Classification along Genre Dimensions

Exploring a Multidisciplinary Problem

Mikael Gunnarsson
This thesis treats the sociotechnical notion of genre as a conflation of a communicative situation and a community of practices involved in producing and using documents. It explores the ways in which documents may be mapped to the sociocultural contexts from which they emanate. In other words, it is concerned with the classification of documents along genre dimensions, with the purpose of supporting information seeking.

The thesis positions itself within Library and Information Science in two parts. Firstly, a theoretical framework for classification along genre dimensions is developed based on relevant theories and practices from Library and Information Science, as well as from sociologically motivated Linguistics, and neighbouring domains. Secondly, a setup for experiments, including feature derivation and reannotation of existing corpora, is designed in order to explore the relationship between text documents and genres, and the extent to which a mapping of documents to genres can be realized in real world applications.

The experimental part of the thesis relies on an existing corpus for genre classification research, used in comparable research, with an addition of a slight extension. In the experiments, combinations of feature sets and target genres are evaluated, using traditional performance estimators for classification performance.

The outcome of the first part of the work indicates that the notion of genre with respect to classification is largely undertheorized in Library and Information Science. We need to know more about the nature of different genres, how to robustly identify the documents of a genre, and the impact genres have on information seeking. Interdisciplinary collaborative research would be most beneficial in these efforts. The results of the experiments of the second part are fairly inconclusive for the evaluation of feature sets, but it can be concluded that the optimal combination of feature sets and target genres is a crucial issue for high performance, and worthy of more investigation.

**Keywords:** Genre, Library and information science, Document studies, Classification, Knowledge organization, Library classification, Text linguistics, Sociolinguistics, Speech act theory, Machine learning, Support vector machines, k-NN classification, K-means clustering