# Genre Classification of Web Documents

**Elizabeth Sugar Boese** and **Adele Howe**
Department of Computer Science
211 UNVSC Building 601 South Howes Street Colorado State University
{boese},{howe}@cs.colostate.edu
970/491-7016
http://www.cs.colostate.edu/~boese/Research

## Abstract

Retrieving relevant documents over the Web is an overwhelming task when search engines return thousands of Web documents. Sifting through these documents is time-consuming and sometimes leads to an unsuccessful search. One problem is that most search engines rely on matching a query to documents based solely on topical keywords. However, many users of search engines have a particular *genre* in mind for the desired documents. The *genre* of a document concerns aspects of the document such as the style or readability, presentation layout, and meta-content such as words in the title or the existence of graphs or photos. By including genre in Web searches, we hypothesize that Web document retrieval could greatly improve accuracy by better matching documents to the user's information needs.

Before implementing a search engine capable of discriminating on both genre and topic, a feasibility analysis of genre classification is needed. Our previous research achieved 91% classification accuracy across ten genres, whereas similar research range between 60 and 85% accuracy. However, the ten genres used in our research were mostly distinct and only *exemplar* Web documents (consisting of only one genre) were chosen. This paper discusses our current work which involves an in-depth analysis of maintaining high accuracy rates among genres that are very similar.

## Background

Genres are differentiated by features of a document. Common features include: words that appear in the document, text statistics (e.g., number of words per sentence) and HTML analysis (e.g., number of tables). Filtering methods are usually applied to narrow down the number of features used for classification. Ideally, these methods choose features that are highly correlated with one genre and independent of the other genres. Filtering increases the speed of classification and can also improve the accuracy of results.

The number of genres analyzed in most research studies is very small, usually less than ten genres. This begs the question whether current genre classification can scale to handle more of the genres that exist on the Web. In our research, we found that classifying documents to highly differentiated genres resulted in high accuracy rates. However, when incorporating similar genres such as *tutorial* and *how-to*, the accuracy rates diminish.

Experiments on genre classification have mostly relied on a flat structure. However, Crowston and Kwasnik (Crowston & Kwasnik 2004) have proposed that genres should be organized into a hierarchical scheme based from general to more specific facets [features], but they have no current experiments on this topic. Rehm presented a genre hierarchy, without any empirical analysis, for his academia corpus, but most of the genres are mixed with topics (e.g., list of personal homepages, list of contact information, list of publications) (Rehm 2002).

## Results of Our Study

Experiments were performed on a corpus of 343 Web documents distributed across ten genres. Documents were selected and manually classified to genres by myself and one or two others, who were either fellow classmates or the two professors for the course. Documents were pruned from the corpus when there were disagreements on the classification of the document. Document features were filtered using a best-first forward search strategy and greedy hill-climbing with backtracking capability. This narrowed the number of features down from over 1,600 to 78 features, which resulted in similar accuracy in less time. Classification was performed using LogitBoost. Table 1 shows the results of experiments, depicting an average correctness of 91.5%.

| Genre | Total | Hits | % Correct | FP | FN |
|---|---|---|---|---|---|
| **Abstract** | 25 | 22 | 88.0% | 2 | 3 |
| **Call for Papers** | 23 | 23 | 100.0% | 1 | 0 |
| **FAQ** | 34 | 31 | 91.2% | 3 | 3 |
| **How-To** | 26 | 21 | 80.8% | 2 | 5 |
| **Hub** | 26 | 22 | 84.6% | 7 | 4 |
| **Job Description** | 34 | 31 | 91.2% | 3 | 3 |
| **C.V./Resume** | 41 | 38 | 92.7% | 0 | 3 |
| **Statistics** | 53 | 44 | 83.0% | 12 | 9 |
| **Syllabus** | 48 | 44 | 91.7% | 2 | 4 |
| **Tech Paper** | 33 | 33 | 100.0% | 1 | 0 |

Table 1: Data set used in experiments: a total of 343 Web documents were used across 10 genres. (Hits = correctly classified, FP = false-positives, FN = false-negatives)

Although the results are very good, there are some points that must be noted. First, only *exemplar* Web documents were chosen for each genre. A document was considered to be *exemplar* if it contained only one genre. This was necessary to avoid the complexity of multi-classification confounding the results. Second, the genres are mostly disparate, which becomes apparent when classifying similar genres. We performed an experiment combining *how-to* and *tutorial* documents into the same genre and classified across the ten genres. This resulted in about 63% accuracy, a degradation of 28% from that shown in Table 1. We hypothesize that combining these two genres led to a multitude of inconsistent features, which led to the degradation of accuracy.

