Incremental genre classification

Andrea Stubbe[1], Christoph Ringlstetter[2,] Tong Zheng[2] and Randy Goebel[2]

[1] CIS, University of Munich (Germany)

[2] AICML, University of Alberta (Canada)

Short Abstract

Most approaches for the partition of document spaces into different genres rely on static training corpora. However, thinking of applications, for example within search engines, static classifiers disregard potentially valuable data available via explicit or implicit user feedback. We provide an initial scenario of incremental genre classification. A taxonomy of user behaviors is applied to develop possible strategies for classifier adaption driven by user feedback simulated using annotated corpus data.

Long Abstract

Genre as a selective dimension of an increasingly less concise document space is receiving more and more attention. An obvious application of genre classification is the refinement of document search. During the public employment of a genre interface, a steady stream of user events will arise. If these behavioral observations can be turned into meaningful data, they can be exploited to adapt the start configuration of the underlying classifiers.

1. Search interface

To give the user the possibility to restrict his document search to certain genres the usual search interface has to be adapted. A very simple explicit adaption is to augment the input window where the user enters his query by a *genretype attribute* analogous to the *filetype attribute* most of the current search engines provide. To enable an explicit feedback functionality, the result page has to be extended with a click box where the user can provide a statement on the genre of a presented webpage. Many variants of the sketched interface are conceivable with a completely *silent interface* as an extreme that is supposed to minimize the cognitive load of the user. Desired genres have then to be deduced from the gestalt of the query combined with locally or globally aggravated knowledge about the user. The feedback of the user with respect to the suggested genre labels has to be deduced from his observable navigation on the result set.

2. User behavior

To further analyze the proposed genre search interface we model four different scenarios of user behavior. We define a **query** as a non-empty set of keywords and a genre label. A **result**

set is a set of ranked documents retrieved by the search engine processing a certain query. Each result document is annotated with a Boolean value referring to the genre selected by the user. According to our interface we define two different kinds of clicks: a **retrieval click**, the selection of a certain document, and an **evaluation click**, a user statement on the genre label of the document. The user's readiness to cooperate on the evaluation of the presented genre labels can be divided into four levels.

1.) **Fully cooperative behavior.** The user provides an evaluation statement for all documents of the result set: each page of the result set turns into correctly labeled data.

2.) **Cooperative behavior.** The user provides an evaluation statement of the annotation labels for the retrieved web pages. Thus, each retrieval click leads to an evaluation click.

3.) **Semicooperative behavior.** The user provides an evaluation statement only for a certain percentage of the visited pages.

4.) **Uncooperative behavior.** The user provides no information. Evaluation statistics can only be derived implicitly from the visiting statistics of the pages themselves.

According to studies of standard search engines, the average number of visited pages per search session is less than 2 and in most cases these 2 pages are retrieved from the first 20 hits of the search results. Consistent with this, we set, on average, a visit of two pages per turn. If both, labeled and unlabeled pages are present, the user visits the labeled pages. If the turn contains not enough labeled pages the user is assumed to be able to derive the desired genre with a certain accuracy from the snippet (*snippet genre recognition factor*). The resulting events are summarized in Table 1. For semi-cooperative behavior all are possible. Cooperative behavior is inconsistent with (iii) and (v), uncooperative excludes (i), (ii) and (iv).

(i)	user visits labeled page and confirms label
(ii)	user visits labeled page and rejects label
(iii)	user visits labeled page without evaluation
(iii.a)	page was correct classified
(iii.b)	page was false classified
(iv)	user visits unlabeled page and sets label
(v)	user visits unlabeled page without setting a label
(v.a)	page was correct negative
(v.b)	page was false negative

Table 1. A taxonomy of feedback events

3. Adaption of the genre classifiers

In previous work we have introduced specialized rule based classifiers that rely on aggressively pruned handcrafted feature sets. A necessary prerequisite to endow these static classifiers with the capability of adaptive response to new information is to rewrite them in disjunctive normal form (DNF). Generally, this implies each alternative rule combination to be linked to the other combinations by a logical **OR**. Within the disjunctive elements only connections by logical **AND** are allowed. Lower and upper bounds of the features' numerical ranges have to be explicit. Below we show a cut-out of the catalog-classifier in its DNF form.

