Elements of a Learning Interface for Genre Qualified Search

Andrea Stubbe[1], Christoph Ringlstetter[2], Randy Goebel[2]
[1] CIS, University of Munich, Oettingenstr 67, D-80538 München, Germany
E-mail: stubbe@cis.uni-muenchen.de
[2] AICML, University of Alberta, Edmonton, Canada T6G 2E8
E-mail: kristof,goebel@cs.ualberta.ca

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. This 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. Current search engines leave this filtering procedure entirely to the user. However, the engineering of next generation retrieval systems has to pay more attention to genre as a selective dimension of an increasingly less concise document space [4, 5]. Automatic classification of document sets, for example into *shopping portals*, *scientific papers* or *personal web pages*, can make a big difference in regards to the number of documents that have to be checked for relevancy and by that significantly reduce the user's cognitive load. With the rising commercialization of the web, abounding for example with “spam shops” that dominate the recall of more casual search interests, the partition of the result set into genre is the only way to deliver access to the relevant documents in numbers above mere coincidence. Thus, a next generation search engine interface must allow the user to qualify her keyword based search by one or more web genres that efficiently constrain the space of potentially relevant documents.

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 to adapt the initial configuration of the underlying classifiers: either to improve the performance of the initial classifiers or to adapt to genre shift. The classification process has to be tuned by permanent learning. All attempts to aquire such data from a running system have to consider the user's level of explicitness and cooperativeness. We formulate different scenarios for information gain representing different degrees of uncertainty. 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 by observing the user retrieving pages not classified as belonging to the initially selected genre.

To investigate the adaptability of different genre classifiers, we will simulate the user feedback on genre labeled result sets using annotated corpus data. Our intention is to give an overview of the challenges of dynamic classifier adaption based on data of different quality. We try to provide an idea about the amount of noise and incompleteness that is tolerable for a successful update functionality.

In Section 3, as a starting point, we describe a hierarchical classification schema of document genres. Section 4 addresses our approach for genre classification. In Section 5 we describe a possible search engine interface that provides features for genre classification. In Section 6 we introduce a taxonomy of user behaviors together with their consequences for gathering information. Section 7 provides strategies for incremental classifier adaption. First experiments on the snippet recognition factor are given in Section 9. In Section 10 we describe the experimental results on classifier adaption. The conclusion comments on future directions of the research dedicated to an improved interface for document search.

2 Related work

Boese and Howe [3] state that users often have a certain genre in mind when conducting a search task. In user studies, Meyer zu Eissen/Stein [15] and Rosso [19, 20] both received overwhelmingly positive feedback (nearly 100%) on the question whether labeling texts according to their genre would be useful in determining the relevance of a document. However, these results so far have not been empirically verified. Rosso presented Google snippets (”surrogates”) with and without genre labels to a group of users. He found no significant difference in the agreement between relevance judgment of labeled and unlabeled snippets and the document as well as in the time users needed to rate the snippets. Joho and Jose [12] provide a scenario to investigate the effects of enriched search result presentations (thumbnails and summaries) for relevance assessment and query reformulation. Except for Rosso's experiment the findings imply that the search interface would be improved by adding further information such as genre labels.

Implicit relevancy feedback is a research topic that with the rising impact of commercial search engines attracted a lot of attention. For a bibliography see [14]. The users' preference of explicit and implicit relevance feedback in dependency of the task complexity and the users' retrieval experience was investigated by White et al. [26, 27], continuing the work of Bell and Ruthven [2]. A first study on the reliabilty of implicit relevance feedback was contributed by Joachims et al. [11].

Although genre classification is still a rater new and specialized field of research, already several authors have presented genre palettes and automatic classifiers. For a discussion see, for example, [15, 22]. With regards to the construction of the genre pallete, the majority of authors follows a top-down approach, often inspired by users studies. An exception is the bottom-up experiment of Nilan et al. [16] that, however, not yet has lead to a stable schema.

3 Document genres

In [25] we introduced a hierarchy of genres that tries to meet the demands of genre focused partition of document spaces.[1] This hierarchy is used as a starting point to model an interface for genre qualified search. The hierarchy, consisting of 8 container and 32 leaf classes, is presented in Table 1. The containers of the hierarchy define a first classification level usable for coarse partition of the search space. The leaf classes provide finer granularity, allowing a highly focused search of web documents. With regards to classification errors, this hierarchical classification schema helps to keep misclassifications within logically acceptable layers. From a user perspective, a misclassification of a commentary into another journalistic genre is by far not as embarrassing as, for example, a misclassification of a shop portal as a scientific article.

