feature selection measure for selecting the n-grams with which to represent Web pages with the centroid classification model. Experiments were run using three different levels of preprocessing on the 7-Genre and KI-04 corpora; these three levels were no preprocessing, removing only the JavaScript code, and removing both the HTML tags and JavaScript code. The n-gram length was varied from 2 to 6 in increments of 1, and for each n-gram length, the Web page profile size was varied from 5 to 50 in increments of 5.

Tables 5.9–5.12 show that although the centroid model achieves high mean classification performance with each of the three preprocessing levels, for both the 7-Genre and KI-04 corpora the centroid model achieves significantly worse classification performance when both the HTML tags and JavaScript code are removed from the Web pages ($p < 0.01$), while there is no significant difference between removing only the JavaScript code from the Web pages and doing no preprocessing at all. This is in contrast to the results of the preprocessing study discussed in Section 5.1 in which n-gram frequency was used as the feature selection measure, and is an indication of the superior ability of the χ^2 statistic to select n-grams with high discriminatory powers.

The presence or absence of JavaScript code in a Web page has no significant influence on the classification performance of the centroid model in these experiments, however JavaScript code is found in only 78% of the 7-Genre Web pages and in only 53% of the KI-04 Web pages. All of the Web pages in each corpus contain HTML tags, and removing these tags is detrimental to the performance of the centroid classification model when the χ^2 statistic is used as a feature selection measure. This is an indication that using the χ^2 statistic as a feature selection measure allows the selection of n-grams from the HTML tags which are helpful in determining the genre of the Web page.

These experiments have shown that the centroid classification model achieves high classification performance when each of three different levels of preprocessing are used, over a variety of n-gram lengths and Web page profile sizes. Based on the results of these experiments, preprocessing Web pages to remove HTML tags and/or JavaScript code is unnecessary, therefore future work with the n-gram representation of Web pages for genre classification will be carried out with no preprocessing of the Web pages.

5.4 Conclusions

This chapter has discussed three studies which furthered the investigation as to whether an n-gram representation of a Web page can be used effectively to classify the Web page by its genre. The major contribution of these studies is to show that an n-gram approach to Web page representation for genre classification with the new centroid model gives high mean classification performance under a variety of conditions, on more than one corpus. The excellent classification results support the hypothesis that Web pages of the same genre share a distribution of n-grams that is similar enough to allow n-gram representations of the Web pages to be used to classify the Web pages by their genres. Using n-gram representations of the Web pages, the centroid model achieved high classification performance on both the 7-Genre and KI-04 corpora, suggesting that n-gram representation is a robust method of representing Web pages for classification, and that it should be tested on more challenging corpora and with other classification models, such as the support vector machine classifier.

The study discussed in Section 5.1 investigated the effects of common data preprocessing steps when using an n-gram approach to Web page representation with the centroid classification model. Experiments were run with three different levels of preprocessing on the 7-Genre and KI-04 corpora; these three levels were no preprocessing, removing only the JavaScript code, and removing both the HTML tags and JavaScript code. The results indicate that for both the 7-Genre and KI-04 corpora, the classification performance of the centroid model increased with each level of preprocessing: doing no preprocessing gave poorer results than did removing the JavaScript code from the Web pages, and removing both the HTML tags and the JavaScript code further improved the classification performance of the model. However, these differences were not statistically significant on the KI-04 corpus, and on the 7-Genre corpus there was no statistically significant difference between removing only the JavaScript code and removing both the HTML tags and the JavaScript code. The results of this study indicate that smaller Web page profiles can be advantageous when the feature set has been reduced by preprocessing, which led to the question of whether reducing the feature set by other means could also be effective.

Section 5.2 described the exploration of feature selection measures, comparing the classification results for the centroid model when n-gram frequency, Information Gain, and the χ^2 statistic were used as feature selection measures. Although all three of the measures performed well, the χ^2 statistic significantly outperformed the other measures on both the 7-Genre and KI-04 corpora. These results illustrate the ability of the χ^2 statistic to identify small feature sets of relatively short n-grams to represent Web pages in the centroid classification model. Based on these results, the χ^2 statistic will be used as the feature selection measure in future studies, the n-gram length will be limited to the range from 2 to 6, and the Web page profile size used in future experiments will be limited to the range from 5 to 50.

The study discussed in Section 5.3 revisited the question of data preprocessing, based on the results of the feature selection experiments. In contrast to the results in Section 5.1 in which n-gram frequency was used as a feature selection measure, this study showed that for both the 7-Genre and KI-04 corpora, the centroid model achieved significantly worse classification performance when both the HTML tags and JavaScript code are removed from the Web pages ($p < 0.01$), while there was no significant difference between removing only the JavaScript code from the Web pages and doing no preprocessing at all. This is an indication that using the χ^2 statistic as a feature selection measure allows the selection of n-grams from the HTML tags which are helpful in determining the genre of the Web page. Based on this study, preprocessing Web pages to remove HTML tags and/or JavaScript code is unnecessary, therefore future work on genre classification will be carried out without preprocessing the Web pages.

Chapter 6 will discuss the next phase of this research, which broadens the scope of the work to include more challenging corpora. This includes an investigation of techniques for setting genre thresholds in order to allow a Web page to belong to more than one genre, or to no genre at all, and a comparison of the classification performance of the centroid model with that of the popular support vector machine approach. Experiments conducted on highly unbalanced corpora, both with and without the inclusion of noise Web pages, will also be presented.

Chapter 6

Classification Experiments: Unbalanced Corpora

The studies presented in Chapter 5 focused on data preprocessing and feature selection measures using the centroid classification model and byte n-gram representations of the Web pages. These experiments were run on the 7-Genre and KI-04 corpora, which are both single-label and balanced. In order to answer the question of whether the proposed classification approach is effective on highly unbalanced corpora, this chapter expands the research focus to include two highly unbalanced single-label corpora, as well as a multi-label, unbalanced corpus.

Section 6.1 examines the classification performance of the centroid model on the extremely unbalanced 15-genre corpus. The research question of whether a threshold value can be determined for each genre in order to allow a Web page to belong to more than one genre (or to no genre at all) is investigated in Section 6.2, which looks at techniques for setting genre thresholds. This section also answers the question of whether the proposed classification model gives comparable performance to a support vector machine approach, by making a comparison of the classification performance of these two models. Section 6.3 investigates the research question of whether the centroid model is effective when noise Web pages that do not belong to any genre in a particular corpus are included in that corpus. The chapter concludes with a summary of the experiments and results, given in Section 6.4.

6.1 Classification Performance on the Unbalanced 15-Genre Corpus

The goal of this thesis is to develop an approach to the problem of Web page genre classification that is effective not only on balanced, single-label corpora, but also on unbalanced and multi-label corpora, which better represent a real world environment. This section presents experiments testing the centroid classification model on the highly unbalanced 15-Genre corpus, while Section 6.2 will test the centroid model on a corpus that is both unbalanced and multi-label. The 15-Genre corpus was created for this thesis by combining the 7-Genre and KI-04 collections as described in Section 3.1.3. This corpus has a Zipf-like distribution, with 15 genres ranging in size from 489 to 30. In these experiments, the n-gram length is varied from 2 to 4, and for each n-gram length the Web page profile size is varied from 15 to 50.

6.1.1 Effect of n-gram Length

Table 6.1 gives the mean classification results for the 15-Genre corpus. The precision, recall, and F1-measure in each case is averaged over Web page profile sizes of 15 to 50, for each n-gram length from 2 to 4. These results indicate that the centroid classification model performs extremely well on this unbalanced corpus.

The effect of the n-gram length on the precision, recall, and F1-measure for the classification performance of the centroid model on the 15-Genre corpus is statistically significant ($p < 0.01$). However, the partial η^2 in each case is less than 0.15, indicating that the proportion of total variability in the precision, recall, and F1-measure is only moderately influenced by the n-gram length. An n-gram length of 2 gives the best precision, recall, and F1-measure values. In terms of the precision, using an n-gram length of 2 gives significantly better results than using an n-gram length of 3, but there is no significant difference between using n-gram lengths of 2 and 4; in terms of the recall and F1-measure, using an n-gram length of 2 gives significantly better results than using n-gram length of 3 or 4.

15-Genre Corpus	Centroid Model		
n-gram Length	Precision	Recall	F1
2	0.972	0.974	0.972
3	0.930	0.883	0.892
4	0.954	0.925	0.929
Average	0.952	0.925	0.929

Table 6.1: Mean classification results for the centroid model on the 15-Genre corpus, averaged over Web page profile sizes of 15 to 50. Standard error ≤ 0.008.

6.1.2 Effect of Web Page Profile Size

Table 6.2 gives the mean precision, recall, and F1-measure for each Web page profile size, averaged over n-gram lengths from 2 to 4. As did Table 6.1, Table 6.2 shows that the centroid classification model performs extremely well on this unbalanced corpus.

In these experiments, the number of n-grams used to represent each Web page ranges from 15 to 50 in increments of 5. The effect of the Web page profile size was not statistically significant on the precision, recall, or F1-measure for the classification performance of the centroid model on the 15-Genre corpus. This is consistent with the results of the study discussed in Section 5.3, and indicates that the centroid model is very stable over these Web page profile sizes.

15-Genre Corpus	Centroid Model		
n-gram Length	Precision	Recall	F1
15	0.947	0.915	0.921
20	0.953	0.925	0.930
25	0.949	0.921	0.926
30	0.950	0.921	0.926
35	0.948	0.922	0.925
40	0.956	0.928	0.932
45	0.959	0.934	0.938
50	0.956	0.933	0.935
Average	0.952	0.925	0.929

Table 6.2: Mean classification results for the centroid model on the 15-Genre corpus, averaged over n-gram lengths of 2 to 4. Standard error ≤ 0.013.

6.1.3 Effect of Genre

Table 6.3 gives a comparison of the mean precision, recall, and F1-measure of the centroid model on the 15-Genre corpus, broken down by genre. The results are averaged over n-gram lengths from 2 to 4 and Web page profile sizes of 15 to 50. These results indicate that some genres are easier to classify than others, but the number of Web pages in the genre does not appear to have an appreciable effect on this outcome.

ANOVA indicates that genre is the leading factor (over n-gram length, and Web page profile size) in predicting the outcome for the precision, recall and F1-measure. Although the partial η^2 computed during the ANOVA is quite small for the precision, at 0.071, it is 0.673 for the recall and 0.588 for the F1-measure. This indicates that although genre accounted for almost 70% of the variance in the recall and almost 60% of the variance in the F1-measure, genre accounted for only about 7% of the variance in the precision. The lesser influence of genre on the precision could be due to a ceiling effect, caused by the high precision values. Although genre is an influential factor in predicting the classification performance, the development of a specific hypothesis about which genres can be better classified than others is outside of the scope of this thesis.

6.1.4 The Unbalanced 15-Genre Corpus: Conclusions

One of the goals of this thesis is to develop an approach to the problem of Web page genre classification that is effective on both balanced and unbalanced corpora, and one of the specific research questions to be answered is whether the proposed centroid classification model is effective on highly unbalanced corpora. Tables 6.1–6.3 indicate that using n-gram representations of Web pages and the χ^2 statistic as a feature selection measure with the centroid model achieves effective classification performance on the highly unbalanced 15-genre corpus. Although the recall values are somewhat lower than the precision values, in applications such as online searching, precision is typically considered to be more important than recall; a relevant Web page not being returned by a search presents less of a concern for most users than a non-relevant Web page that is returned. In the next study, discussed in Section 6.2, the centroid model is tested on a corpus that is both unbalanced and multi-label.

Genre	Number of Web Pages	Precision	Recall	F1
TABLE	30	0.943	0.986	0.960
SITEMAP	33	0.708	0.444	0.511
HOTLIST	36	1.000	0.912	0.939
CHECKLIST	39	0.999	0.916	0.944
DISCUSSION	45	1.000	0.872	0.925
ARTICLE	48	0.976	0.964	0.968
HELP	54	0.981	0.831	0.897
DOWNLOAD	60	0.933	0.981	0.954
SEARCH PAGE	69	0.947	0.989	0.966
ONLINE NEWSPAPER FRONT PAGE	81	0.996	0.998	0.997
FAQ	99	1.000	1.000	1.000
BLOG	123	0.895	1.000	0.941
LINK COLLECTION	162	0.914	1.000	0.954
SHOPPING	243	0.997	0.990	0.993
HOMEPAGE	489	0.994	0.984	0.989
Average	107	0.952	0.925	0.929

Table 6.3: Mean classification results by genre for the centroid model on the 15-Genre corpus, averaged over Web page profile sizes of 15 to 50 and n-gram lengths of 2 to 4. Standard error ≤ 0.017.