Similar genres can be identified from a confusion matrix. For our data, Figure 1 shows that six *statistics* Web documents were mis-classified as *hub* genre, and four *hub* documents as *statistics*. Further analysis of the corpus revealed that many of the documents classified as *statistics* contained a set of links, especially when each data item in the table of statistics was a link for more information.

```
  a  b  c  d  e  f  g  h  i  j  <- classified
 21  2  0  0  0  1  0  0  1  0 | a=Abstract
  1 22  0  0  0  0  0  0  0  0 | b=Cfp*
  0  0 31  3  0  0  0  0  0  0 | c=FAQ
  0  1  1 20  0  1  0  2  1  0 | d=How-to
  0  0  0  1 22  0  0  3  0  0 | e=Hub
  0  0  0  2  0 32  0  0  0  0 | f=Job desc
  0  0  0  0  0  0 41  0  0  0 | g=Resume/C.V.
  0  0  1  0  3  0  1 47  1  0 | h=Statistics
  0  0  0  0  0  0  1  2 45  0 | i=Syllabus
  0  0  0  0  0  0  0  0  0 33 | j=Tech paper
 22 25 33 26 25 34 43 54 48 33 | Totals
```

Figure 1: Confusion matrix analyzing the 78 features selected through feature selection. LogitBoost correctly classified 314 out of 343 instances (91.5%) with stratified 10-fold cross-validation. *(Cfp = call for papers)

## Current Directions

The feasibility of genre classification of Web documents has been substantiated by many others (Crowston & Kwasnik 2004; Rehm 2002; Rogati & Yang 2002). These studies have highlighted issues in genre classification. The foremost is the lack of a standard set of Web genres. Researchers have postulated any number of Web genres, ranging from seven to over a hundred. This leads to a lack of pre-classified data sets to utilize for research experiments. Most researchers built their own set by surfing the Internet or from search engine results. Although useful for feasibility studies, scaling the results to practical use requires analyzing more genres and specifically, similar genres.

Given that we can collect a good corpus, our research is now directed towards distinguishing similar genres. Preliminary results on *how-to* and *tutorial* genres showed that these genres needed to be separate to achieve high accuracy rates in classification. We will also address the issue of multi-genre documents, such as *statistics* documents that also resemble the *hub* genre. We will compare the accuracy rates of the different genres, recognizing that some genres (e.g., *resumes*) will be easier to classify due to the relatively strict format and other genres (e.g., *tutorial*) will have a lower maximum classification accuracy. In addition, we want to find other genres that may be combined (e.g. *homepages* : company, organization, faculty, student and personal).

Our goal is to determine the level of granularity and specificity required for genre classification of Web documents by creating a hierarchy of genres. A partial hierarchy scheme, as depicted in Table 2, could integrate into applications such as search engines and directory structures. A search engine could organize results in a genre outline. Directory structures such as Yahoo! and Google could use the genre hierarchy as part of the directory structure, so that surfers desiring a particular genre could more easily access relevant documents. This research leads in to our long-term goal of developing a search engine that integrates genre classification to improve usability.

| **Macro** | **Micro *(sub-micro)*** |
|---|---|
| Form | interactive<br>non-interactive |
| Help | faq/q&a<br>how-to<br>tutorial<br>definition<br>lecture notes / presentation |
| Homepage | personal<br>academic *(faculty, student)*<br>corporate<br>organizational |
| List | links/hub/sitemap<br>downloads<br>other non-link list *(publications, citations, table of contents, inventory)*<br>statistics |

Table 2: Partial hierarchy of Web document genres divided into three levels: macro, micro, and sub-micro genres. Sub-micro genres are in parentheses next to the Micro-level.

## References

Crowston, K., and Kwasnik, B. 2004. A framework for creating a facetted classification for genres: Addressing issues of multidimensionality. In *Proceedings of the 37th Annual Hawaii International Conference on System Sciences.* Oahu, HI: IEEE Computer Society.

Rehm, G. 2002. Towards automatic web genre identification. In *Proceedings of the Hawaiian International Conference on System Sciences.* Oahu, HI: IEEE Computer Press.

Rogati, M., and Yang, Y. 2002. High-peforming feature selection for text classification. In *Proceedings of ACM Conference on Information and Knowledge Management (CIKM) '02.* McLean, Virginia: ACM Press.