Even though we are aware that the concept of genre sometimes applies to parts of a document instead of the whole [18], we determine genre on the level of a complete web page because, so far, the page is the basic unit for search tasks. To meet the challenge of mixed documents, we allow the classification of one document into multiple classes.

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 (journalistic, private,...), catalog (e-commerce shops, ...), faq (service pages, hobby related)* and two belonging to the same container, the journalistic genres *news* and *interview*.

4 Static genre classification

As we argued above, genre classification helps to recognize unwanted documents and thus partitions the document space into relevant and non-relevant documents. A kernel issue underlying document classification is the selection of features.

4.1 Features

Many kinds of features were considered to organize the 32 leaf genres, including HTML, form, vocabulary, parts of speech, complex patterns, and combinations of all these. 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. When we put all genre specific features together, the result is a global feature set with 200 different features.

A.Journalism	**C. Information**	D.3 protocol
A.1 commentary	C.1 science report	**E Directory**
A.2 review	C.2 explanation	E.1 person
A.3 portrait	C.3 recipe	E.2 catalog
A.4 marginal note	C.4 faq	E.3 resource
A.5 interview	C.5 lexicon, word list	E.4 timeline
A.6 news	C.6 biling. dictionary	**F. Communic.**
A.7 feature	C.7 presentation	F.1 mail,talk
A.8 reportage	C.8 statistics	F.2 for.,guestb.
B. Literature	C.9 code	F.3 blog
B.1 poem	**D. Documentation**	F.4 formular
B.2 prose	D.1 law	**G. Nothing**
B.3 drama	D.2 official report	G.1 nothing

Table 1: *A hierarchy of genres*

textlength, forms
$length > 200 \wedge length < 6500 \wedge headlines < 3 \wedge sent > 1$
personal pronouns
$(pronoun2ndP^{norm} < 0.3 \vee pronoun2ndP^{norm} < 0.9) \wedge$
$dirSpeech > 6 * pronoun2ndP) \wedge$
$((pronoun2ndP - 3) * dirSpeech \leq 0 \vee pronoun2ndP < 3)$
part of speech
$verb \geq 5 \wedge adj < 20 \wedge adjPositivNegativ < 0.4$
textual qualities
$causalVocab < 4 \wedge timeMarkers > 0 \wedge$
$names < 15 \wedge questionmarks^{norm} < 0.01$
numbers
$ordNumbers^{norm} < 1.5 \wedge ordNumbers < 3$
spoken/written text
$(contractions < 0.4 \vee dirSpeech > 0) \wedge contr./dirSpeech < 0.2$
tense
$verbsPastTense < 0.18 \wedge verbsPastT. > verbsPresentT. \wedge$
$verbsIngForms > verbsPresentTense$

Table 2: *The rule based classifier for the* news *genre.With the superscript* norm *indicating normalization according to text length.*

4.2 Specialized classifiers

For our specialized genre classifiers, we conducted an aggressive pruning of possible features. The goal was to allow only a small set of significant and natural features for each single classifier. Feature selection was organized on training corpora comprising 20 prototype documents for each genre. 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.[2]. As an example, the specialized classifier of the genre *news* is defined by the conjunction presented in Table 2.

[1] The hierarchy extends previous work by [6, 7].

[2] For additional information about the process of creating the classifiers, please see [25]

5 Search interface

The usual search interface has to be enhanced to give the user the possibility to restrict his document search to certain genres. A genre attribute could be introduced as an additional optional criterion for experienced searchers, analogous to the *filetype attribute* most of the current search engines provide.The yielding graphical interface is shown in Figure 1. To enable an explicit feedback functionality, the result page has to be extended for example with radio-boxes where the user can provide input on the genre of a presented web page (Figure 1). Many variants of the sketched interface are conceivable with a completely *silent interface* as an extreme minimum in the spectrum of interaction that is supposed to minimize the cognitive load of the user. This is an issue especially if more complex search tasks have to be carried out [2, 27]. For the implicit case genres have to be deduced from the gestalt of the query combined with locally or globally aggregated knowledge about the user. The feedback of the user with respect to the suggested genre labels has to be deduced or induced from his observable navigation on the result set [14, 26].