6.2 Multi-labeling with the Unbalanced 20-Genre Corpus

As discussed in Section 2.5, the difficulty of assigning a single genre label to a Web page has been acknowledged by many of the researchers who have conducted surveys and user studies about Web page genre [33, 34, 84, 98, 99, 108]. The goal of this thesis is to develop an approach to the problem of Web page genre classification that is effective not only on balanced, single-label corpora, but also on unbalanced and multi-label corpora, which better represent a real world environment. The specific research question of whether a threshold value can be determined for each genre in order to allow a Web page to belong to more than one genre (or to no genre at all) is investigated by comparing the results of experiments in which different techniques are used for setting genre thresholds. This section also answers the question of whether the proposed centroid classification model gives comparable performance to a support vector machine approach, by making a comparison of the classification performance of these two models. The experiments discussed in this section are run on the 20-Genre corpus, which is both multi-label and unbalanced. See Section 3.1.4 for a detailed description of the corpus.

6.2.1 Setting Thresholds

In order to classify a Web page as belonging to more than one genre, or as not belonging to any known genre, the centroid classification model is modified to include threshold values that are computed for each genre. If the distance between the Web page profile and a genre profile is less than or equal to the threshold, then the Web page is labeled as belonging to that genre. If the distance is greater than the threshold, then the Web page is deemed not to belong to that genre. This addition of genre thresholds to the centroid classification model changes the architecture of the model from that of one 20-way classifier to that of twenty 2-way classifiers. Two methods of setting the genre thresholds are investigated: the distribution threshold method and the optimal threshold method. The classification performance of the centroid model using each of these methods is compared with the performance of a support vector machine; n-gram representations of the Web pages are used in each case.

Distribution Threshold Method

One method of determining the threshold for each Web page genre is to base the threshold on the distribution of the Web pages that belong to that genre, in the training set. A normal, or Gaussian, distribution is a distribution which is bell-shaped, with a peak at the mean.

For each genre in the corpus, the distance is computed between the centroid profile for that genre and the profile of each Web page of that genre in the training set. If this set of distances has a normal distribution, that genre threshold is set at the 85th percentile (one standard deviation above the mean), such that 85% of the Web pages belonging to the genre (in the training set) fall within the threshold. If the set of distances is not normally distributed, the threshold is set at the 75th percentile. The 75th percentile was chosen as a cutoff based on the popular use of the semi-interquartile range statistic for non-normal data. This statistic is also appropriate for normal data, but is less powerful than the standard deviation statistic for such distributions [120]. Because each genre threshold is set based on the distribution of the Web pages belonging to the genre, this will be referred to as the *distribution threshold method*.

Optimal Threshold Method

Another method of setting each genre threshold is to first order all of the Web pages in the training set according to their distance from a particular genre profile, in ascending order. This ordered list of Web pages from the training set can then be stepped through one Web page at a time, such that at each step, the current Web page is labeled as belonging to the genre in question, and the accuracy of the classification thus far is computed. In this manner, the optimal threshold for each genre, based on the training data, can be determined. This process is then repeated for each genre in the corpus. This method of setting the genre thresholds will be referred to as the *optimal threshold method*, because the method gives a set of fixed thresholds, one for each genre, that give the optimal classification accuracy on the training set. The pseudocode for this algorithm is given in Figure 6.1.

```
**Input**  : set of training document profiles, centroid profile for genre $A$
**Output**: optimal threshold for genre $A$

**Initialize:**

$PA$ = centroid profile for genre $A$
$NA$ = number of documents in genre $A$
$N$ = total number of training documents
$numCorrect = 0$
$optAccuracy = 0$
$opti = 1$

**Sort:**

sort all training document profiles into array $d\,[1 \cdots N]$ according to their
distance from $PA$, such that $\texttt{distance}\,(PA, d\,[\,i\,]) \leq \texttt{distance}\,(PA, d\,[i+1])$

**Compute Optimal Threshold:**

for ($i = 1$ to $N$) do
    if ($d\,[\,i\,] \in A$) then
        $numCorrect = numCorrect + 1$
    end
    $currentAccuracy = (numCorrect + ((N - NA) - (\,i\, - numCorrect)))\,/\,N$
    if ($currentAccuracy \geq optAccuracy$) then
        $optAccuracy = currentAccuracy$
        $opti = i$
    end
    if ($numCorrect == NA$) then
        break
    end
end
if ($i == N$) then
    return $\texttt{distance}\,(PA, d\,[N])$
else
    return $(\texttt{distance}\,(PA, d\,[opti]) + \texttt{distance}\,(PA, d\,[opti+1]))\,/\,2$
end
```

Figure 6.1: Optimal threshold algorithm.

Support Vector Machine Method

An alternative method of assigning more than one label to a Web page is to use the support vector machine (SVM) classification model, rather than the centroid classification model. The SVM method is a popular and well-known supervised machine learning approach developed by Vladimir Vapnik and his co-workers at AT&T Bell Labs [29, 134]. The SVM method performs classification by constructing an N-dimensional hyperplane that optimally separates the data, making the assumption that the larger the distance between the data and the hyperplane, the better the performance of a classifier will be. The vectors near the hyperplane are the support vectors.

Although the Naïve Bayes classifier is often used as a baseline for comparisons in text classification problems, the SVM classifier has been shown to outperform the Naïve Bayes classifier [55, 102, 110, 141], therefore the SVM model was chosen as a basis of comparison for the centroid classification model. This comparison serves two purposes. The first purpose of this comparison is to evaluate the n-gram representation of Web pages on a classification model other than the k-NN and centroid models. The second purpose of this comparison is to answer the research question of whether or not the proposed classification model gives comparable performance to the SVM approach, using the same n-gram representations of the Web pages. Because these experiments are not focused on achieving the best possible performance using the SVM classifier, the default parameter settings are used for the SVM classifier. Had the focus of these experiments been on optimizing the performance of the SVM classifier, it seems likely that the classifier could have achieved even better performance.

For the experiments using the SVM method in this thesis, the sequential minimal optimization (SMO) method is used to find the separating hyperplane. The SMO method essentially breaks up large quadratic programming problems into a series of the smallest possible problems, which are then solved analytically [90].

For these experiments, multiple binary SVM classifiers are trained individually and the outputs of the classifiers are combined for classification of multiple genres; thus, for a classification problem with twenty genres, twenty SVM classifiers are trained using the conventional *one-against-all* approach. With the one-against-all approach, each binary SVM classifier separates one class from all the rest of the training data.

This means that the ith SVM classifier, where $1 \leq i \leq |G|$ and $|G|$ is the number of genres, is trained using all of the training examples of the ith genre with positive labels, and using all of the training examples from the remaining genres with negative labels.

6.2.2 Experiments

These experiments are carried out on the multi-label 20-Genre corpus, using the centroid classification model with each of the two threshold methods described in Section 6.2.1, as well as the SVM classifier discussed in the same section. In each case, the Web pages are represented by profiles composed of byte n-grams and their associated frequencies.

Based on the results discussed in Chapter 5, the χ^2 statistic is used as the feature selection measure, and no preprocessing is performed on the Web pages. Also based on the results in Chapter 5, the n-gram length is varied from 2 to 4 in increments of 1, and for each n-gram length, the Web page profile size is varied from 15 to 50 in increments of 5.

As discussed in Section 3.1.4, 3-fold cross-validation is used for each experiment with this corpus. The results for all of the iterations are then averaged to give the final results. This helps provide robustness against overfitting and gives additional strength to the statistical analysis.

6.2.3 Overall Results

Tables 6.4 and 6.5 give the mean classification results for the three classification methods on the 20-Genre corpus, averaged over Web page profile sizes of 15 to 50 and n-gram lengths of 2 to 4 respectively. In each case, the precision, recall, and F1-measure values that have been averaged over Web page profile sizes or n-gram lengths are themselves macro-averages; macro-averages give equal weight to each genre, whereas micro-averages give equal weight to each Web page.

Based on both ANOVA and Scheffé post hoc testing, the classification performance of both the SVM approach and the centroid model with the optimal threshold method is significantly better than that of the centroid model when the distribution threshold method is used, in terms of the precision, recall, and F1-measure, with $p < 0.01$ for each metric. The p-value refers to the probability that the observed results could have occurred by chance if the means of two groups are equal. A small p-value leads to the conclusion that the means are different, with a p-value of less than 0.05 considered to indicate statistical significance.

Based on the precision, there is no statistically significant difference between the SVM method and the centroid model with the optimal threshold method, however in terms of the recall and F1-measure, the classification performance of the latter method is significantly better than that of the SVM method ($p < 0.01$). Thus, the new centroid classification model achieves classification performance comparable to that of the SVM model, using a less complex and more easily scalable approach.

20-Genre Corpus	Distribution Threshold Method			Optimal Threshold Method			Support Vector Machine		
n-gram Length	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1
2	0.987	0.409	0.546	0.992	0.768	0.855	0.989	0.720	0.808
3	0.961	0.374	0.504	0.990	0.764	0.851	0.998	0.702	0.798
4	0.916	0.358	0.481	0.989	0.731	0.823	0.998	0.662	0.768
Average	0.955	0.380	0.510	0.990	0.754	0.843	0.995	0.695	0.791

Table 6.4: Mean classification results on the 20-Genre corpus, averaged over Web page profile sizes of 15 to 50. The χ^2 statistic is used as a feature selection measure. Standard error ≤ 0.005.

20-Genre Corpus	Distribution Threshold Method			Optimal Threshold Method			Support Vector Machine		
n-gram Length	Prec.	Rec.	F1	Prec.	Rec.	F1	Prec.	Rec.	F1
15	0.926	0.368	0.494	0.987	0.725	0.816	0.995	0.691	0.789
20	0.971	0.371	0.499	0.989	0.725	0.816	0.996	0.695	0.791
25	0.950	0.371	0.498	0.991	0.764	0.852	0.995	0.694	0.790
30	0.964	0.383	0.515	0.991	0.765	0.852	0.996	0.696	0.793
35	0.962	0.381	0.513	0.994	0.767	0.855	0.994	0.696	0.792
40	0.957	0.388	0.521	0.991	0.765	0.852	0.994	0.697	0.793
45	0.951	0.388	0.519	0.991	0.764	0.851	0.996	0.696	0.792
50	0.956	0.391	0.522	0.988	0.761	0.848	0.994	0.696	0.792
Average	0.955	0.380	0.510	0.990	0.754	0.843	0.995	0.695	0.791

Table 6.5: Mean classification results on the 20-Genre corpus, averaged over n-gram lengths of 2 to 4. The χ^2 statistic is used as a feature selection measure. Standard error ≤ 0.010.

6.2.4 Effect of n-gram Length

Table 6.4 gives the mean classification results for the 20-Genre corpus. The precision, recall, and F1-measure in each case is averaged over Web page profile sizes of 15 to 50, for each n-gram length from 2 to 4. Table 6.4 shows that as the n-gram length is increased, the precision, recall and F1-measure values decrease for each method, with the exception of the precision values for the SVM method. Although n-gram lengths of 3 and 4 give the best precision for the SVM method, the results of Scheffé post hoc multiple comparison tests show that for each method overall, n-grams of length 2 are the best choice.

The effect of the n-gram length on the precision, recall, and F1-measure for each method is statistically significant ($p < 0.01$), however the partial η^2, computed during the ANOVA, is less than 0.10 in each case. This indicates that the proportion of total variability in the precision, recall, and F1-measure for each method is only slightly influenced by the n-gram length.

6.2.5 Effect of Web Page Profile Size

Table 6.5 gives the mean precision, recall, and F1-measure for each Web page profile size, for each method. These results are averaged over n-gram lengths from 2 to 4. In these experiments, the number of n-grams used to represent each Web page ranges from 15 to 50 in increments of 5, however the effect of the Web page profile size was not statistically significant on the precision of any of the three methods. The effect of the Web page profile size was found to be statistically significant on the recall and F1-measure for the centroid model when the optimal threshold method was used ($p < 0.01$), however the partial η^2 was less than or equal to 0.05, indicating that the Web page profile size accounted for at most 5% of the overall variance of the recall and F1-measure.

6.2.6 Effect of Genre

Table 6.6 gives a comparison of the mean precision and recall for each genre, for each of the three methods tested in these experiments. Table 6.6 also gives the best results of Vidulin et al. [136] using the same corpus.