(currency > 3.1 ∧ currency < 100,000 ∧ form > 0.1 ∧ form < 100,000 ∧ rel-curr. > 1.51 ∧ rel-curr. < 100,000)

∨

(currency > 5.1 ∧ currency < 100,000 ∧ form > 0 ∧ form < 100,000 ∧ rel-curr. > 5.1 ∧ rel-curr. < 19.9)

To establish comparability, for all features the adaptions of the ranges are normalized to values within the interval [0..1]. The general adaption algorithm to process available information on the genre of an input file, given the premise of a static feature space, has to distinguish between two different situations:

a.) *False negative:* A document of genre N_i has not been recognized as N_i. For every disjunctive element of the classifier in DNF form, we compute the sum of the required range adaptions to achieve a correct classification of the input document. The element with the minimum sum is selected and its ranges are temporarily adapted.

Constraint: Generally, the files in the *relevant history* that are classified correctly attendant on the classifier adaption (*new correct positives*) have to outnumber the files that are now falsely classified (*new false positives*).

b.) *False positive:* A document of genre N_j has been falsely recognized as genre N_i. We identify elements of the disjunction that have approved the input document as N_i. Within the elements, we look for the smallest sum of adaptions that prevent the positive classification of the document.

Constraint: Again, the number of files for the *relevant* history that are classified correctly attendant on the classifier adaption (*new correct negatives*) has to be larger than the number of files that are now falsely classified (*new false negatives*).

Uncooperative user behavior: the challenge with uncooperative user behavior is to investigate the degree to which we can derive knowledge from events that do not involve explicit user statements. In practice and in literature the *lingering time* is used to substitute explicit relevancy judgments of a user. If the user stays at a retrieved webpage for a time longer than a certain threshold τ, the page is assumed to be relevant. The probability of the correctness of this assumption, $P(\text{relevant}(x)|\text{time}(x) > \tau)$, is estimated using relative frequencies within controlled user data. The inference from relevancy to the users' evaluation of genre labeling introduces additional difficulties. Either a correctly labeled page can be irrelevant for the user or an incorrectly labeled page can be relevant. To the best of our knowledge, the probabilities that judgments on the labeling derived by document relevancy are correct, $P(\text{genre}(x) = \text{label}(x)|\text{relevant}(x))$, $P(\text{genre}(x) \neq \text{label}(x)|\neg\text{relevant}(x))$, have so far not been investigated.

4. Experiments

In a first series of experiments on the incremental adaption of three example classifiers, *blog, catalog*, and *faq*, we used the corpus provided by Marina Santini for the positive examples, each split into 160 documents for training and 40 documents for measuring recall. For the training/test with negative examples we used a controlled corpus of 31 different genres. From the training corpora we randomly generated 48 result sets consisting of 20 documents, each containing ~3 documents of the desired genre, as the basis for the simulation of user behavior.

4.1. Experiments for the fully cooperative user

The *fully cooperative user* provides the interface with complete information about the binary classification of the presented data. In Table 2 we present the results for the adaption of the rule based classifiers and of an SVM-classifier. For one genre, *faq,* the SVM did not converge. Summarized, a significant improvement of the classification can be achieved by using fully labeled data. However, a fully cooperative user can only be expected if he has a very high personal interest in the improvement of the classification. To reconcile to a realistic search environment, we have to gradually adapt this concept.

Genre	$Recall^{Train}$	$Fallout^{Train}$	$Recall^{Test}$	$Fallout^{Test}$	$Recall^{Test\text{-}SVM}$	$Fallout^{Test\text{-}SVM}$
Blog	70.00 (61.25)	1.80 (0.50)	72.50 (57.50)	1.85 (0.13)	72.50 (65.00)	2.14 (1.07)
Catalog	58.75 (40.00)	1.32 (0.59)	52.50 (40.00)	1.19 (0.27)	47.50 (42.50)	1.37 (0.31)
FAQ	90.50 (41.50)	3.36 (1.33)	77.50 (52.50)	4.29 (1.20)	-	-

Table 2: Fully cooperative user. Results for adapted classifiers and start configuration (in brackets).