In extension to [1], 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 user events: a **retrieval click**, the watching of a certain document, and an **evaluation click**, a user statement on the genre label of the document. The evaluation click has the following value set: true (1), false (0) and unspecified (0.5). An unspecified evaluation slot can mean two different things: the user is unable to specify the genre of the document or, more probable, he is uncooperative in doing so. If we abstract from questions of query refinement, we can look upon a query, its result, and the click events as a unit denoted as a **turn**. Four cases of annotated results, presented by the search engine, have to be distinguished.

page $x \in N_i$	labeled as N_i	**correct positive**
page $x \in N_i$	not labeled as N_i	**false negative**
page $x \notin N_i$	labeled as N_i	**false positive**
page $x \notin N_i$	not labeled as N_i	**correct negative**

6 User behavior

To analyze the dynamics of a genre search interface, we model different scenarios concerning the user's readiness to cooperatively evaluate the presented genre label.[3] The user's behavior can be divided into four levels.

- **Fully cooperative behavior.** The user retrieves all web pages of the result set and provides an evaluation statement of the annotation labels for the retrieved web pages. Thus, each page of the result set turns into correctly labeled data.

- **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.

Fig. 1: *Example of a graphical interface for genre constrained search and explicit user feedback. For the silent interface, only a label with the genre is displayed.*

(I)	**visited pages**
(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
(II)	**unvisited pages**
(vi)	labeled page that was not visited
(vi.a)	correct positive
(vi.b)	false positive
(vii)	unlabeled page was not visited
(vii.a)	correct negative
(vii.b)	false negative

Table 3: *A taxonomy of feedback events*

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

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

These different attitudes towards evaluation of the system interfere with the principal user behavior - pages watched per turn, ratio between labeled and unlabeled pages visited - and constitute the fundamental user events summarized in Table 3. For semi-cooperative behavior, all events are possible whereas cooperative behavior is inconsistent with (I.iii) and (I.v.b) and uncooperative behavior excludes events (I.i), (I.ii) and (I.iv).

[3] As we will see, the fourth behavior is equivalent to the feedback mode of the silent interface.

7 Adaption of the specialized genre classifiers

A necessary prerequisite to endow our 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 cross-section of the catalog-classifier in its original and DNF form.

Cut-out original catalog classifier

$$\begin{array}{l} currency > 3 \wedge formular > 0 \wedge currency^{Rel} > 1.5 < \\ \quad \vee \\ currency^{Rel} > 1.5 \wedge currency^{Rel} < 20 \wedge currency > 5 \end{array}$$

Cut-out DNF catalog classifier

$$\begin{array}{l} currency \geq 3.1 \wedge currency \leq POS\_INF \wedge \\ formular \geq 0.1 \wedge formular \leq POS\_INF \wedge \\ currency^{Rel} \geq 1.51 \wedge currency^{Rel} \leq POS\_INF \\ \quad \vee \\ currency \geq 5.1 \wedge currency \leq POS\_INF \wedge \\ formular \geq 0 \wedge formular \leq POS\_INF \wedge \\ currency^{Rel} \geq 5.1 \wedge currency^{Rel} \leq 19.9 \end{array}$$

To achieve a correct classification of the input document, 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:

1. *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 if the sum does not exceed a maximum threshold that prevents adaption to outliers or deliberately wrong feedback.
   Constraint: The performance of this temporarily adapted classifier is then computed for all documents seen so far to find out whether the changes lead to an overall improvement. Generally, the files 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). In particular, for the purpose of modeling *genre-shift*, i.e. the modification of feature-value sets that determine a genre, a temporal discount factor can be introduced.[4] In the same way, preference for higher precision or recall can be influenced by appropriate weighting. In our experiments, we used the positive evolution of the *F1-measure* as the constraint for rule adaption.

2. *False positive:* A document of genre N_j has been falsely recognized as genre N_i. We identify elements of the disjunction that have confirmed 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.if the sum does not fall bellow a minimum threshold that prevents adaption to outliers or deliberately wrong feedback.
   Constraint: Generally, 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*).

8 User behavior and information gain

Given the taxonomy of feedback events introduced in section 6, the question arises of how information can be derived under the conditions of an increasingly realistic model of user behavior. Two major problems have to be faced: the loss of information, particularly important for the use of annotated data in test environments involving users, and the introduction of noise.

8.1 Fully cooperative User

The *fully cooperative user* provides the interface with complete information about the binary classification of the presented data. All documents of the result set are labeled whether they belong to the desired genre or not. In this way, the provided information is equivalent to a completely labeled additional dataset. 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.