Vidulin et al. represent Web pages using a set of 502 genre features that are a combination of HTML-based, URL-based, and text-based features. They used J48, the Weka implementation of the C4.5 algorithm, for constructing the classifier and bagging ensembles. The results from Vidulin et al. that are shown in Table 6.6 are for their use of the bagging algorithm, using 10-fold cross-validation.

20-Genre Corpus	Vidulin et al. [136]		Distribution Threshold Method		Optimal Threshold Method		Support Vector Machine	
Genre	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.	Prec.	Rec.
OFFICIAL	0.73	0.27	1.000	0.486	0.988	0.924	0.996	0.752
SHOPPING	0.72	0.33	0.694	0.137	1.000	0.590	1.000	0.715
PROSE FICTION	0.69	0.30	0.833	0.190	0.966	0.751	1.000	0.878
ADULT	0.78	0.71	1.000	0.379	1.000	0.671	1.000	0.253
FAQ	0.98	0.73	1.000	0.495	0.998	0.971	1.000	0.840
POETRY	0.76	0.61	1.000	0.348	0.996	0.910	0.993	0.874
ENTERTAINMENT	0.69	0.27	0.897	0.271	0.969	0.653	0.992	0.693
SCIENTIFIC	0.85	0.51	0.917	0.307	0.989	0.955	1.000	0.860
BLOG	0.83	0.56	0.986	0.497	0.998	0.762	1.000	0.291
GATEWAY	0.45	0.12	0.960	0.431	0.986	0.587	0.994	0.712
ERROR MESSAGE	0.87	0.68	0.986	0.541	0.993	0.981	0.998	0.895
COMMUNITY	0.76	0.55	0.969	0.440	0.979	0.792	1.000	0.735
USER INPUT	0.83	0.57	1.000	0.440	0.995	0.756	1.000	0.691
CHILDREN'S	0.81	0.48	1.000	0.599	0.996	0.860	0.969	0.130
PERSONAL	0.72	0.16	0.994	0.434	0.999	0.572	0.993	0.734
COMMERCIAL/PROMO.	0.40	0.04	0.978	0.514	0.978	0.787	0.982	0.599
CONTENT DELIVERY	0.64	0.23	0.917	0.099	0.989	0.450	0.982	0.805
JOURNALISTIC	0.62	0.36	1.000	0.312	0.994	0.903	1.000	0.858
INFORMATIVE	0.30	0.09	0.958	0.252	1.000	0.699	0.995	0.816
INDEX	0.63	0.37	1.000	0.433	0.991	0.514	0.997	0.768
Average	0.70	0.40	0.955	0.380	0.990	0.754	0.995	0.695

Table 6.6: Mean classification results by genre on the 20-Genre corpus. The results for the centroid and SVM models are averaged over Web page profile sizes of 15 to 50 and n-gram lengths of 2 to 4. Standard error ≤ 0.016.

The results in Table 6.6 indicate, not surprisingly, that some genres are easier to classify than others. ANOVA indicates that genre is the leading factor (over method, n-gram length, and Web page profile size) in predicting the outcome for the precision, and is second only to method in predicting the outcome for the recall and F1-measure. Although genre is an influential factor in predicting the classification performance, a specific hypothesis about which genres can be better classified than others has not been developed. There is no observable trend that relates this influence on variability to the number of Web pages in each genre, and as shown in Table 6.6, the results on a particular genre vary depending on the classification method; a genre that is well classified by one method may be very poorly classified by another method. The variability between genres is likely to be caused by a factor that has not been explored as part of the current research, such as the length of the Web pages in each genre, or the homogeneity of each genre.

Of the three classification methods tested in these experiments, only one labeled any of the Web pages as not belonging to any recognized genre. The SVM and optimal threshold methods both supplied at least one genre label for each Web page, however the distribution threshold method did not. In this case, an average of almost 50% of the Web pages were labeled as not belonging to any of the 20 known genres in the corpus. This indicates that the genre thresholds were being set too strictly, and helps account for the very low recall values returned by this method, however setting the thresholds less strictly results in low precision. The Web pages that were unidentifiable by the distribution threshold method were of all genres, and the number of Web pages from each genre that were unidentifiable tended to be in proportion with the size of the genre.

6.2.7 Multi-labeling with the 20-Genre Corpus: Conclusions

This section has investigated techniques for setting genre thresholds in order to allow a Web page to belong to more than one genre, or to no genre at all, and a comparison of the classification performance of the centroid model with that of the popular SVM approach was also made. The results in Tables 6.4–6.6 show that each method has a much higher precision than recall, averaged over all 20 genres. This means that the genre labels assigned by the classifiers are quite accurate, but that these machine learning classifiers are not assigning as many labels as did the human annotators when the corpus was constructed. The mean precision and recall for these methods exceeds the best results reported by Vidulin et al. [136].

Based on the results presented in this section, it is reasonable to conclude that the n-gram representation of Web pages allows the effective classification of those Web pages by genre, even when the Web page belongs to more than one genre. The combination of the centroid classification model and the optimal threshold method is more successful than the SVM method in classifying this multi-label corpus. The results of these experiments also showed that in general, as the length of the n-grams used to represent the Web pages was increased, the classification performance for each model decreased. Based on these results, a short n-gram length will be used in future work. The results also indicated that over the range of 15 to 50, the number of n-grams used to represent each Web page has only a slight impact on the classification results, therefore this range is considered sufficient for use in future work.

The major contribution of this study is to show that byte n-gram Web page representations can be used effectively, with more than one classification model, to classify Web pages by genre, even when the Web pages belong to more than one genre. The study discussed in Section 6.3 will test the centroid model on a highly unbalanced corpus to which noise has been added. For the purpose of this study, noise will be defined as any Web page belonging to a Web page genre that is not one of the genres in the corpus. The optimal threshold method will be used as part of the centroid model to identify noise Web pages during the classification process.

6.3 The Unbalanced Syracuse Corpus and Noise

The goal of this thesis is to develop an approach to the problem of Web page genre classification that is effective not only on balanced corpora, but also on unbalanced corpora which better represent a real world environment. This section presents a study testing the centroid model on the highly unbalanced Syracuse corpus. This corpus of 1985 Web pages has a Zipf-like distribution, with 24 genres ranging in size from 489 to 30. In order to investigate the research question of whether the centroid model is effective when noise Web pages that do not belong to any genre in a particular corpus are included in the corpus, this study compares the classification performance of the centroid model on this corpus with that of the same model when 750 noise Web pages have been added to the corpus. This gives a total of 2735 Web pages. For the purpose of this thesis, a noise Web page is a Web page that does not belong to any genre in the corpus in question. The noise Web pages that are added in this case are those that belong to the 94 genres in the original corpus that contained fewer than 30 Web pages each; see Section 3.1.5 for more details about the corpus.

Because the Web pages in this corpus have only one genre label, the centroid model is modified from Section 6.2 such that each Web page is assigned at most one label. The optimal threshold method is used to determine thresholds for each genre, and a Web page is assigned the label of the genre to which it is most similar, within these thresholds. If the Web page is not within the threshold for any genre, it is labeled as a noise Web page. In these experiments, the n-gram length is varied from 2 to 4, and for each n-gram length the Web page profile size is varied from 15 to 50.

6.3.1 Effect of n-gram Length

Table 6.7 gives the mean classification results for the centroid model on the Syracuse corpus, with and without noise Web pages. The precision, recall, and F1-measure in each case is averaged over Web page profile sizes of 15 to 50, for each n-gram length from 2 to 4. These results indicate that the addition of the noise Web pages decreases the precision of the classification, but increases the recall; this means that the noise Web pages are less likely to be mislabeled as belonging to another genre than are the non-noise Web pages. In each case, an n-gram length of 2 gives the best precision, recall, and F1-measure values.

The effect of the n-gram length on the precision, recall, and F1-measure for the classification performance of the centroid model on the Syracuse corpus, both with and without noise Web pages, is statistically significant ($p < 0.01$). For the precision, the partial η^2 in each case is less than 0.10, indicating that the proportion of total variability in the precision is only slightly influenced by the n-gram length. For the recall and F1-measure, however, the partial η^2 in each case is at least 0.38, indicating that the n-gram length has a more pronounced effect on these measures.

Syracuse Corpus	Without Noise Web Pages			With Noise Web Pages		
n-gram Length	Precision	Recall	F1	Precision	Recall	F1
2	0.998	0.947	0.970	0.990	0.949	0.966
3	0.993	0.821	0.876	0.980	0.828	0.875
4	0.963	0.737	0.810	0.954	0.747	0.812
Average	0.985	0.835	0.885	0.975	0.841	0.884

Table 6.7: Mean classification results for the centroid model on the Syracuse corpus, averaged over Web page profile sizes of 15 to 50. Standard error ≤ 0.005.

6.3.2 Effect of Web Page Profile Size

Table 6.8 gives the mean precision, recall, and F1-measure for each Web page profile size from 15 to 50, averaged over n-gram lengths from 2 to 4. Table 6.8 shows that the classification performance of the centroid model is very stable over these Web page profile sizes.

The effect of the Web page profile size was not statistically significant on the precision of the classification performance of the centroid model on the Syracuse corpus, with or without noise Web pages, although it was statistically significant on the recall and F1-measure ($p < 0.01$). The p-value refers to the probability that the observed results could have occurred by chance if the means of two groups are equal, with a small p-value leading to the conclusion that the means are different.

Despite this statistical significance, the partial η^2 was at most 0.023, indicating that less than 3% of the variability in the recall and F1-measure could be attributed to the Web page profile size. ANOVA indicates that varying the Web page profile size from 25 to 50 causes no statistically significant difference on the precision, recall, and F1-measure, while using Web page profile sizes of 15 and 20 does cause a significant difference in these measures ($p < 0.01$).

Syracuse Corpus	Without Noise Web Pages			With Noise Web Pages		
Web Page Profile Size	Precision	Recall	F1	Precision	Recall	F1
15	0.979	0.805	0.861	0.966	0.813	0.859
20	0.978	0.818	0.871	0.968	0.825	0.870
25	0.985	0.834	0.885	0.974	0.840	0.884
30	0.987	0.843	0.893	0.977	0.849	0.892
35	0.989	0.847	0.896	0.980	0.853	0.896
40	0.986	0.845	0.891	0.976	0.851	0.891
45	0.983	0.843	0.891	0.974	0.849	0.890
50	0.992	0.845	0.894	0.981	0.851	0.892
Average	0.985	0.835	0.885	0.975	0.841	0.884

Table 6.8: Mean classification results for the centroid model on the Syracuse corpus, averaged over n-gram lengths of 2 to 4. Standard error ≤ 0.009.

6.3.3 Effect of Genre

Table 6.9 gives a comparison of the mean precision, recall, and F1-measure of the centroid model on the Syracuse corpus with and without noise Web pages, broken down by genre. The results are averaged over n-gram lengths from 2 to 4 and Web page profile sizes of 15 to 50. As was the case with the 15-Genre and 20-Genre corpora, these results indicate that some genres are easier to classify than others, but the number of Web pages in the genre does not appear to have an appreciable effect on this outcome. ANOVA indicates that genre is the leading factor (over n-gram length, and Web page profile size) in predicting the outcome for the precision, recall and F1-measure.

