

Elements of a Learning Interface for Genre Qualified Search

Andrea Stubbe¹, Christoph Ringlstetter², and Randy Goebel²

¹ CIS, University of Munich

² AICML, University of Alberta

Abstract. Even prior to content, the genre of a web document leads to a first coarse binary classification of the recall space in relevant and non-relevant documents. Thinking of a genre search engine, massive data will be available via explicit or implicit user feedback. These data can be used to improve and to customize the underlying classifiers. A taxonomy of user behaviors is applied to model different scenarios of information gain. Elements of such a learning interface, as for example the implications of the *lingering time* and the *snippet genre recognition factor*, are discussed.

1 Introduction

Given a web user's information need, even prior to content, the genre of a web page leads to a first coarse binary classification of the recall space in immediately rejected documents and such that require further processing. A search engine that provides automatic classification into genres, as for example, *shopping portals*, *scientific papers* or *personal web pages*, could make a big difference in regards to the number of documents that have to be checked for relevancy.

If such an interface is available in public, a steady stream of user events will arise. These behavioral observations have to be turned into meaningful data either to improve the performance of the initial classifiers or to adapt to genre shift. All attempts to acquire such data from a running system have to consider the user's level of explicitness and cooperativeness. Discussed in detail are the aspects of a silent genre interface where the user's statements on the genre of a document are only provided implicitly. In that connection, two qualities play a decisive role: first, knowledge about the implications of the *lingering time*, the time a user spends with a certain web page, will help to improve the precision of the genre classifiers; second, the *snippet genre recognition factor*, the percentage of documents whose genre a user can identify by only referring to the snippet, influences possible improvements of recall. To investigate the adaptability of genre classifiers for different scenarios, we simulate user feedback on genre labeled result sets by annotated corpus data.

2 Genre Qualified Search

To enable genre qualified search we need a classification schema, features, classifiers, and a search interface.

Document genres. The technical term “genre” refers to the partition of documents into distinct classes of texts with similar function and form. In [1] we introduced a genre hierarchy consisting of 8 container 32 leaf classes that tries to meet the demands of genre focused search. With regards to the main purpose of this study, the adaption of classifiers by user data, we exemplified results by five genres, three rather distinct ones: *blog*, *catalog*, *faq*, and the two journalistic genres *news* and *interview*.

Features. Feature selection was organized on training corpora comprising 20 prototype documents for each genre. Together over 200 different features were considered to organize the 32 leaf genres, including form, vocabulary, and parts of speech, complex patterns, and combinations of all these [1]. Examples of features are content-to-code-ratio, average line length, number of names, positive adjectives, dates, or bibliographic references. An example of a high level structure is a *casual style of writing* that can be recognized by the number of contractions (e.g. “won’t”) and the use of vague, informal, and generalizing words.

Specialized classifiers. To reach minimal description we introduced specialized features for each single genre classifier [1]. The features were arranged into a conjunction of single rules, applying a human supervised selection process that prevents overfitting by statistical coincidence on small training samples.¹ In Table 1 we show a cross-section of the specialized classifier of the genre *news*.

Table 1. Cross-section of the rule based classifier for the *news* genre

textlength in characters, forms

$length > 200 \wedge length < 6500 \wedge headlines < 3 \wedge sent > 1$

part of speech

$verbs \geq 5 \wedge adjectives < 20 \wedge adjectivesPositivNegativ < 0.4$

spoken/written text

$(contractions < 0.4 \vee dirSpeech > 0) \wedge contractions/dirSpeech < 0.2$

tense

$verbsPastTense < 0.18 \wedge verbsPastTense > verbsPresentTense \wedge$

$verbsIngForms > verbsPresentTense$

Search interface. To enable genre qualified search, an explicit interface has to provide functionalities for genre input and user feedback. Since “most users are reluctant to additional work”[2], a more realistic variant is a *silent interface* that minimizes the cognitive load of the user. Desired genres have then to be deduced from the query combined with locally or globally aggregated knowledge about the user. The feedback of the user is derived from his observable navigation on the result set.

3 Adaption of the Specialized Genre Classifiers

To make our static classifiers adaptive we had to rewrite them in disjunctive normal form (DNF). Alternative rules are combined by a logical *OR*; within the disjunctive

¹ For research purposes the corpora, features, and classifiers are available at <http://www.cis.uni-muenchen.de/~andrea/genre/>.