8.2 Cooperative User

A rational cooperative user will retrieve pages of the desired genre and will give feedback as to whether they were correctly classified. If not enough positively labeled pages are available, it can be assumed that the user will try to identify documents of the desired genre by the snippet information (s.b.), retrieve pages, and give feedback on the genre. With prior knowledge about the underlying genre distribution and recall/precision of the basic classifiers, we can model probabilities of the occurrence of useful events for classifier adaption.

According to studies of standard search engines [8, 24], the average number of visited pages per search session is less than two and in most cases these two pages are retrieved from the first 20 hits of the search results. As is immediately clear, given a fair amount of genre labeled documents, an average number of only two retrieved pages per turn leads to a strong preference of events that can help to improve precision. To increase the number of negative examples, under the condition of cooperative user behavior, we can force the user at the cost of immediate performance to provide more useful information. To prevent the bias to precision related examples (positives) we initially lower the ranges, thus deteriorating precision and enlarging recall. Since we know that the feedback will help us to improve precision, we can recover to a higher level of F1

[4] *Genre-shift* can happen globally, within the web community, or locally, for certain user aggregates or a single user.

performance. A shortcoming of this solution is that we can lose the cooperation of the users altogether if we frustrate them with too weak performance.

8.3 Semi-cooperative User

Under the assumption of non-sabotage behavior, the semi-cooperative case only reduces the amount of available new data for the adaption process and can be modeled by the cooperative case if enough explicite feedback data is available. Otherwise it will be modeled by the uncooperative case.

8.4 Uncooperative User

With uncooperative user behavior, only the *lingering time*, an implicit source of information, is available to generate user statements. The lingering time of the user on a retrieved result page, depending on genre, topic, and model exogenous factors, is transformed into a binary signal: if it exceeds a certain threshold τ, a positive relevancy signal for the document is assumed. Otherwise the document is considered as non-relevant.

If we abstract from model exogenous events, a negative signal means that the document is irrelevant either because of the wrong topic or wrong genre. We could use such a signal to derive evaluation data on genre classification for the cases of false positives and correct negatives. Unfortunately, in a realistic scenario the precision of a search engine with regards to topic seems to be far from perfect. This prevents us from gathering reliable data on the correctness of the genre classification via a negative relevancy signal.

This leaves the case where the lingering time exceeds the threshold and a positive relevancy signal is hypothesized. This hypothesis is incorrect if the user stays on the web page because of exogenous factors. The two probabilities, $P(relevant(x)|time(y) > \tau)$ and $P(\neg relevant(x)|time(y) > \tau)$, can be estimated by frequency counts during a controlled user study.

If a document is actually relevant, this case again can be further divided into relevancy of topic with and without the document being of the desired genre, $P(c(x) = label(x)|relevant(x)), P(c(x) \neq label(x)|relevant(x))$.[5] For a rational user of the genre search interface, we expect the relevant cases that come with the wrong genre to be much rarer than those with the correct genre, $P(c(x) = label(x)|relevant(x)) >> P(c(x) \neq label(x)|relevant(x))$.

After collecting data for the estimation of these basic probabilities, the problem of data loss and introduction of noise for the four a posteriori events of genre recognition can be modeled:

(1) Document of desired genre. The case of a document x actually being of desired genre, $c(x) = N^{desired}$ is subdivided into correct positives, $c(x) = label(x)$, and false negatives, $c(x) \neq label(x)$. [6]

(1.1) Correct positive. To get a positive relevancy signal for cases where the correct genre has been recognized the topic must be relevant. Insofar, we have to expect data loss with a factor of $1 - precision(topic(x))$ and a small data gain via accidental confirmations by an exogenous event.

(1.2) False negative. The more interesting case, however, is the case where the document is of the desired genre but was not recognized. Data gained for this case can improve the recall of the classifiers. As to the confirmation by exogenous events the probability is the same as for case (1.1). A difference exists concerning the loss of data. Not only is data lost by irrelevant topic but also by the user not recognizing the document as being of the desired genre. The problem lies in the indirect access to the document only enabled via the document's snippet. Data loss is additional in the size of the *snippet recognition factor* (s.b.).

For both cases (1.1,1.2), we get no introduction of noise since noise could only be introduced by a negative lingering signal. However, as mentioned, negative signals are not reliable and because of that are left out of consideration.

(2) Document not of desired genre. As for the documents of a genre different to that desired, $c(x) \neq N^{desired}$, we have false positives $c(x) \neq label(x)$ and correct negatives $c(x) = label(x)$.