Syracuse Corpus		Without Noise Web Pages			With Noise Web Pages		
Genre	Genre Size	Prec.	Rec.	F1	Prec.	Rec.	F1
INDEX TO MISC. RESOURCES	30	0.786	0.460	0.506	0.767	0.460	0.505
TABLE OF CONTENTS	33	0.993	0.828	0.888	0.993	0.828	0.888
BIBLIOGRAPHIC RECORD	34	1.000	0.833	0.901	0.965	0.833	0.885
BIOGRAPHY	34	1.000	0.707	0.814	1.000	0.707	0.814
ENCYCLOPEDIA ENTRY	35	0.999	0.803	0.887	0.979	0.803	0.878
TUTORIAL AND HOW-TO	35	0.930	0.677	0.762	0.930	0.677	0.762
DEFINITION/DESCRIPTION	35	1.000	0.577	0.698	0.995	0.577	0.697
OTHER BIOGRAPHY	37	1.000	0.873	0.927	1.000	0.873	0.927
LESSON PLAN	37	0.997	0.562	0.654	0.997	0.562	0.654
PRESS RELEASE	39	1.000	0.861	0.921	0.999	0.861	0.920
DIRECTORY OF COMPANIES	39	1.000	0.844	0.896	1.000	0.844	0.896
ABOUT A PROGRAM	48	0.983	0.741	0.823	0.980	0.741	0.822
ABOUT AN ORGANIZATION	58	0.999	0.924	0.957	0.999	0.924	0.957
FACTS-AND-FIGURES PAGE	58	0.998	0.868	0.922	0.998	0.868	0.922
DISCUSSION FORUM	61	0.997	0.917	0.953	0.996	0.917	0.952
COMPANY/ORG. HOMEPAGE	74	0.983	0.861	0.913	0.983	0.861	0.912
ADVERTISING	75	0.998	0.898	0.940	0.994	0.898	0.938
MAGAZINE ARTICLE	101	1.000	0.941	0.969	1.000	0.941	0.969
RECIPE	125	1.000	0.973	0.986	1.000	0.973	0.986
BLOG	135	0.999	0.981	0.990	0.996	0.981	0.988
DIRECTORY RESOURCES/LINKS	142	0.981	0.975	0.977	0.981	0.975	0.977
OTHER ARTICLE	177	1.000	0.971	0.985	1.000	0.971	0.985
NEWS STORY	193	0.996	0.975	0.985	0.996	0.975	0.985
PRODUCT/SERVICE PAGE	350	0.998	0.987	0.992	0.998	0.987	0.992
NOISE	750	n/a	n/a	n/a	0.821	0.996	0.898
Average	109	0.985	0.835	0.885	0.975	0.841	0.884

Table 6.9: Mean classification results by genre for the centroid model on the Syracuse corpus. The results are averaged over Web page profile sizes of 15 to 50 and n-gram lengths of 2 to 4. Standard error ≤ 0.015.

6.3.4 The Effect of Noise

As shown in Tables 6.7–6.9, the addition of 750 noise Web pages to the Syracuse corpus resulted in a slight decrease in the precision of the centroid classification model, and a slight increase in the recall. This means that the noise Web pages are less likely to be mislabeled as belonging to another genre than are the non-noise Web pages.

Of the 1985 non-noise Web pages, an average of 170 pages (8.6%) were erroneously labeled as unclassifiable noise Web pages by the centroid model; of the 750 noise Web pages, an average of 3 pages (0.4%) were erroneously labeled as non-noise pages. The number of Web pages erroneously labeled as noise increases from 3.6% to 13.3% as the n-gram length was increased from 2 to 4, whereas the number of noise pages erroneously given genre labels decreases from 0.85% to 0.03% as the n-gram length was increased from 2 to 4. This suggests that the proportion of noise Web pages expected to appear in a corpus could influence the choice of the n-gram length to be used for the Web page profiles.

6.3.5 The Syracuse Corpus and Noise: Conclusions

This section presented a study testing the centroid model on the highly unbalanced Syracuse corpus, with and without the presence of noise. Tables 6.7–6.9 indicate that the centroid classification model achieves excellent classification performance on this unbalanced corpus, even in the presence of noise Web pages. This result gives further support to the use of the optimal threshold method for determining thresholds for each genre; the method works well not only when Web pages may belong to more than one genre, such as with the 20-Genre corpus, but also when the Web pages belong to only one genre, or to no recognized genre, as with the Syracuse corpus. This study concludes the experiments with the centroid classification model. Section 6.3.5 summarizes the studies discussed in this chapter, and the conclusions that have been drawn from this work.

6.4 Conclusions

The goal of this thesis was to develop an approach to the problem of Web page genre classification that is effective not only on balanced, single-label corpora, but also on unbalanced and multi-label corpora, which better represent a real world environment. Chapter 5 investigated the use of the centroid classification model on balanced corpora, and showed that the model gives state-of-the-art classification performance on both of the corpora that were tested. The studies discussed in this chapter have expanded the research focus to include more challenging corpora than the balanced corpora that were used in Chapter 5. The major contribution of these studies is to show that an n-gram approach to Web page representation for genre classification with the new centroid model gives high mean classification performance under a variety of conditions, on more than one type of corpus.

The experiments discussed in Section 6.1 examined the classification performance of the centroid model on the extremely unbalanced 15-Genre corpus. The results of these experiments indicated that with the centroid classification model, using n-gram representations of Web pages and the χ^2 statistic as the feature selection measure achieved effective classification performance on the unbalanced corpus. The recall values were somewhat lower than the precision values, however in applications such as online searching, precision is typically considered to be more important than recall.

Section 6.2 described the investigation of techniques for setting genre thresholds in order to allow a Web page to belong to more than one genre, or to no genre at all, and a comparison of the classification performance of the centroid model with that of the popular support vector machine approach was made. Based on the results of this study, the conclusion was drawn that the n-gram representation of Web pages allows the effective classification of those Web pages by genre, even when the Web page belongs to more than one genre. The combination of the centroid classification model and the optimal threshold method was more successful than the SVM method in classifying this multi-label corpus, using a less complex and more easily scalable approach. The results of these experiments also showed that in general, as the length of the n-grams used to represent the Web pages was increased, the classification performance for each model decreased, suggesting that a short n-gram length should be used in future work. The experimental results also indicated that over the range of 15 to 50, the number of n-grams used to represent each Web page has only a slight impact on the classification results, therefore this range was determined to be sufficient for use in future work.

The study discussed in Section 6.3 looked at the effect of noise Web pages on the classification performance of the centroid model on the highly unbalanced Syracuse corpus. The results of this study indicated that the centroid classification model achieves excellent classification performance on this corpus, even in the presence of noise Web pages. This result gives further support to the use of the optimal threshold method for determining thresholds for each genre; the method works well not only when Web pages may belong to more than one genre, such as with the 20-Genre corpus, but also when the Web pages belong to only one genre, or to no recognized genre, as with the Syracuse corpus.

The studies discussed in this chapter have shown that the proposed centroid model for Web page genre classification is effective on highly unbalanced corpora, even in the presence of noise, and that the optimal threshold method can be used successfully to determine a threshold for each genre that allows a Web page to be assigned to more than one genre, or to no genre at all.

The next and final chapter, Chapter 7, gives a summary of the research contributions of this thesis, discusses the limitations of the research, and suggests directions for future work.

Chapter 7

Conclusion

The goal of this thesis was to develop an approach to the problem of Web page genre classification that is effective not only on balanced, single-label corpora, but also on unbalanced and multi-label corpora, which better represent a real world environment. This chapter provides a brief summary of the research and results that have been presented, discusses the main contributions of this research, and describes potential future work.

7.1 Summary

Chapter 4 discussed an investigation of the feasibility of using n-gram representations of Web pages for the task of Web page genre classification. The results of these experiments indicated that an n-gram approach to Web page representation for genre classification was feasible, and that the proposed centroid model, which gave classification results in the same range as those of other researchers on the 7-Genre corpus, warranted further exploration.

The n-gram representation of the Web pages achieved good performance on both the k-NN and centroid classification models, suggesting that the method is a robust way of representing Web pages for classification. The results of this study also revealed that increasing the Web page profile size beyond a base number of 500 did not appreciably affect the performance of either the k-NN or centroid models, however these experiments showed that increasing the n-gram length beyond a base size of 2 did have an impact on the precision and recall for each model. Although the centroid model achieved high classification performance with word, character, and byte n-gram representations of Web pages, the use of byte n-grams gave the best overall performance. The results of this work also indicated that the Kešelj distance measure given in Equation 4.3 gave the best performance, when compared with several other distance measures, in determining the similarity between n-gram profiles.

Chapter 5 discussed studies which furthered the investigation of the best set of parameters for the proposed approach, including n-gram length and Web page profile size. The effect of data preprocessing was also explored, as was the use of the χ^2 statistic and Information Gain as feature selection measures. The experiments in these studies broadened the research scope to include both the 7-Genre and KI-04 corpora.

The study discussed in Section 5.1 investigated the effects of common data preprocessing steps when using an n-gram approach to Web page representation with the centroid classification model. The results of this study indicated that for both the 7-Genre and KI-04 corpora, the classification performance of the centroid model increased with each level of preprocessing: doing no preprocessing gave poorer results than did removing the JavaScript code from the Web pages, and removing both the HTML tags and the JavaScript code further improved the classification performance of the model. These results also indicated that smaller Web page profiles can be advantageous when the feature set has been reduced by preprocessing, which led to the question of whether reducing the feature set by other means could also be effective.

Section 5.2 described the exploration of measures for feature selection, comparing the classification results for the centroid model when n-gram frequency, Information Gain, and the χ^2 statistic were used as feature selection measures. Although all three of the measures performed well, the χ^2 statistic significantly outperformed the other measures on both the 7-Genre and KI-04 corpora.

The study discussed in Section 5.3 revisited the question of data preprocessing, based on the results of the feature selection experiments. In this study, the χ^2 statistic was used as the feature selection measure, allowing the use of smaller Web page profiles than were used in the data preprocessing study discussed in Section 5.1. In contrast to the results in Section 5.1 in which n-gram frequency was used as a feature selection measure, this study showed that for both the 7-Genre and KI-04 corpora, the centroid model achieved significantly worse classification performance when both the HTML tags and JavaScript code were removed from the Web pages, while there was no significant difference between removing only the JavaScript code from the Web pages and doing no preprocessing at all. This is an indication that using the χ^2 statistic as a feature selection measure allows the selection of n-grams from the HTML tags which are helpful in determining the genre of the Web page. All of the results in this study were superior to the results of the data preprocessing study discussed in Section 5.1. Based on these results, preprocessing Web pages to remove HTML tags and/or JavaScript code was determined to be unnecessary when features are selected using the χ^2 statistic.

Chapter 6 presented experiments which broadened the scope of the work to include more challenging corpora. This included an investigation of techniques for setting genre thresholds in order to allow a Web page to belong to more than one genre, or to no genre at all. A comparison of the classification performance of the centroid model with that of the SVM approach was also made, and experiments conducted on highly unbalanced corpora, both with and without the inclusion of noise Web pages, were presented.

The study discussed in Section 6.1 examined the classification performance of the centroid model on the extremely unbalanced 15-Genre corpus. The results of this study indicated that with the centroid classification model, using n-gram representations of Web pages and the χ^2 statistic as a feature selection measure achieved effective classification performance on the unbalanced corpus.

Section 6.2 described the investigation of techniques for setting genre thresholds in order to allow a Web page to belong to more than one genre, or to no genre at all. The classification performance of the centroid model was compared with that of the SVM approach; the combination of the centroid classification model and the optimal threshold method was more successful than the SVM method in classifying this multi-label corpus. The results of these experiments also showed that in general, as the length of the n-grams used to represent the Web pages was increased, the classification performance for each model decreased, suggesting that a short n-gram length of 2 or 3 should be used. The experimental results also indicated that over the range of 15 to 50, the number of n-grams used to represent each Web page has only a slight impact on the classification results, therefore this range was determined to be sufficient for use in future work.

The study discussed in Section 6.3 looked at the effect of noise Web pages on the classification performance of the centroid model on the highly unbalanced Syracuse corpus. The results of this study indicated that the centroid classification model achieved excellent classification performance on this corpus, even in the presence of noise Web pages.

The research described in these chapters demonstrated that the proposed n-gram Web page representation and centroid classification model provide an approach to the problem of Web page genre classification that is effective not only on balanced, single-label corpora, but also on highly unbalanced and multi-label corpora, even in the presence of noise Web pages.

7.2 Research Contributions

The hypothesis of this thesis is that a byte n-gram representation of a Web page can be used effectively to classify the Web page by its genre(s). The goal of this thesis was to develop an approach to the problem of Web page genre classification that is effective not only on balanced, single-label corpora, but also on unbalanced and multi-label corpora, which better represent a real world environment. This thesis research developed n-gram representations for Web pages and Web page genres, and based on these representations, a new approach to the classification of Web pages by genre was developed.

Specifically, the investigation of whether there is a simple, easily scalable method for Web page classification was addressed by the following research questions, in the context of a set of controlled experiments.

Web Page and Web Page Genre Representation

The broad question of how Web pages and Web page genres should be represented was answered in the context of the following specific research questions.

1. **Do Web pages of the same genre share a distribution of n-grams that is similar enough to allow n-gram representations of the Web pages to be used for Web page genre classification?**

In the studies discussed in Chapters 4–6, the use of n-gram representations of Web pages achieves good classification performance with three different classification models, allowing the conclusion that Web pages of the same genre do in fact share a distribution of n-grams that is similar enough to allow n-gram representations of the Web pages to be used for Web page genre classification.

