On the Impact of Features and Classifiers for Measuring Knowledge Gain during Web Search - A Case Study

Wolfgang Gritz[1], Anett Hoppe[1,2] and Ralph Ewerth[1,2]

[1]TIB – Leibniz Information Centre for Science and Technology, Hannover, Germany
[2]L3S Research Center, Leibniz University Hannover, Germany

Abstract

Search engines are normally not designed to support human learning intents and processes. The field of Search as Learning (SAL) aims to investigate the characteristics of a successful Web search with a learning purpose. In this paper, we analyze the impact of text complexity of Web pages on predicting knowledge gain during a search session. For this purpose, we conduct an experimental case study and investigate the influence of several text-based features and classifiers on the prediction task. We build upon data from a study of related work, where 104 participants were given the task to learn about the formation of lightning and thunder through Web search. We perform an extensive evaluation based on a state-of-the-art approach and extend it with additional features related to textual complexity of Web pages. In contrast to prior work, we perform a systematic search for optimal hyperparameters and show the possible influence of feature selection strategies on the knowledge gain prediction. When using the new set of features, state-of-the-art results are noticeably improved. The results indicate that text complexity of Web pages could be an important feature resource for knowledge gain prediction.

Keywords

Textual Complexity, Knowledge Gain, Search as Learning, Learning Resources, Web-based Learning

1. Introduction

Conventional information retrieval systems are usually designed to satisfy an information need. The research area Search as Learning (SAL), on the other hand, deals with the assumption that search sessions can also be driven by a learning intention. Research in the area of SAL is not only concerned with the ranking of search results, but also with the detection or prediction of the learning intention or even the knowledge state and knowledge gain [1, 2].

Vakkari [3] presented a survey of features which indicate the user's knowledge and learning needs, but also knowledge gain during the search process. More recently, a wide variety of features were considered, including resource-based (based on text or multimedia content) or behavioral features. For example, Syed and Collins-Thompson [4] have considered document retrieval features to improve learning outcome for short- and long-term vocabulary learning. Collins-Thompson et al. [5], on the other hand, have studied different query types and found a correlation between the variety of intrinsic query types and knowledge gain. Pardi et al. [6] further examined the time spent on Web pages with primarily textual or video content and learning outcome. One finding was that the time spent on text-based Web pages had a greater impact on knowledge gain than time spent on video-based Web pages. Gadiraju et al. [7] explored the influence of behavioral features on the learning outcome, and found a positive correlation between the average complexity of user queries and their knowledge gain. Recently, some approaches have been suggested that combine several types of features [8, 9]. For example, Otto et al. [9] studied the effect on knowledge gain prediction, when complexity and linguistic features are complemented with multimedia features. They achieved slight improvements by adding multimedia features, e.g., representing the amount of image and video data on the screen or the image type (infographics, outdoor photography, etc.).

A crucial aspect of learning is the appropriateness of the text for the reader. In his survey, Collins-Thompson [10] has summarized studies that deal with the automatic assessment of the reading difficulty of texts. Hancke [11] has previously analyzed lexical, syntactic, and morphological features for German, while Kurdi et al. [12] investigated features that allow for conclusions about the complexity of English texts.

In this paper, we investigate the influence of text complexity of Web pages on knowledge gain prediction in a comprehensive experimental case study. For this purpose, we present a large set of text-based features of various types and, furthermore, analyze the impact of different classifiers and feature selection strategies on knowledge gain prediction. First, the experimental results show that state-of-the-art results [9] can be significantly improved and, second, that the textual complexity of Web pages can

Proceedings of the CIKM 2021 Workshops, November 1–5, Gold Coast, Queensland, Australia
wolfgang.gritz@tib.eu (W. Gritz); anett.hoppe@tib.eu (A. Hoppe); ralph.ewerth@tib.eu (R. Ewerth)
0000-0003-1668-3304 (W. Gritz); 0000-0002-1452-9509 (A. Hoppe); 0000-0003-0918-6297 (R. Ewerth)

be a valuable predictor for the classification of knowledge gain. Our contributions can be summarized as follows:

- A large set of features describing textual complexity of Web pages is presented.
- We conduct an extensive, systematic evaluation including multiple classifiers, hyperparameter analysis and optimization, as well as feature selection strategies and analyze their impact on knowledge gain prediction.
- We demonstrate that the state-of-the-art-results can be improved, even when only considering textual complexity features.

The remainder of this paper is structured as follows: In Section 2 the experimental setup and the process of extraction is described. Experimental results are reported in Section 3 and the impact of text complexity features is analyzed. Finally, a summary of the main results and an outlook is given in section 4.