(2.1) False positive. The problem with the data gain for false positives is that they can be amplified by a positive lingering signal. For the relevant documents, this is the portion where the topic is relevant and the genre is not, $P(relevant(x)|time(x) > \tau)P(c(x) \neq label(x)|relevant(x))$. For the non-relevant documents where the genre was falsely identified as the desired, this is the portion that is amplified by an exogenous event, $P(\neg relevant(x)|time(x) > \tau)P(c(x) \neq label(x)|\neg relevant(x))$.

(2.2) Correct negative. For the last case, the correct negatives, in a rational environment where the user only retrieves documents that he assumes to be of the wanted genre, the introduction of noise depends on *the snippet recognition factor*. If a document is retrieved via misrecognition of the snippet, it can be wrongly confirmed by an exogenous event.

Since for the introduction of noise the correlation between relevancy and *lingering time* and furthermore between relevancy and genre relevancy is crucial and so far, to the best of our knowledge, no experimental results are available, in this paper we can only give experiments on the question of how robust classifiers are against the introduction of noise. For the other central parameter of implicit user feedback, the *snippet genre recognition factor*, we give first experiments in the next section.

9 Experiments on the snippet genre recognition factor

If a user retrieves a document from the result set despite it not having been positively labeled, for the rational case this means that the user concludes it does fall among the desired genre. Since the document's snippet is the communicative act of the search engine to feature the results of a user query, it is fundamental for the implicit user feedback how well the user performs in deriving the genre of a document from its snippet. A typical snippet can be found in Figure 1. To the best of our knowledge no literature has been established on the problem of the *snippet recognition factor*. We give first experiments to open the discussion. To this end, we used our annotated genre corpus [25]: we

[5] Note that the case with irrelevant topic and correct genre falls into the category of non-relevant documents.

[6] For the case of multiple desired genres, this has to be rewritten to $c(x) \in \bigcup N_i^{desired}$.

$Genre$	$Precision$	$Recall$
A.1 commentary	42.86	48.00
A.2 review	68.42	52.00
A.3 portrait	84.21	64.00
A.4 marginal note	45.00	36.00
A.5 interview	90.90	40.00
A.6 news	32.35	44.00
A.7 feature	34.48	40.00
A.8 reportage	35.71	40.00

Table 4: *Users' recognition of journalistic genres by snippets. Precision and recall in percent.*

$Genre$	$Precision$	$Recall$
E.2 catalog	90.57	87.27
C.4 faq	98.67	82.22
F.3 blog	62.50	90.90
A.6 news	77.65	69.47

Table 5: *Users' recognition of the genres blog, catalog and faq by snippets. Precision and recall in percent.*

selected a document and set up a query to a search engine (Google). The query was a combination of several keywords that the engine would use to construct the snippet and a defining N-Gram to make sure the selected document of the genre corpus would be retrieved.

For experiment (1) we chose the eight journalistic genres of our hierarchy and retrieved five snippets for each of them. These 40 snippets were presented to five users with the request to classify them within a time range of < 15 sec each. Table 4 shows the results. With an overall precision of 54.24% and a recall of 45.50%, the results point to a high amount of data loss. The low recognition rate could also lead to some amount of noise introduced by a combination of falsely retrieved documents and exogenous induced lingering.

On the other hand, the genres *interview* and *portrait* seem to be identifiable with high accuracy. The variations are caused by the differences in the communication of the document genre by the snippet.

For experiment 2 we chose the more distinct genres: *blog*, *catalog* and *faq*. Here 20 snippets of each genre were presented to the five users. Additionally, we mixed in 10 *news* documents. The results are summarized in Table 5. These more distinct genres seem to be much easier to distinguish.[7] Only for *blog* and *news* a higher number of documents is confused. For *faq* pages the editors of the pages take care that the acronym occurs in the heading of the page. This heading is then communicated by the search engine as the heading of the snippet which makes it very easy for the users to recognize the genre. With restrictions this is also true for *blog, interview* and *portrait*.

The data so far shows that for certain genres a significant amount of noise and data loss has to be predicted while for others the recognition rate is nearly perfect. We plan to conduct a comprehensive user study for the complete hierarchy.

10 Experiments on classifier adaption

Different user attitudes towards system evaluation interfere with the principal user behavior: how many pages visited per turn and the ratio between labeled and unlabeled pages visited. Consistent with [8, 24], we set, on average, a number of two retrieved pages per turn. If both labeled and unlabeled pages are present, the user visits the labeled pages. If the turn derives only unlabeled pages, the user is assumed to be able to derive the desired genre with a certain accuracy from the snippet (*snippet genre recognition factor*).