2. **Does there exist a set of parameters for the proposed approach, with regard to n-gram type, length, and number, that will allow state-of-the-art classification of Web pages by their genres?**

The studies discussed in Chapters 4–6 incrementally refined the centroid classification model and the parameters in order to achieve state-of-the-art classification performance; comparisons with the results of other researchers are given in Tables 4.6 and 6.6.

3. **Can a threshold value be determined for each genre in order to allow a Web page to belong to more than one genre, or to no genre at all?**

The study discussed in Section 6.2 investigated techniques for setting genre thresholds in order to allow a Web page to belong to more than one genre, or to no genre at all. The combination of the centroid classification model with the optimal threshold method gave the best classification results on the multi-label 20-Genre corpus, and this combination, as discussed in Section 6.3, also achieved excellent classification performance on the Syracuse corpus, even in the presence of noise Web pages.

Classification Method

The broad question of what classification method should be used was addressed by the following specific research questions.

1. **Can an appropriate distance measure be found that computes the similarity between Web page and Web page genre profiles, leading to a classification method based on the nearest profile?**

The results of the study discussed in Section 4.6 indicate that using the Kešelj distance measure, given in Equation 4.3, to compute the similarity between Web page profiles and Web page genre profiles gave the best performance when compared with several other distance measures. The use of this distance measure gave very successful results in subsequent work with the centroid classification model, performing Web page classification based on the nearest genre profile.

2. **Does the proposed classification model give comparable performance to the SVM approach, using the same n-gram representations of the Web pages?**

The study discussed in Section 6.2 compared the classification performance of the centroid model with that of the SVM approach, using the same n-gram representations of the Web pages. The combination of the centroid classification model with the optimal threshold method was more successful than the SVM method in classifying the multi-label 20-Genre corpus.

3. **Is the proposed approach effective on highly unbalanced corpora?**

The studies discussed in Chapter 6 presented experiments testing the centroid classification model on the highly unbalanced 15-Genre and Syracuse corpora, as well as the 20-Genre corpus, which is both unbalanced and multi-label. The results of these studies indicated that the centroid model achieved effective classification performance on each corpus.

4. **Given noise Web pages that do not belong to any genre in a particular corpus, is the proposed approach effective when these noise pages are included in the corpus?**

The study discussed in Section 6.3 looked at the effect of noise Web pages on the classification performance of the centroid model on the highly unbalanced Syracuse corpus. The results of this study indicate that the centroid classification model achieves excellent classification performance on this corpus, even in the presence of noise Web pages. The addition of 750 noise Web pages to the Syracuse corpus resulted in a slight decrease in the precision of the centroid model, and a slight increase in the recall, indicating that the noise Web pages added to the Syracuse corpus were less likely to be mislabeled than were the non-noise Web pages.

7.2.1 Theoretical Contributions

The research reported in this thesis developed an n-gram based representation of Web pages and a centroid based classification model for the task of classifying Web pages by genre. Although the use of n-grams is common in many research areas, their use had not been extensively explored in the context of Web page classification. The studies described in this thesis have shown that it is an effective representation for Web page classification, based on the corpora on which these experiments were run. The studies show that the use of short n-grams and small Web page profiles is sufficient to represent Web pages for classification; this insight can be of benefit to other researchers in any of the many areas of study in which n-gram representations are utilized.

The exploration of feature selection measures provides insight into the benefit of using a theoretically based measure to select features, and demonstrates the ability to produce compact yet high quality Web page representations. Although these experiments focused on Web page classification, they provide a guideline that may be applicable to many other classification tasks.

The development of the centroid classification model for Web page genre classification provides a classification model that is easily scalable; adding another genre to the model requires only the creation of a centroid from the training data for the new genre. Such a scalable model is very desirable in information retrieval systems. The studies in this thesis, having demonstrated the effectiveness of this simple model under a variety of conditions, provide a foundation on which other researchers can build.

The refinement of the centroid classification model with the development of a method for computing a threshold for each Web page genre is also of value to other researchers, and demonstrates the adaptability of this simple model for challenging environments.

7.2.2 Applied Contributions

This thesis proposed a practical n-gram representation of Web pages and a centroid classification model that could be applied in a real information retrieval system, either on the server-side or on the client-side. The results of the studies discussed in this thesis provide a guide to creating a compact Web page representation that allows high classification performance with a variety of classification models and Web page corpora. Of particular interest to real world applications is the investigation of data preprocessing, which demonstrated that when an appropriate feature selection measure is used, no preprocessing of the Web pages is necessary. These results have the potential to be directly applied in information retrieval systems.

7.2.3 Methodological Contributions

The methodological approach used to conduct the series of studies discussed in this thesis provides a methodological contribution to the research community. The use of an n-gram representation of Web pages for use in Web page genre classification was proposed, and a prototype of a centroid classification model using this representation was developed, evaluated, and refined. This model can be used and further modified by other researchers for use in Web page or document classification work. Descriptions of the classification model and experiments are available to the research community [78, 79, 80, 81, 82].

7.3 Limitations

Although this thesis has addressed some of the challenges of a real world environment, the lack of a large benchmark Web page genre corpus on which to test the n-gram based classification approach is a limitation of this work. The corpora that have been used in this research are very small and structured in comparison with a real world environment such as the World Wide Web. As well, all of the Web pages used in these experiments have been HTML pages. As Santini et al. [109] point out, the omission of other file formats, such as Word documents, PDF and Postscript files, and Web pages containing Flash objects, skews the diversity of the genres in the existing Web page corpora. These limitations make it impossible to generalize the results of this research to the World Wide Web as a corpus, therefore the focus has been on an incremental refinement of the classification model, with comparisons to the results of other researchers on the same corpora wherever possible.

In order to conduct this research in a timely manner, choices had to be made as to what approaches would be investigated. For example, only three feature selection methods were investigated, and only three classification models were tested with the n-gram Web page representation. Although a range of byte, character, and word

n-gram Web page representations were explored, in each case the n-grams were of fixed length; the use of variable-length n-grams was left for future work.

7.4 Future Work

All of the experiments discussed in this thesis were conducted using profiles containing fixed-length n-grams to represent Web pages. It is possible that incorporating variable-length n-grams in these Web page profiles could improve the performance of the classification model, and this would be an interesting extension to the current research.

Future research could also explore the use of Naïve Bayes classifiers. Although the SVM classifier has been shown to outperform the Naïve Bayes classifier [55, 102, 110, 141], more recent approaches to Bayesian classification, such as Multinomial Naïve Bayes and Weight-normalized Complement Naïve Bayes [94] can give performance comparable to that of the SVM method.

Although the study discussed in Section 6.2 compared the classification performance of the centroid model with that of the SVM approach, using the same n-gram representations of the Web pages, no optimization was performed for the SVM classifier. The use of n-gram representations of Web pages for the task of Web page genre classification could be further explored, using the SVM approach. For example, as discussed by Valentini and Masulli [133], the use of error correcting output coding (ECOC) to combine the results of multiple classifiers, such as those used in the *one-against-all* approach of the SVM experiments in Section 6.2, could improve the classification performance.

The use of ECOC methods to improve classification performance is not limited to use on with the SVM approach. For example, Berger [16] suggests that this approach is appropriate for both the k-NN and Naïve Bayes classifiers, while Ghani [47] demonstrated that the approach can dramatically reduce the text classification error using Naïve Bayes classifiers. Future work could explore the use of ECOC methods for the task of Web page genre classification on a variety of classification models.

Future work could also include a study of ensemble classification, in which a number of classifiers based on different Web page representations and/or classification models are combined [38]. This approach to Web page genre classification has been successfully explored by, for example, Kanaris and Stamatatos [57, 57], and Jebari and Ounalli [53, 54].

All of the Web pages in the corpora used in this research were English language HTML Web pages, using a relatively small number of Web page genres. Expanding these studies to include corpora of other languages, file formats, and genres would help generalize the results of the current research.

Although the n-gram Web page representation and centroid classifier perform extremely well on the Web page genre corpora used in this research, the model does not directly address the issues of the evolution of existing Web page genres and the emergence of new genres. These issues are analogous to those of the appearance and mutation of computer viruses or email spam. Future research could investigate a variety of methods for recognizing and measuring Web page genre drift, and could explore techniques for detecting the emergence of new Web page genres.

Bibliography

[1] 20-Genre Web Page Genre Collection Download Site.
http://dis.ijs.si/mitjal/genre/.

[2] 7-Genre Web Page Collection and KI-O4 Web Page Collection Download Site.
http://www.itri.brighton.ac.uk/~Marina.Santini/#Download.

[3] Predictive Analytics SoftWare (PASW) Statistics Information Site.
http://www.spss.com/statistics/.

[4] Stopword List. http://www.lextek.com/manuals/onix/stopwords2.html.

[5] Text::Ngrams Perl Module Download Site.
http://users.cs.dal.ca/~vlado/srcperl/Ngrams/.

[6] WebKB Collection Download Site.
http://www.cs.cmu.edu/afs/cs.cmu.edu/project/theo-20/www/data/.

[7] WEGA Genre Firefox Add-on Synopsis and Download Site.
http://www.uni-weimar.de/cms/index.php?id=8793.

[8] L.A. Adamic and B.A. Huberman. Zipf's Law and the Internet. *Glottometrics*, 3(1):143–150, 2002.

[9] R. Albert and A.L. Barabási. Statistical Mechanics of Complex Networks. *Reviews of Modern Physics*, 74(1):47–97, 2002.

[10] E. Alpaydin. *Introduction to Machine Learning.* MIT press, 2004.

[11] M.F. Amasyali and B. Diri. Automatic Turkish Text Categorization in Terms of Author, Genre and Gender. In *Proceedings of the 11th International Conference on Applications of Natural Language to Information Systems (NLDB 2006)*, volume 3999 of *Lecture Notes in Computer Science*, pages 221–226. Springer-Verlag, 2006.

[12] A.P. Asirvatham and K.K. Ravi. Web Page Classification based on Document Structure. *IEEE National Convention*, 2001.

[13] R. Aslıyan, K. Gunel, and T. Yakhno. Detecting Misspelled Words in Turkish Text Using Syllable n-gram Frequencies. 2007.

[14] R.A. Baeza-Yates and B. Ribeiro-Neto. *Modern Information Retrieval.* Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1999.

[15] R.E. Bellman. *Adaptive Control Processes: A Guided Tour.* Princeton University Press, 1961.

[16] A. Berger. Error-correcting Output Coding for Text Classification. In *Proceedings of IJCAI-99 Workshop on Machine Learning for Information Filtering*, 1999.

[17] E.S. Boese and A.E. Howe. Effects of Web Document Evolution on Genre Classification. In *Proceedings of the 14th ACM International Conference on Information and Knowledge Management (CIKM '05)*, pages 632–639, New York, NY, USA, 2005. ACM Press.

[18] K.S. Bordens and B.B. Abbott. *Research Design and Methods: A Process Approach*. McGraw-Hill, sixth edition, 2005.

[19] P. Braslavski. Combining Relevance and Genre-Related Rankings: an Exploratory Study. In G. Rehm and M. Santini, editors, *Proceedings of the International Workshop Towards Genre-Enabled Search Engines: The Impact of Natural Language Processing*, pages 1–4, September 2007.

[20] I. Bretan, J. Dewe, A. Hallberg, N. Wolkert, and J. Karlgren. Web-Specific Genre Visualization. In *WebNet'98*, 1998.

[21] W.B. Cavnar. Using an n-gram-based Document Representation with a Vector Processing Retrieval Model. In *Overview of the Third Text REtrieval Conference (TREC-3)*, pages 269–278, 1994.

[22] W.B. Cavnar and J.M. Trenkle. N-gram-based text categorization. In *Proceedings of the 3rd Annual Symposium on Document Analysis and Information Retrieval, SDAIR-94*, 1994.

[23] C. Chekuri, M.H. Goldwasser, P. Raghavan, and E. Upfal. Web Search Using Automatic Classification. In *Proceedings of the 6th International World Wide Web Conference (WWW1997)*, 1997.

[24] G. Chen and B. Choi. Web Page Genre Classification. In *Proceedings of the 2008 ACM Symposium on Applied Computing (SAC '08)*, pages 2353–2357, New York, NY, USA, 2008. ACM.

[25] B.Y.M. Cheng, J.G. Carbonell, and J. Klein-Seetharaman. A Machine Text-inspired Machine Learning Approach for Identification of Transmembrane Helix Boundaries. *Lecture Notes in Computer Science*, pages 29–37.

[26] B.Y.M. Cheng, J.G. Carbonell, and J. Klein-Seetharaman. Protein Classification Based on Text Document Classification Techniques. *Proteins: Structure, Function and Bioinformatics*, 2004.

