Towards Practical Genre Classification of Web Documents

George Ferizis
CSIRO ICT Centre, Sydney NSW, Australia
george.ferizis@csiro.au

Peter Bailey
CSIRO ICT Centre, ANU Canberra, Australia
peter.bailey@csiro.au

ABSTRACT

Classification of documents by genre is typically done either using linguistic analysis or term frequency based techniques. The former provides better classification accuracy than the latter but at the cost of two orders of magnitude more computation time. While term frequency analysis requires much less computational resources than linguistic analysis, it returns poor classification accuracy when the genres are not sufficiently distinct. A method that removes or approximates the expensive portions of linguistic analysis is presented. The accuracy and computation time of this method is then compared with both linguistic analysis and term frequency analysis. The results in this paper show that this method can significantly reduce the computation of both time of linguistic analysis and term frequency analysis, while retaining an accuracy that is higher than that of term frequency analysis.

Categories and Subject Descriptors

I.2.7 [**Artificial Intelligence**]: Natural Language Processing—*Text Analysis*; I.5.2 [**Pattern Recognition**]: Design Methodology—*Classifier design and evaluation*
; H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing

General Terms

Design, Experimentation, Performance

Keywords

Genre classification, term frequency, linguistic

1. INTRODUCTION

Queries submitted to search engines rarely contain information about the desired document genre. An example query for the term *Robert Ludlum*, returns documents that range from biographies and interviews to online shops. Guided navigation, through the use of genre classification, may increase the relevancy of search results as it provides to the user the ability to associate a genre with the search terms that they provided.

Genre classification groups a set of documents into smaller sets according to some predefined genre classes. Genre classification differs from text classification as it discriminates between the style of the documents as opposed to the latter which discriminates between the topic of the documents. Text classification techniques typically use the frequency of terms in the documents to discriminate between documents of different topics. Intuitively, documents on the same or similar topics will contain certain terms in common more frequently than documents on other topics. Similar techniques have been tried for classifying genre with varying degrees of success.

Previous classification techniques have either used term frequency analysis [3, 4] or a more robust linguistic approach that involves POS(Part of Speech) tagging [2]. While it has been observed in previous literature [4, 5] that these linguistic approaches do provide better accuracy than term frequency approaches, it has also been observed by Kessler et al [3] that the POS tagging requires significant computational time. Our experiments as shown in Table 1 show that POS tagging using the Brill POS tagger [1] contributes to97% of the total time spent classifying a document.

Table 1: Timing analysis of Karlgren algorithm

Stage	Time Spent(s)	% of Total Time
POS tagging	450	97.2
Extraction and analysis of variables	13	2.8
Applying classifier	0.1	0.02

In this paper a technique is presented that is based on the Karlgren and Cutting [2] algorithm. It does away with POS tagging by approximating some POS features that are critical to the accuracy of the classifier. It is shown that this method returns more accurate results than term frequency based techniques, and for a small sacrifice in accuracy provides two orders of magnitude greater speed performance than POS based linguistic analysis. It is worth noting that no claim is made about the innovative nature of this approach in this paper. It has been selected to demonstrate the computational overhead POS tagging introduces to linguistic classification techniques for little accuracy benefit, while demonstrating that even approximations to POS gives superior results to term frequency classification techniques.

2. ALGORITHM AND RESULTS

The C4.5 decision trees produced by the Karlgren algorithm indicate that the most important linguistic features for genre classification that are determined from POS are

WWW 2006, May 23–26, 2006, Edinburgh, Scotland.
ACM 1-59593-332-9/06/0005.

present participle frequency, *adverb frequency* and *noun frequency*. In the algorithm presented in this paper *noun frequency* is ignored, while *present participle frequency* and *adverb frequency* are approximated with heuristics.

The *present participle frequency* was approximated by selecting all words with a length greater than 5 characters and ending with the suffix **-ing**. *Adverb frequency* was approximated by selecting all words of length greater than 4 and ending with the string **-ly**, as well as using a list of the 50 most common adverbs to appear in a training corpus as determined by the Brill POS tagger.