To conduct the experiments for classifier adaption, we used annotated genre data. In the first experiment on the incremental adaption of three example classifiers, *blog, catalog,* and *faq*, we used the corpus provided by Marina Santini [21, 22] split into 160 documents for training and 40 documents for measuring recall. For the training/testing with negative examples we used 620 documents of 31 different genres for training, enlarged by a random sample of 360 web pages, and 620 documents for the measuring of fallout. From the training corpora, we randomly generated 48 result sets to simulate the user behavior. Each set consisted of 20 documents, containing on average 3 documents of the desired genre.[8]

In the second experiment we used a collection of 400 documents for the two journalistic genres: *interview* and *news*. For the negative examples we used a corpus of 1,000 random web pages from the Spirit Collection [13].

In addition to the adaption of our rule based classifiers, we give experiments on the performance of an SVM-classifier provided with an extended training set [9, 10]. To reach comparability for each genre, we used only the aggressively pruned feature set of the specialized classifiers.

For the 32 genres of the hierarchy, our initial classifiers showed on average a recall of 60.5% and a precision of 65.4% [25]. The performance of the single classifiers used in the experiments on feedback are given in the respective tables.

10.1 Fully cooperative user

The *fully cooperative user* provides the interface with complete information about the binary classification of the presented data, establishing a completely labeled additional dataset. The results of the adaption process of the rule based classifiers are shown in Table 6. We give results for recall (R) and fall-out (F) for the original and for the adapted classifiers. Recall is the percentage of the genre set that is recognized, $f(label(x) = N_i | x \in N_i)$; fallout is the percentage of the documents in the general data set of distinct genre that are falsely recognized, $f(label(x) = N_i | x \notin N_i)$.

In Table 7 we present the results of the adaption of a SVM-classifier [9, 10]. Despite there have been proposals to incrementally adapt SVMs by estimating a neighborhood of the new data [17] most of the implementations [28, 10] do not provide such a feature and a recomputation of the complete data is needed. The adaptive results for the aggressively pruned feature sets come close to the adapted rule based classifiers. For one genre, *faq*, the algorithm did not converge and a forced termination led to an extremely

[7] Note that compared to Experiment 1, a higher baseline has to be taken into account since in Experiment 2 only four different genres are classified instead of eight.

[8] Since topic was of no interest for this paper, it is reasonable to randomly generate the result sets.

Genre	$Recall^{Test-Rule}$	$Fallout^{Test-Rule}$
blog	72.50(57.50)	1.85(0.13)
catalog	52.50(40.00)	1.19(0.27)
faq	77.50(52.50)	4.29(1.20)
interview	67.50(55.00)	2.26(1.61)
news	30.00(5.00)	12.00(1.50)

Table 6: *Fully cooperative case:Results for Recall and Fallout (in percent) of the adapted and the original classifiers (in parentheses). Test set for the first three genres, homogeneous with the data of the adaption process (Santini corpus).*

Genre	$Recall^{Test-SVM}$	$Fallout^{Test-SVM}$
blog	72.50(65.00)	2.14(1.07)
catalog	47.50(42.50)	1.37(0.31)

Table 7: *Fully cooperative case:Results for Recall and Fallout (in percent) of an SVM classifier trained on the the extended and the original data set (in parentheses).*

poor performance. We omit these results. Summarized, even assuming a static feature space, a significant improvement of the classification can be achieved by using fully labeled data.

10.2 Cooperative User

A rational cooperative user will retrieve pages of the desired genre and will give feedback whether or not 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 pages, and give feedback on the genre. In an experimental run for the *faq* corpus, out of the 48 result sets a feedback of 60 correct positives, 28 false positives, 6 false negatives, and 2 correct negatives emerged. Interestingly, for our experimental design, the rule classifiers can be improved significantly even by this small number of additional examples (Table 8). This phenomenon can be described as a case of active learning [23] in that only a few interesting examples are enough to adapt the borders of a classifier.[9] Also for the two converging SVM classifiers the small amount of additional training examples led to an improvement of F1 values (Table 9).

As a result of the experiments, we can state that by only doing a fraction of the labeling we nearly get the same improvements as for the completely labeled data set provided by a fully cooperative user.

10.3 Uncooperative user

Since the semicooperative case only reduces the amount of available data, for the experiments, we skipped this case and switched to uncooperative user behavior. With uncooperative user behavior, only the *lingering time*, an implicit source of information, is available to generate user statements.