[27] M. Clark and S. Watt. Classifying XML Documents by Using Genre Features. In *Proceedings of the 18th International Conference on Database and Expert Systems Applications (DEXA '07)*, pages 242–248, 2007.

[28] G. Cleuziou and C. Poudat. On the Impact of Lexical and Linguistic Features in Genre-and Domain-Based Categorization. *Lecture Notes in Computer Science*, 4394, 2007.

[29] C. Cortes and V. Vapnik. Support Vector Networks. *Machine Learning*, 20:273–297, 1995.

[30] M. Craven, D. DiPasquo, D. Freitag, A. McCallum, T. Mitchell, K. Nigam, and S. Slattery. Learning to Construct Knowledge Bases from the World Wide Web. *Artificial Intelligence*, 118(1-2):69–113, 2000.

[31] K. Crowston and B.H. Kwasnik. Can Document-genre Metadata Improve Information Access to Large Digital Collections? *Library Trends*, 52(2):345–361, 2003.

[32] K. Crowston and B.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 (HICSS-37)*. IEEE Computer Society, 2004.

[33] K. Crowston and M. Williams. Reproduced and Emergent Genres of Communication on the World-Wide Web. In *Proceedings of the 30th Hawaii International Conference on System Sciences (HICSS-30)*, pages 30–39. IEEE Computer Society, 1997.

[34] K. Crowston and M. Williams. Reproduced and Emergent Genres of Communication on the World Wide Web. *The Information Society*, 16(3):201–215, 2000.

[35] J.F. da Silva, G. Dias, S. Guilloré, and J.G.P. Lopes. Using LocalMaxs Algorithm for the Extraction of Contiguous and Non-contiguous Multiword Lexical Units. In *Proceedings of the 9th Portuguese Conference on Artificial Intelligence: Progress in Artificial Intelligence*, pages 113–132. Springer, 1999.

[36] J.F. da Silva and J.G.P. Lopes. A Local Maxima Method and a Fair Dispersion Normalization for Extracting Multiword Units. In *Proceedings of the 6th Meeting on the Mathematics of Language*, pages 369–381, 1999.

[37] N. Dewdney, C. VanEss-Dykema, and R. MacMillan. The Form is the Substance: Classification of Genres in Text. In *Proceedings of the Workshop on Human Language Technology and Knowledge Management*, pages 1–8, Morristown, NJ, USA, 2001. Association for Computational Linguistics.

[38] T.G. Dietterich. Ensemble Methods in Machine Learning. *Lecture Notes in Computer Science*, pages 1–15, 2000.

[39] G. Doddington. Automatic Evaluation of Machine Translation Quality Using n-gram Co-occurrence Statistics. In *Proceedings of the 2nd International Conference on Human Language Technology Research*, pages 138–145. Morgan Kaufmann Publishers Inc., 2002.

[40] L. Dong, C. Watters, J. Duffy, and M. Shepherd. Binary Cybergenre Classification Using Theoretic Feature Measures. In *International Conference on Web Intelligence (WI 2006)*, pages 313–316, 2006.

[41] L. Dong, C. Watters, J. Duffy, and M. Shepherd. An Examination of Genre Attributes for Web Page Classification. In *Proceedings of the 41st Hawaii International Conference on System Sciences (HICSS-41)*. IEEE Computer Society, 2008.

[42] L.B. Eriksen and C. Ihlström. Evolution of the Web News Genre - The Slow Move Beyond the Print Metaphor. In *Proceedings of the 33rd Hawaii International Conference on System Sciences (HICSS-33)*. IEEE Computer Society, 2000.

[43] A. Finn and N. Kushmerick. Learning to Classify Documents According to Genre. *Journal of American Society for Information Science and Technology*, 57(11):1506–1518, 2006.

[44] A. Finn, N. Kushmerick, B. Smyth, F. Crestani, M. Girolami, and C.J. van Rijsbergen. Genre Classification and Domain Transfer for Information Filtering. In *Proceedings of ECIR-02, 24th European Colloquium on Information Retrieval Research*. Springer Verlag, Heidelberg, DE, 2002.

[45] G. Forman. An Extensive Empirical Study of Feature Selection Metrics for Text Classification. *The Journal of Machine Learning Research*, 3:1289–1305, 2003.

[46] M. Ganapathiraju, D. Weisser, R. Rosenfeld, J. Carbonell, R. Reddy, and J. Klein-Seetharaman. Comparative n-gram Analysis of Whole-genome Protein Sequences. *Proceedings of the 2nd International Conference on Human Language Technology Research*, pages 76–81, 2002.

[47] R. Ghani. Using Error-Correcting Codes for Text Classification. In *Proceedings of the 17th International Conference on Machine Learning (ICML 2000)*, 2000.

[48] M. Grcar, B. Fortuna, D. Mladenic, and M. Grobelnik. kNN versus SVM in the Collaborative Filtering Framework. In *Proceedings of the Workshop on Knowledge Discovery in the Web WebKDD*, pages 21–24. Springer, 2005.

[49] S. Gupta, G. Kaiser, S. Stolfo, and H. Becker. Genre Classification of Websites Using Search Engine Snippets. Technical Report CUCS-004-05, Columbia University, Department of Computer Science, 2005.

[50] J. Houvardas and E. Stamatatos. N-gram Feature Selection for Authorship Identification. In *Proceedings of the 12th International Conference on Artificial Intelligence: Methodology, Systems, Applications*, pages 77–86. Springer, 2006.

[51] M.N. Hwang. Protein Sequence Search Based on n-gram Indexing. *Bioinformatics*, 1(6):53–57, 2006.

[52] C. Jebari. Combining Classifiers for Flexible Genre Categorization of Web Pages. In G. Rehm and M. Santini, editors, *Proceedings of the International Workshop Towards Genre-Enabled Search Engines: The Impact of Natural Language Processing*, pages 5–12, September 2007.

[53] C. Jebari and H. Ounalli. A New Approach for Flexible Document Categorization. In *Proceedings of the World Academy of Science, Engineering and Technology*, volume 20, pages 32–35, April 2007.

[54] C. Jebari and H. Ounalli. Genre Categorization of Web Pages. In *Proceedings of the Seventh IEEE International Conference on Data Mining Workshops (ICDMW 2007)*, pages 455–464, Washington, DC, USA, 2007. IEEE Computer Society.

[55] T. Joachims, C. Nedellec, and C. Rouveirol. Text Categorization with Support Vector Machines: Learning with Many Relevant Features. In *Proceedings of the 10th European Conference on Machine Learning (ECML-98)*, pages 137–142. Springer, 1998.

[56] I. Kanaris and E. Stamatatos. Webpage Genre Identification Using Variable-Length Character n-Grams. In *Proceedings of the 19th IEEE International Conference on Tools with Artificial Intelligence (ICTAI 2007)*, pages 3–10, 2007.

[57] I. Kanaris and E. Stamatatos. Learning to Recognize Webpage Genres. *Information Processing & Management*, 45(5):499–512, 2009.

[58] J. Karlgren and D. Cutting. Recognizing Text Genres with Simple Metrics Using Discriminant Analysis. In *Proceedings of the 15th Conference on Computational Linguistics*, pages 1071–1075, Morristown, NJ, USA, 1994. Association for Computational Linguistics.

[59] A. Kennedy and M. Shepherd. Automatic Identification of Home Pages on the Web. In *Proceedings of the 38th Hawaii International Conference on System Sciences (HICSS-38)*. IEEE Computer Society, 2005.

[60] B. Kessler, G. Numberg, and H. Schütze. Automatic Detection of Text Genre. In *Proceedings of the 35th Annual Meeting on Association for Computational Linguistics*, pages 32–38, Morristown, NJ, USA, 1997. Association for Computational Linguistics.

[61] V. Kešelj, E. Milios, A. Tuttle, S. Wang, and R. Zhang. DalTREC 2005 Spam Track: Spam Filtering using n-gram-based Techniques. In *Proceedings of the 14th Text REtrieval Conference (TREC 2005)*, 2005.

[62] V. Kešelj, F. Peng, N. Cercone, and T. Thomas. N-gram-based author profiles for authorship attribution. In *Proceedings of the Conference Pacific Association for Computational Linguistics, PACLING'03*, pages 255–264, 2003.

[63] Y. Kim and S. Ross. Automating Metadata Extraction: Genre Classification. In *Proceedings of the UK e-Science All Hands Meeting 2006: Achievements, Challenges, and New Opportunities*, 2006.

[64] Y. Kim and S. Ross. Detecting Family Resemblance: Automated Genre Classification. *Data Science Journal*, 6:172–183, 2007.

[65] Y. Kim and S. Ross. "The Naming of Cats": Automated Genre Classification. *International Journal of Digital Curation*, 2(1):49–61, 2007.

[66] Y. Kim and S. Ross. Examining Variations of Prominent Features in Genre Classification. In *Proceedings of the 41st Hawaii International Conference on System Sciences (HICSS-41)*. IEEE Computer Society, 2008.

[67] B. Kwasnik and K. Crowston. Introduction to the Special Issue: Genres of Digital Documents. *Information Technology & People*, 18(2):76–88, 2005.

[68] B. Kwasnik, K. Crowston, J. Rubleske, and Y.-L. Chun. Building a Corpus of Genre-Tagged Webpages for an Information-Access Experiment. In *Proceedings of the Colloquium "Towards a Reference Corpus of Web Genres", held in conjunction with Corpus Linguistics 2007*, July 2007.

[69] Y.B. Lee and S.H. Myaeng. Text Genre Classification with Genre-revealing and Subject-revealing Features. In *Proceedings of the 25th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '02)*, pages 145–150, New York, NY, USA, 2002. ACM Press.

[70] Y.B. Lee and S.H. Myaeng. Automatic Identification of Text Genres and Their Roles in Subject-based Categorization. In *Proceedings of the 37th Hawaii International Conference on System Sciences (HICSS-37)*. IEEE Computer Society, 2004.

[71] R. Levering, M. Cutler, and L. Yu. Using Visual Features for Fine-Grained Genre Classification of Web Pages. In *Proceedings of the 41st Hawaii International Conference on System Sciences (HICSS-42)*. IEEE Computer Society, 2008.

[72] C.S. Lim, K.J. Lee, and G.C. Kim. Automatic Genre Detection of Web Documents. *Natural Language Processing-IJCNLP 2004: First International Joint Conference, Hainan Island, China, March 22-24, 2004, Revised Selected Papers*, 2005.

[73] C.S. Lim, K.J. Lee, and G.C. Kim. Multiple Sets of Features for Automatic Genre Classification of Web Documents. *Information Processing and Management*, 41(5):1263–1276, 2005.

[74] H. Liu and V. Kešelj. Combined Mining of Web Server Logs and Web Contents for Classifying User Navigation Patterns and Predicting Users' Future Requests. *Data & Knowledge Engineering*, 61(2):304–330, 2007.

[75] H. Lodhi, C. Saunders, J. Shawe-Taylor, N. Cristianini, and C. Watkins. Text Classification using String Kernels. *The Journal of Machine Learning Research*, 2:419–444, 2002.

[76] J. Marino, R. Banchs, J.M. Crego, A. de Gispert, P. Lambert, J.A. Fonollosa, and M. Ruiz. Bilingual N-gram Statistical Machine Translation. In *Proceedings of Machine Translation Summit X*, pages 275–282, 2005.

[77] J. Marino, R. Banchs, J.M. Crego, A. de Gispert, P. Lambert, J.A.R. Fonollosa, and M.R. Costa-jussa. N-gram-based Machine Translation. *Computational Linguistics*, 32(4):527–549, 2006.

[78] J.E. Mason. An *n*-gram Based Approach to the Classification of Web Pages by Genre. In *Proceedings of the Grace Hopper Celebration of Women in Computing, (GHC 2009)*, 2009.

[79] J.E. Mason, M. Shepherd, and J. Duffy. An N-gram Based Approach to Automatically Identifying Web Page Genre. In *Proceedings of the 42nd Hawaii International Conference on System Sciences (HICSS-42)*. IEEE Computer Society, 2009.

[80] J.E. Mason, M. Shepherd, and J. Duffy. Classifying Web Pages by Genre: A Distance Function Approach. In *Proceedings of the 5th International Conference on Web Information Systems and Technologies, (WEBIST 2009)*, 2009.