4.2. Experiments for the cooperative user

A rational cooperative user will retrieve pages of the desired genre and will give feedback whether they were correctly classified. If not enough positively labeled pages are available, it can be assumed that the user will try to derive the missing label from the snippets, retrieve the pages, and give feedback on the genre. As is immediately clear, the assumed user behavior of only retrieving two documents leads to a strong preference of events that can help to improve precision. The phenomenon of classification improvement despite of the data loss can be described as a case of *active learning* in that only a few interesting examples are sufficient to successfully adapt the borders of a classifier.

Genre	$Recall^{Train}$	$Fallout^{Train}$	$Recall^{Test}$	$Fallout^{Test}$	$Recall^{Test\text{-}SVM}$	$Fallout^{Test\text{-}SVM}$
Blog	81.25 (61.25)	6.40 (0.50)	83.40 (57.50)	6.36 (0.13)	72.50 (65.00)	2.14 (1.07)
Catalog	60.00(40.00)	1.73 (0.59)	52.50 (40.00)	1.06 (0.27)	45.00 (42.50)	1.98 (0.31)
FAQ	85.00 (41.50)	1.33 (1.33)	75.00 (52.50)	1.91 (1.20)	-	-

Table 3: Cooperative user. Results for adapted classifiers and start configuration (in brackets).

4.3 Experiments for the uncooperative User

With uncooperative user behavior, the lingering time of the user on a retrieved result page, depending on genre, topic and model exogenous factors, is transformed into a binary relevancy signal. A negative signal means that the document is irrelevant either because of wrong topic or because of wrong genre. Unfortunately, in a realistic scenario the topic precision is poor which prevents us from gathering reliable data on genre by a negative relevancy signal. This leaves the case where the lingering time exceeds the threshold. To get a positive relevancy signal for cases where the document is of the desired genre the topic must be relevant. Insofar, we have to expect data loss for correct positives and false negatives with a factor of 1 – precision(topic(x)) and a small data gain via accidental confirmations by exogenous events. As for the documents of a genre different than that desired, we have false positives and correct negatives that can be amplified by a positive lingering signal caused by relevancy because of topic or by exogenous events. For our experiments we worked with deliberate probabilities of 0.1 for the lingering time caused by an exogenous event, 0.95 for a relevant document being of relevant topic *and* relevant genre, and a topic precision of 0.5. With these values we get a data loss of 45% for the correct positives and the false negatives and an defilement with 12% noise for the retrieved negatives. For the experiment with faq we received out of 48 result sets with 20 documents each, 0 feedback examples for false positives, 40 for correct positives, 6 for false negatives, 0 for correct negatives, 1 noisy example for correct positives and 7 noisy examples for false negatives. For both classifier types we get reduced but fairly robust improvements despite of the data loss and the defilement with noise.

Genre	$Recall^{Train}$	$Fallout^{Train}$	$Recall^{Test}$	$Fallout^{Test}$	$Recall^{Test\text{-}SVM}$	$Fallout^{Test\text{-}SVM}$
Blog	70.00 (61.25)	1.84 (0.50)	72.50 (57.50)	2.26 (0.13)	57.50 (65.00)	2.14 (1.07)
Catalog	56.25 (40.00)	1.22 (0.59)	52.50 (40.00)	0.97 (0.27)	45.00 (42.50)	0.92 (0.31)
FAQ	79.37 (41.50)	1.33 (1.33)	67.50 (52.50)	1.91 (1.20)	-	-

Table 4: Uncooperative user. Results for adapted classifiers and start configuration (in brackets).

5 .Conclusion

We have introduced an initial framework for the steady improvement of a genre search interface exploiting data of observed user events. Our next goals are to extend the simulation to more genres by collecting additional genre corpora and then to implement a prototype of a genre interface to collect real data for the estimation of now assumed values for the correlation between *lingering time* and correct genre and the *genre snippet recognition factor*. Finally, we want to extend the classifier adaption with respect to a dynamic feature space.