Genre	$Recall^{Test-Rule}$	$Fallout^{Test-Rule}$
blog	83.40(57.50)	6.36(0.13)
catalog	52.50(40.00)	1.06(0.27)
faq	75.00(52.50)	1.91(1.20)
interview	65.00(55.00)	1.93(1.61)
news	25.00(5.00)	8.00(1.50)

Table 8: *Cooperative case: Results for Recall and Fallout (in percent) of the adapted and the original classifiers (in parentheses) for the test set for the first three genres, homogeneous with the data of the adaption process (Santini corpus).*

Genre	$Recall^{Test-SVM}$	$Fallout^{Test-SVM}$
blog	72.50(65.00)	2.14(1.07)
catalog	45.00(42.50)	1.98(0.31)

Table 9: *Cooperative case: Results for Recall and Fallout (in percent) for an SVM classifier trained on the the extended and the original dataset.*

For our experiments on the classifier adaption we deliberately introduced the basic probabilities of 0.9 for the lingering time exceeding τ with given document relevancy and 0.95 for a relevant document being of relevant topic *and* relevant genre. The topic precision was set to 0.5.

Those values lead to a data loss of 45% for the correct positives and the false negatives. By the exogen factors we get a 12% introduction of noise, wrong positive amplification of documents that are not of the desired genre, for the chosen correct negatives and the false positives. For the introduced probabilities the adaption of the specialized rule based classifiers leads to the results summarized in Table 10.

For the experiment with *faq* we received 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.

In Table 11 we present the results of the adaption of the SVM-classifier applied in a soft-margin version. For both classifier types we observed fairly robust improvements despite the data loss and introduction of noise.

Genre	$Recall^{Test-Rule}$	$Fallout^{Test-Rule}$
blog	72.50(57.50)	2.26(0.13)
catalog	52.50(40.00)	0.97(0.27)
faq	67.50(52.50)	1.91(1.20)
interview	60.00(55.00)	1.77(1.61)
news	10.00(5.00)	4.50(1.50)

Table 10: *Uncooperative case:Results for Recall and Fallout (in percent) of the adapted and the original classifiers (in parentheses). Test set for the first three genres, homogeneous with the data of the adaption process (Santini corpus).*

[9] Note that we worked with a snippet recognition of 100%; if this parameter is reduced, we loose false negative examples that help to improve recall.

Genre	$Recall^{Test-SVM}$	$Fallout^{Test-SVM}$
blog	57.50(65.00)	2.14(1.07)
catalog	45.00(42.50)	0.92(0.31)

Table 11: *Uncooperative case:Results for Recall and Fallout (in percent) of SVM classifiers trained on the extended and the original datasets.*

11 Conclusion

We introduced elements for the steady improvement of a genre search interface. The interface exploits data derived from observations of user behavior based on a taxonomy of feedback events. For experiments with corpus based simulated user events, we could achieve significant improvements of the original classifier setup. The improvements showed a remarkable stability against noise and data loss caused by miscategorized user events for more realistic, less cooperative user models.

With regards to the snippet recognition factor, we infer from first experiments that the ability to identify the genre of a document by its snippet varies significantly between the genres. Overall, the recognition accuracy seems high enough to derive data from events where the user chooses a document that was not classified as the desired genre.

Our future goals are to provide a prototype of a genre interface to collect data for the estimation of currently assumed probabilities, as, for example, the correlation between lingering time and the correctness of genre classification by snippets, and to extend the classifier adaption from static to dynamic feature space.

References

[1] R. Baeza-Yates, C. Hurtado, M. Mendoza, and G. Dupret. Modeling user search behavior. In *LA-WEB '05: Proceedings of the Third Latin American Web Congress*, page 242, Washington, DC, USA, 2005. IEEE Computer Society.

[2] D. J. Bell and I. Ruthven. Searcher's assessments of task complexity for web searching. In *ECIR*, pages 57–71, 2004.

[3] E. S. Boese and A. E. Howe. Effects of web document evolution on genre classification. In *CIKM '05: Proceedings of the 14th ACM international conference on Information and knowledge management*, pages 632–639, New York, NY, USA, 2005. ACM Press.

[4] A. Broder. A taxonomy of web search. *SIGIR Forum*, 36(2):3–10, 2002.

[5] A. Z. Broder and A. C. Ciccolo. Towards the next generation of enterprise search technology. *IBM Syst. J.*, 43(3):451–454, 2004.