[81] J.E. Mason, M. Shepherd, and J. Duffy. Classifying Web Pages by Genre: An *n*-gram Based Approach. In *Proceedings of the International Conference on Web Intelligence, (WI'09)*. IEEE Computer Society, 2009.

[82] J.E. Mason, M. Shepherd, J. Duffy, V. Kešelj, and C. Watters. An *n*-gram Based Approach to Multi-labeled Web Page Genre Classification. In *Proceedings of the 43rd Hawaii International Conference on System Sciences (HICSS-43)*. IEEE Computer Society. To appear.

[83] A. Mehler, R. Gleim, and A. Wegner. Structural Uncertainty of Hypertext Types: An Empirical Study. In G. Rehm and M. Santini, editors, *Proceedings of the International Workshop Towards Genre-Enabled Search Engines: The Impact of Natural Language Processing*, pages 13–19, September 2007.

[84] S. Meyer zu Eissen and B. Stein. Genre Classification of Web Pages. In *Proceedings of the 27th German Conference on Artificial Intelligence (KI-2004)*. Springer, 2004.

[85] Y. Miao, V. Kešelj, and E. Milios. Document Clustering Using Character n-grams: A Comparative Evaluation with Term-based and Word-based Clustering. In *Proceedings of the 14th ACM International Conference on Information and Knowledge Management (CIKM'05)*, pages 357–358. ACM New York, NY, USA, November 2005.

[86] N.S. Mitić, G.M. Pavlović-Lažetić, and M.V. Beljanski. Could n-gram Analysis Contribute to Genomic Island Determination? *Journal of Biomedical Informatics*, 2008.

[87] M. Nelson and J.S. Downie. Informetric Analysis of a Music Database. *Scientometrics*, 54(2):243–255, 2002.

[88] W.J. Orlikowski and J. Yates. Genre Repetoire: The Structuring of Communicative Practices in Organizations. *Administrative Science Quarterly*, 39(4):541, 1994.

[89] K. Papineni, S. Roukos, T. Ward, and W.J. Zhu. BLEU: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 311–318. Association for Computational Linguistics, 2001.

[90] J. Platt. Sequential Minimal Optimization: A Fast Algorithm for Training Support Vector Machines. *Advances in Kernel Methods-Support Vector Learning*, 208, 1999.

[91] M.F. Porter. An Algorithm for Suffix Stripping. *Program: Electronic Library and Information Systems*, 40(3):211–218, 2006.

[92] G. Rehm. Towards Automatic Web Genre Identification. In *Proceedings of the 37th Hawaii International Conference on System Sciences (HICSS-37)*. IEEE Computer Society, 2002.

[93] G. Rehm, M. Santini, A. Mehler, P. Braslavski, R. Gleim, A. Stubbe, S. Symonenko, M. Tavosanis, and V. Vidulin. Towards a Reference Corpus of Web Genres for the Evaluation of Genre Identification Systems. In *Proceedings of the 6th Language Resources and Evaluation Conference (LREC 2008)*, May 2008.

[94] J.D. Rennie, L. Shih, J. Teevan, and D. Karger. Tackling the Poor Assumptions of Naïve Bayes Text Classifiers. In *Proceedings of International Conference on Machine Learning*, pages 616–623, 2003.

[95] G.F. Roberts. The Home Page as Genre: A Narrative Approach. In *Proceedings of the 31st Hawaii International Conference on System Sciences (HICSS-31)*, pages 78–86. IEEE Computer Society, 1998.

[96] M. Robnik-Šikonja and I. Kononenko. Theoretical and Empirical Analysis of ReliefF and RReliefF. *Machine Learning*, 53(1):23–69, 2003.

[97] A. Rosmarin. *The Power of Genre.* University of Minnesota Press, 1985.

[98] M. Rosso. *Using Genre to Improve Web Search.* PhD thesis, University of North Carolina, 2005.

[99] M. Rosso. User-based identification of Web genres. *Journal of the American Society for Information Science and Technology*, 59(7), 2008.

[100] D. Roussinov, K. Crowston, M. Nilan, B. Kwasnik, Jin Cai, and Xiaoyong Liu. Genre based navigation on the web. In *Proceedings of the 34th Hawaii International Conference on System Sciences*, 2001.

[101] M. Santini. Common Criteria for Genre Classification: Annotation and Granularity. In *Proceedings of the Workshop on Text-based Information Retrieval (TIR-06) held in conjunction with ECAI*, 2006.

[102] M. Santini. Identifying Genres of Web Pages. In *Proceedings of TALN*, 2006.

[103] M. Santini. Interpreting Genre Evolution on the Web: Preliminary Results. In *EACL 2006 Workshop: NEW TEXT-Wikis and Blogs and Other Dynamic Text Sources*, 2006.

[104] M. Santini. Some issues in Automatic Genre Classification of Web Pages. *JADT 06-Actes des 8 Journées internationales d'analyse statistiques des donnés textuelles*, 2006.

[105] M. Santini. Towards a Zero to Multi Genre Classification Scheme. *Journée ATALA Typologies de textes pour le traitement automatique*, 9, 2006.

[106] M. Santini. *Automatic Identification of Genre in Web Pages.* PhD thesis, University of Brighton, 2007.

[107] M. Santini. Characterizing Genres of Web Pages: Genre Hybridism and Individualization. In *Proceedings of the 40th Hawaii International Conference on System Sciences (HICSS-40).* IEEE Computer Society, 2007.

[108] M. Santini. Zero, Single, or Multi? Genre of Web Pages Through the Users' Perspective. *Information Processing and Management*, 44(2):702–737, 2008.

[109] M. Santini, A. Mehler, and S. Sharoff. Riding the Rough Waves of Genre on the Web: Concepts and Research Questions. In A. Mehler, S. Sharoff, and M. Santini, editors, *Genres on the Web: Computational Models and Empirical Studies*, pages 3–32. Submitted to Springer, Berlin/New York, 2009.

[110] M. Santini, R. Power, and R. Evans. Implementing a Characterization of Genre for Automatic Genre Identification of Web Pages. In *Proceedings of the COLING/ACL on Main conference poster sessions*, pages 699–706, Morristown, NJ, USA, 2006. Association for Computational Linguistics.

[111] M. Santini and M. Rosso. Testing a Genre-Enabled Application: A Preliminary Assessment. In *Proceedings of the 2nd BCS-IRSG Symposium on Future Directions in Information Access*, 2008.

[112] V. Shanks and H.E. Williams. Fast categorisation of Large Document Collections. *Proceedings of the 8th International Symposium on String Processing and Information Retrieval (SPIRE 2001*, pages 194–204, 2001.

[113] C.E. Shannon. A Mathematical Theory of Communication. *Bell System Technical Journal*, 27:379–423 and 623–656, July and October 1948.

[114] S. Sharoff. In the Garden and in the Jungle: Comparing Genres in the BNC and Internet. In *Proceedings of the Colloquium "Towards a Reference Corpus of Web Genres", held in conjunction with Corpus Linguistics 2007*, July 2007.

[115] D. Shen, Z. Chen, Q. Yang, H.J. Zeng, B. Zhang, Y. Lu, and W.Y. Ma. Webpage Classification Through Summarization. In *Proceedings of the 27th Annual International Conference on Research and Development in Information Retrieval (SIGIR '04)*, pages 242–249, New York, NY, USA, 2004. ACM.

[116] M. Shepherd and C. Watters. The Evolution of Cybergenres. In *Proceedings of the 31st Hawaii International Conference on System Sciences (HICSS-31)*, volume 2, pages 97–109. IEEE Computer Society, 1998.

[117] M. Shepherd and C. Watters. The Functionality Attribute of Cybergenres. In *Proceedings of the 32nd Hawaii International Conference on System Sciences (HICSS-32)*. IEEE Computer Society, 1999.

[118] M. Shepherd and C. Watters. Identifying Web Genre: Hitting A Moving Target. In *Proceedings of the 13th International World Wide Web Conference (WWW2004). Workshop on Measuring Web Search Effectiveness: The User Perspective*, 2004.

[119] M. Shepherd, C. Watters, and A. Kennedy. Cybergenre: Automatic Identification of Home Pages on the Web. *Journal of Web Engineering*, 3(3&4):236–251, 2004.

[120] M.R. Spiegel, J.J. Schiller, and R.A. Srinivasan. *Schaum's Outline of Theory and Problems of Probability and Statistics*. McGraw-Hill Professional, second edition, 2000.

[121] SPSS Inc., Chicago, IL. *SPSS Base 15.0 User's Guide*, 2006.

[122] E. Stamatatos. Ensemble-based Author Identification Using Character n-grams. In *Proceedings of the 3rd International Workshop on Text-based Information Retrieval*, pages 41–46, 2006.

[123] E. Stamatatos, N. Fakotakis, and G. Kokkinakis. Text Genre Detection Using Common Word Frequencies. In *Proceedings of the 18th Conference on Computational Linguistics*, pages 808–814, Morristown, NJ, USA, 2000. Association for Computational Linguistics.

[124] E. Stamatatos, N. Fakotakis, and G.K. Kokkinakis. Automatic Text Categorization in Terms of Genre, Author. *Computational Linguistics*, 26(4):471–495, 2000.

[125] J. Steffen. N-gram Language Modeling for Robust Multi-lingual Document Classification. In *The 4th International Conference on Language Resources and Evaluation (LREC2004)*, 2004.

[126] B. Stein and S. Meyer zu Eissen. Retrieval Models for Genre Classification. *Scandinavian Journal of Information Systems*, 20(1):93–119, 2008.

[127] A. Stubbe, C. Ringlstetter, and R. Goebel. Elements of a Learning Interface for Genre Qualified Search. In G. Rehm and M. Santini, editors, *Proceedings of the International Workshop Towards Genre-Enabled Search Engines: The Impact of Natural Language Processing*, pages 21–28, September 2007.

[128] A. Stubbe, C. Ringlstetter, and K.U. Schulz. Genre as Noise: Noise in Genre. *International Journal on Document Analysis and Recognition*, 10(3):199–209, 2007.

[129] I.S.H. Suyoto and A.L. Uitdenbogerd. Simple efficient n-gram indexing for effective melody retrieval. In *Proceedings of the First Annual Music Information Retrieval Evaluation eXchange*, September 2005.

[130] J. Swales. *Genre Analysis*. Cambridge University Press, New York, 1990.

[131] S. Symonenko. Recognizing Genre-Like Regularities in Website Content Structure. In G. Rehm and M. Santini, editors, *Proceedings of the International Workshop Towards Genre-Enabled Search Engines: The Impact of Natural Language Processing*, pages 29–36, September 2007.

[132] A. Tomović, P. Janičić, and V. Kešelj. N-Gram-based Classification and Unsupervised Hierarchical Clustering of Genome Sequences. *Computer Methods and Programs in Biomedicine*, 81(2):137–153, 2006.

[133] G. Valentini and F. Masulli. Ensembles of Learning Machines. *Lecture Notes in Computer Science*, pages 3–22, 2002.

[134] V. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag, New York, 1995.

[135] V. Vidulin, M. Luštrek, and M. Gams. Evaluation of Different Approaches to Training a Genre Classifier. In *Proceedings of the 29th International Conference on Artificial Intelligence and Pattern Recognition*, pages 515–520, 2007.

[136] V. Vidulin, M. Luštrek, and M. Gams. Training the Genre Classifier for Automatic Classification of Web Pages. In *Proceedings of the 29th International Conference on Information Technology Interfaces (ITI 2007)*, pages 93–98, June 2007.

[137] V. Vidulin, M. Luštrek, and M. Gams. Using Genres to Improve Search Engines. In G. Rehm and M. Santini, editors, *Proceedings of the International Workshop Towards Genre-Enabled Search Engines: The Impact of Natural Language Processing*, pages 29–36, September 2007.

[138] P. Wastholm, A. Kusma, and B.B. Megyesi. Using Linguistic Data for Genre Classification. In *Proceedings of the Swedish Artificial Intelligence and Learning Systems Event SAIS-SSLS*, 2005.

[139] Z. Wei, D. Miao, J. Chauchat, and C. Zhong. Feature Selection on Chinese Text Classification Using Character n-Grams. *Lecture Notes in Computer Science*, 2008.

[140] W. Wibowo and H.E. Williams. Simple and Accurate Feature Selection for Hierarchical Categorisation. In *Proceedings of the 2002 ACM Symposium on Document Engineering*, pages 111–118, New York, NY, USA, 2002. ACM Press.

[141] Y. Yang and X. Liu. A Re-Examination of Text Categorization Methods. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 42–49. ACM New York, NY, USA, 1999.