Genre classification experiments were run over a random sample of documents from the genres: *editorial*, *reportage*, *scientific* and *speeches*. This test was used to determine if the reduced linguistic approach had the same difficulty as the term frequency based approach when classifying documents between genres which are not distinct. Both the reports and editorial were from the same source to ensure that the stylistic differences between various sources could not be used to discriminate between the genres. The algorithm in this paper was compared to Karlgren and Cuttings algorithm and a simple term frequency based algorithm that used the 500 most common words in the training corpus as features.

All experiments were run on a dual CPU 2 GIIz AMD Opteron 64 system running Linux with 8 GB of memory. In all experiments only one processor was used, so the results obtained should be comparable to results that would be obtained on a single CPU system.

2.1 Runtime

Table 2 shows the number of documents that each method could classify per second. It shows that removing the POS variables from the linguistic techniques significantly improves the number of documents that can be classified each second. It also shows that the reduced linguistic approach does have a better runtime than the term frequency approach, however this may be caused by the implementation of the algorithm as the runtimes are fairly similar.

Table 2: Average classification times

Algorithm	Documents/second
POS based linguistic	1.6
Term frequency	145
Reduced linguistic	430
Reduced linguistic with adverbs	238

2.2 Classifying arbitrary genres

The graph in figure 1 shows the results for all three approaches with varying training set sizes. The results show that the approximation of POS features only results in a small reduction in accuracy.

A comparison of the C4.5 decision trees produced by the reduced linguistic approach and the POS approach shows that they are fairly similar. Any differences in accuracy can be attributed to the inaccuracy of the *adverb* approximating, or from not considering the *noun* linguistic feature.

3. CONCLUSION

This paper has shown that with a small sacrifice in accuracy it is possible to classify documents without using computationally expensive POS tagging. The experiments also show that the reduced linguistic approach classifies more accurately than term frequency based approaches.

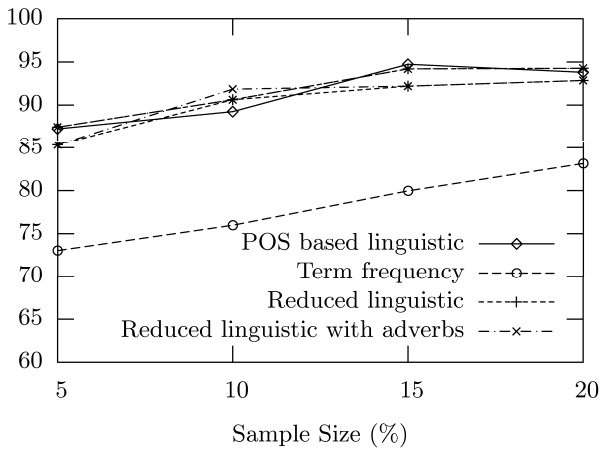

Figure 1: Accuracy of arbitrary genre classification

The experiments also show that it may be possible to produce a genre classification algorithm that has is accurate and efficient enough to be applied to large collections of documents. This has many potential applications both on the Web at large and within enterprises.

Future work may look at increasing the accuracy of classification by using techniques to rapidly detect nouns in a document, and techniques to improve the detection of adverbs, at the cost of some computation time. Further work is also planned to measure how hypertext information can improve the classification accuracy for web collections

4. REFERENCES

[1] E. Brill. A simple rule-based part-of-speech tagger. In *Proceedings of ANLP-92, 3rd Conference on Applied Natural Language Processing*, pages 152–155, Trento, IT, 1992.

[2] 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.

[3] B. Kessler, G. Nunberg, and H. Schütze. Automatic detection of text genre. In P. R. Cohen and W. Wahlster, editors, *Proceedings of the Thirty-Fifth Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, pages 32–38, Somerset, New Jersey, 1997. Association for Computational Linguistics.

[4] E. Stamatatos, G. Kokkinakis, and N. Fakotakis. Automatic text categorization in terms of genre and author. *Comput. Linguist.*, 26(4):471–495, 2000.

[5] M. Wolters and M. Kirsten. Exploring the use of linguistic features in domain and genre classification. In *Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics*, pages 142–149, Morristown, NJ, USA, 1999. Association for Computational Linguistics.