[6] K. Crowston and M. Williams. Reproduced and emergent genres of communication on the world-wide web. In *30th Hawaii International Conference on System Sciences (HICSS) (6)*, pages 30–39, 1997.

[7] J. Dewe, J. Karlgren, and I. Bretan. Assembling a balanced corpus from the internet. In *Proceedings of 11th Nordic Conference of Computational Linguistics*, Copenhagen, Denmark, 1998.

[8] B. J. Jansen, A. Spink, J. Bateman, and T. Saracevic. Real life information retrieval: A study of user queries on the web. *SIGIR Forum*, 32(1):5–17, 1998.

[9] T. Joachims. Text categorization with support vector machines: learning with many relevant features. In *Proc. 10th European Conference on Machine Learning ECML-98*, pages 137–142, 1998.

[10] T. Joachims. A statistical learning learning model of text classification for support vector machines. In *SIGIR '01: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 128–136, New York, NY, USA, 2001. ACM Press.

[11] T. Joachims, L. Granka, B. Pan, H. Hembrooke, and G. Gay. Accurately interpreting clickthrough data as implicit feedback. In *SIGIR '05: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 154–161, New York, NY, USA, 2005. ACM Press.

[12] H. Joho and J. M. Jose. A comparative study of the effectiveness of search result presentation on the web. In *Advances in Information Retrieval, 28th European Conference on Information Retrieval*, pages 302–313. LNCS, Springer, 2006.

[13] H. Joho and M. Sanderson. The spirit collection: an overview of a large web collection. In *SIGIR Forum, 38(2)*, pages 57–61, 2004.

[14] D. Kelly and J. Teevan. Implicit feedback for inferring user preference: a bibliography. *SIGIR Forum*, 37(2):18–28, 2003.

[15] S. Meyer zu Eissen and B. Stein. Genre classification of web pages. In *KI*, pages 256–269, 2004.

[16] M. Nilan, J. Pomerantz, and S. Paling. Genres from the bottom up: What has the web brought us? In *American Society for Information Science and Technology Annual Meeting*, pages 330–339, 2001.

[17] L. Ralaivola and F. d'Alché Buc. Incremental support vector machine learning: A local approach. In *ICANN '01: Proceedings of the International Conference on Artificial Neural Networks*, pages 322–330, London, UK, 2001. Springer-Verlag.

[18] G. Rehm. Towards automatic web genre identification. In *HICSS '02: Proceedings of the 35th Annual Hawaii International Conference on System Sciences (HICSS'02)-Volume 4*, Big Island, Hawai, USA, 2002. IEEE Computer Society.

[19] M. A. Rosso. *Using Genre to Improve Websearch*. PhD thesis, University of North Carolina Chapitl Hill, 2005.

[20] M. A. Rosso. What type of page is this?: genre as web descriptor. In *JCDL '05: Proceedings of the 5th ACM/IEEE-CS joint conference on Digital libraries*, pages 398–398, New York, NY, USA, 2005. ACM Press.

[21] M. Santini. Linguistic facets for genre and text type identification: A description of linguistically-motivated features. Technical report, Information Technology Research Institute, University of Brighton, 2005.

[22] M. Santini. Common criteria for genre classification: Annotation and granularity. In *Workshop on Text-based Information Retrieval (TIR-06)*, Riva del Garda, Italy, 2006.

[23] G. Schohn and D. Cohn. Less is more: Active learning with support vector machines. In *ICML '00: Proceedings of the Seventeenth International Conference on Machine Learning*, pages 839–846, San Francisco, CA, USA, 2000. Morgan Kaufmann Publishers Inc.

[24] C. Silverstein, H. Marais, M. Henzinger, and M. Moricz. Analysis of a very large web search engine query log. *SIGIR Forum*, 33(1):6–12, 1999.

[25] A. Stubbe, C. Ringlstetter, and K. U. Schulz. Genre as noise: Noise in genre. *International Journal on Document Analysis and Recognition*, 2007. To appear.

[26] R. White, J. Jose, C. van Rijsbergen, and I. Ruthven. A simulated study of implicit feedback models. In *Proceedings of the European Conference on Information Retrieval (ECIR04)*, pages 311–328, 2004.

[27] R. White, I. Ruthven, and J. Jose. A study of factors affecting the utility of implicit relevance feedback. In *SIGIR '05: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 35–42, 2005.

[28] I. H. Witten and F. Eibe. Data mining: practical machine learning tools and techniques. 2nd edition. Morgan Kaufmann, San Francisco. http://www.cs.waikato.ac.nz/ ml/weka, 2005.