[142] Y. Yang and J.O. Pedersen. A Comparative Study on Feature Selection in Text Categorization. In *Proceedings of the 14th International Conference on Machine Learning (ICML '97)*, pages 412–420, San Francisco, CA, USA, 1997. Morgan Kaufmann Publishers Inc.

[143] J. Yates, W. Orlikowski, and J. Rennecker. Collaborative Genres for Collaboration: Genre Systems in Digital Media. In *Proceedings of the 30th Hawaii International Conference on System Sciences (HICSS-30)*. IEEE Computer Society, 1997.

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

[145] P.C.K. Yeung, L. Freund, and C.L.A. Clarke. X-Site: A Workplace Search Tool for Software Engineers. In *Proceedings of the 30th Annual International*

ACM SIGIR Conference on Research and Development in Information Retrieval. ACM New York, NY, USA, 2007.

[146] R. Zhang, M. Shepherd, J. Duffy, and C. Watters. Automatic Web Page Categorization using Principal Component Analysis. In *Proceedings of the 38th Hawaii International Conference on System Sciences (HICSS-40)*. IEEE Computer Society, 2007.

[147] S. Zhou and J. Guan. Chinese Document Classification Based on n-grams. In *Proceedings of the 3rd International Conference on Computational Linguistics and Intelligent Text Processing (CICLING-02)*, pages 405–414. Springer, 2002.

[148] G.K. Zipf. *Selected Studies of the Principle of Relative Frequency in Language.* Harvard University Press, Cambridge, MA, USA, 1932.

[149] G.K. Zipf. *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology.* Addison-Wesley Press, Cambridge, MA, USA, 1949.

Appendix A

As discussed in Section 3.3, in order to interpret the experimental results of the research conducted for this thesis, the data are subjected to statistical analysis. All of the experiments conducted for this thesis are run using k-fold cross-validation. For the 7-Genre and KI-04 corpora, 10-fold cross-validation is performed, while 3-fold cross-validation is performed with the 15-Genre, 20-Genre, and Syracuse corpora. With this method, the Web pages in a particular corpus are partitioned randomly into k groups of equal size and distribution. For each of the k cross-validation iterations, one of the partitions is used as the test (validation) set, and the other $k-1$ partitions make up the training set; over the k iterations, each of the k partitions is used exactly once as the test set. The k-fold cross-validation method allows the simulation of k experiments, thus increasing the strength of the statistical analysis. In general, the power of a statistical test increases as the number of observations increases [18].

The statistical tests conducted as part of this thesis were run using the Statistical Package for the Social Sciences (SPSS) software, which has since been re-branded as the Predictive Analytics SoftWare (PASW) [3].

This section gives a brief description of the steps taken in running the Analysis of Variance (ANOVA) and Scheffé post hoc testing, using version 15.0 of the SPSS software. The SPSS software provides a graphical environment with dropdown menus and dialog boxes. The basic steps in using SPSS, as given in the SPSS User's Guide [121], are as follows.

- enter the data in the SPSS data editor
- select a procedure from the menus
- select the variables for the analysis
- select the appropriate options for the analysis
- run the procedure and look at the results

Data Format

For the statistical tests conducted as part of this thesis, the data were formatted in tables. Tables A.1 and A.2 give examples of the table formats. Table A.1 illustrates the format for the analysis of more than one method; Table A.2 shows an example in which the statistical analysis is performed on a single method. Because cross-validation is performed, there would be more than one entry in each table for each combination of independent variables. For the 7-Genre and KI-04 corpora, 10-fold cross-validation is performed, while 3-fold cross-validation is performed with the 15-Genre, 20-Genre, and Syracuse corpora. Thus, in each table there would be results for each of these cross-validation runs. For example, when comparing three methods using 3-fold cross validation, using three n-gram lengths, eight Web page profile sizes, and twenty genres, the number of rows in Table A.1 would be $3 \times 3 \times 3 \times 8 \times 20 = 4320$. In the same case, if the statistical analysis were run on only one of the methods, as shown in Table A.2, the number of rows would be $3 \times 3 \times 8 \times 20 = 1440$. The goal is to have the number of observations large enough that the statistical analysis can detect differences in the data, but small enough that weak relationships with little theoretical or practical value do not achieve statistical significance [18].

Method ID	$n$gram Length	Web Page Profile Size	Genre	Precision	Recall	F1-measure
1	2	15	1	1.00000000	.88888889	.94117647
1	2	15	2	1.00000000	.22727273	.37037037
⋮	⋮	⋮	⋮	⋮	⋮	⋮
1	2	15	19	.00000000	.00000000	.00000000
1	2	15	20	1.00000000	.40789474	.57943926
⋮	⋮	⋮	⋮	⋮	⋮	⋮
3	4	50	1	1.00000000	.84210526	.91428571
3	4	50	2	1.00000000	.59090909	.74285714
⋮	⋮	⋮	⋮	⋮	⋮	⋮
3	4	50	19	1.00000000	.78666667	.88059702
3	4	50	20	1.00000000	.80263158	.89051095

Table A.1: Example: Data format for analyzing the results of more than one method.

$n$gram Length	Web Page Profile Size	Genre	Precision	Recall	F1-measure
2	15	1	1.00000000	1.00000000	1.00000000
2	15	2	1.00000000	.57142857	.72727273
⋮	⋮	⋮	⋮	⋮	⋮
2	15	19	1.00000000	.70270270	.82539682
2	15	20	1.00000000	.52702703	.69026549
⋮	⋮	⋮	⋮	⋮	⋮
4	50	1	1.00000000	.88888889	.94117647
4	50	2	1.00000000	.45000000	.62068966
⋮	⋮	⋮	⋮	⋮	⋮
4	50	19	1.00000000	.69863014	.82258065
4	50	20	.97500000	.54166667	.69642857

Table A.2: Example: Data format for analyzing the results of one method.

Opening a Data File

A variety of types of data files can be opened in the SPSS environment, including spreadsheets, database tables, and delimited text files. To open a data file in SPSS, from the menus, make the following choices.

```
File
   Read text data
      Data...
```

In the `Open Data` dialog box, select the file to be opened and click `Open`.

Selecting a Procedure

The statistical tests conducted for this thesis used the general linear model univariate analysis, which provides regression analysis and analysis of variance (ANOVA) for one dependent variable by one or more factors and/or variables [121]. As explained in Section 3.3.1, ANOVA is based on the concept of analyzing the variance that appears in groups of data by testing differences in the means for statistical significance [18]. The total observed variance is partitioned according to the factors assumed to be responsible for producing that variation in the data. This is accomplished by partitioning the total variance into the component that is due to true random error, and the components that are due to differences between means. These latter variance components, or sources of variance, are then tested for statistical significance. If the difference is significant, the null hypothesis of no differences between the means is rejected, and the alternative hypothesis, that the means are different from each other, is accepted. ANOVA also allows the detection of interaction effects between variables.

To select the general linear model univariate analysis in SPSS, as used for this thesis, make the following choices from the menus.

```
Analyze
  General Linear Model
    Univariate...
```

The `Univariate` dialog box will now be open.

Selecting Variables for Analysis

An independent variable, or fixed factor, is a variable whose values for a particular experiment are fixed by the experimenter [18]. In the experiments conducted for this thesis, the independent variables included *n*-gram length, Web page profile size, and genre. Dependent variables are variables whose values are observed and recorded as part of the experimental results [18]. In the experiments conducted for this thesis, the dependent variables included precision, recall, and the F1-measure.

Using the univariate general linear model, a separate analysis is performed for each dependent variable. To select the variables for the general linear model univariate analysis in SPSS, as used for this thesis, make the following choices in the `Univariate` dialog box.

```
  Select one dependent variable from the list on the left, and
  click the arrow to the left of the Dependent Variable textbox.
For each fixed factor,
  select one fixed factor from the list on the left, and
  click the arrow to the left of the Fixed Factor textbox.
The next step is to select the options for the analysis.
```

Selecting Options for the Analysis

To select the options for the general linear model univariate analysis in SPSS, as used for this thesis, make the following choices in the `Univariate` dialog box.

- click `Model...`
  - select `Full factorial`
  - select `Sum of squares:  Type III`
  - select `Include intercept in model`
  - click `Continue`
- click `Post Hoc...`
  - For each factor of interest under `Factor(s):`
    - select one fixed factor from the list on the left
    - click the arrow to the left of the `Post Hoc Tests for:` textbox
  - select `Scheffe`
  - click `Continue`
- click `Options...`
  - For each factor of interest under `Factor(s) and Factor Interactions:`
    - select one fixed factor from the list on the left
    - click the arrow to the left of the `Display Means for:` textbox
  - select `Descriptive statistics`
  - select `Estimates of effect size`
  - select `Observed power`
  - click `Continue`
- click `OK` to run the procedure

Looking at the Results

In SPSS, the results of a procedure are opened in a new window referred to as the viewer. The viewer is divided into two panes, with an outline of the contents on the left, and the statistical tables and output on the right.

As explained in Section 3.3.1, when a significant effect is reported, it is typically expressed in terms of a *p*-value. The *p*-value refers to the probability (ranging from zero to one) that the observed results (or results more extreme) could have occurred by chance, if in reality the null hypothesis is true. A small p-value leads to the conclusion that the means are different; a p-value of less than 0.05 is considered to indicate statistical significance, and is reported in the form $p < 0.05$. When the p-value indicates statistical significance, the smaller this value is, the more confidence can be given to the corresponding interpretation of the results [18]. In the tables in SPSS, the p-value is found in the column labeled `Sig`.

The effect size is the degree to which changing an independent variable (such as n-gram length) affects the value of a dependent variable (such as classification accuracy) [18]. The partial Eta squared (η^2) indicates the proportion of the total variability that is attributable to a particular factor, or independent variable, when controlling for other factors. In the output from the SPSS procedure described here, this value is found in the `Tests of Between Subjects Effects` table, in the column labeled `Partial Eta Squared`.

As discussed in Section 3.3.2, a statistically significant effect in ANOVA is often succeeded with a follow-up, or post hoc test. This can be done in order to assess which groups are different from which other groups, or to test various other hypotheses. Post hoc tests such as Scheffé's test most commonly compare every group mean with every other group mean. For the statistical analysis in this thesis, Scheffé's test is used because it is considered to be very conservative, meaning that it is more difficult to achieve statistical significance with this test than with less conservative tests, such as the Tukey or Duncan post hoc tests [18]. In the output from the SPSS procedure described here, the various tables resulting from the Scheffé post hoc testing are found under the heading `Post Hoc Tests`.

Appendix B

7-Genre Corpus	Genre Centroid Size for Web Page Profile Size of		Number of Unique n-grams for Web Page Profile Size of	
n-gram Length	500	2500	500	2500
2	2113	2846	4017	5249
3	4644	16254	11881	41961
4	12231	40916	43433	136432
5	20564	72417	92774	297182
6	27360	103258	142553	498839

Table B.1: Centroid genre profile sizes and the total number of unique n-grams used, for Web page profile size of 500 and 2500, with the 7-Genre corpus.

7-Genre Corpus	Character n-grams No Stopword Removal			Character n-grams With Stopword Removal		
n-gram Length	Precision	Recall	F1	Precision	Recall	F1
2	0.855	0.825	0.840	0.848	0.831	0.839
3	0.870	0.832	0.850	0.859	0.812	0.835
4	0.929	0.924	0.927	0.908	0.901	0.904
5	0.932	0.929	0.930	0.930	0.926	0.928
6	0.934	0.931	0.932	0.931	0.926	0.928
Average	0.904	0.888	0.896	0.895	0.879	0.887

Table B.2: Classification results for character n-grams without stopword removal, and with stopword removal, using the centroid model on the 7-Genre corpus, with a Web page profile size of 500. Standard error ≤ 0.007.

7-Genre Corpus	Byte n-grams No Stopword Removal			Byte n-grams With Stopword Removal		
n-gram Length	Precision	Recall	F1	Precision	Recall	F1
2	0.867	0.832	0.849	0.850	0.817	0.833
3	0.917	0.908	0.912	0.924	0.916	0.920
4	0.930	0.929	0.929	0.938	0.935	0.936
5	0.937	0.933	0.935	0.933	0.929	0.931
6	0.934	0.939	0.932	0.934	0.930	0.932
Average	0.917	0.906	0.912	0.916	0.905	0.910

Table B.3: Classification results for byte n-grams without stopword removal, and with stopword removal, using the centroid model on the 7-Genre corpus, with a Web page profile size of 500. Standard error ≤ 0.007.