2. Experimental Setup and Text-based Features

We use data from a study [13] in which participants were asked to acquire knowledge about the formation of thunder and lightning. The topic has already proven useful in previous work [14, 15]. It is a phenomenon that is generally known and requires both factual and procedural knowledge. On the Web, many sources exist on the subject, explaining it in diverse ways (texts, graphics, videos, etc.). The participants were asked to do a Web search for a maximum of 30 minutes; but were allowed to end the search earlier if they felt they had learned everything important. We could use data from $N = 104$ participants (88 female, 16 male, average age of 22.7 ± 2.7 years), for which the visited Web pages were downloaded during the experiment. The participants were recruited over a local recruitment portal composed of students from the University of Tübingen. Students were compensated with 16€ per person for participating in the study. None of the participants had former expertise in meteorology.

2.1. Technical Setup of the Study

While plenty of data were collected during the study (data sources such as eye and mouse tracking information), here, we focus on the text content of the visited Web pages. During the Web search, all visited Web pages of the participants were tracked and recorded via the "ScrapbookX" (1.5.14)[1] and "ScrapbookXAutosave" (1.4.3)[2] plugins.

[1] https://github.com/danny0838/firefox-scrapbook

[2] https://github.com/danny0838/firefox-scrapbook-autosave

2.2. Knowledge Gain Measurement

To measure knowledge gain, the participants were asked to solve a 10-item multiple choice test one week before (t1) and immediately after (t2) the Web search. The knowledge gain is subsequently defined as the difference between the numbers of correct answers of t2 and t1. The potential range of values for the knowledge gain is therefore $[-10, 10]$. The average value in t1 was 5.24 ± 1.80 respectively 7.46 ± 1.43 in t2. The average knowledge gain was 2.22 ± 1.78 and lies in the range of $[-3, 6]$.

2.3. Feature Extraction

In the study, the participants performed free Web searches, such that realistic search and browsing behavior could be recorded. Since we focus on the textual complexity of the visited pages, other page types like search engine result pages and video-based contents are filtered. For this purpose, we used a keyword-based approach and omitted pages which contained the following keyterms in their URL: "google.", "youtu", "ecosia", "RDSIndex","universitaetsbibliothek", "meteoros", "webcam" and "learningsnacks". For all remaining pages, we extracted all displayed text without further processing. This can lead to the fact that e.g., tables or advertisements are in the analyzed texts. We decided against any further preprocessing in order to minimize the bias in the data set.

2.4. Website Features

To assess the complexity of text on Web pages, we extract eight different types of features:

- syntactical features
- readability scores
- part of speech (POS) density
- lexical richness
- lexical variation
- lexical sophistication
- syntactic constituents features
- connectives

Since the study was conducted in German, we mainly rely on the Common Text Analysis Platform (CTAP) tool [16], which currently provides 218 different complexity features for the German language. In total, we extract 248 features from each Web page. Below we give a short description of each feature group. For a complete overview consider the appendix[3].

The *syntactic features* group consists of basic text statistics such as the number of letters, syllables, words, and sentences. Moreover, the average length of each

[3] https://github.com/molpood/IWILDS_Complexity_Feature_List/

element is considered, like sentence length in letters or word length in syllables, as well as the standard deviation. In addition, we calculate the average reading time of the Web pages by assuming 180 words per minute [17].

The second group of features consists of well-known *readability scores* that aim to estimate the skills a reader must have to understand the text. The features are based on combinations of the *syntactic features* (automated readability index (ARI), Coleman-Liau index, Flesch-Kincaid grade, Flesch reading ease) and partly on difficult or complex words. They are given either by a list (Dale-Chall readability score, Gunning fog) or by words with three or more syllables (SMOG index). For example, the formula for ARI is as follows:

$$\text{ARI} = 4.71 \cdot \frac{|\text{characters}|}{|\text{words}|} + 0.5 \cdot \frac{|\text{words}|}{|\text{sentences}|} - 21.43$$

In the case of the ARI, the result is a human-interpretable numerical value on a scale of 1-14 (1: Kindergarten, 14: Professor).

The *POS density* group reflects the density of different word types like adjectives or verbs in the website text. It is based on the tokenization of the text and calculates the different number of word types (e.g., adjectives or verbs) in relation to all tokens, e.g.,

$$\text{density}_{\text{adjectives}} = \frac{|\text{adjectives}|}{|\text{tokens}|}$$

The fourth group *lexical richness* is very similar. Here, the number of non-duplicated tokens is set in relation to all tokens. In addition to the fraction $\frac{\text{types}}{\text{tokens}}$, various variations such as the logarithm or square root are applied to the numerator and denominator.

The *lexical variation* group examines the subset of lexical words (LW) consisting of nouns, verbs, adjectives and adverbs. The class puts the number of individual components in relation to the number of lexical words, e.g., the lexical variation lv_adjectives for adjectives:

$$\text{lv\_adjectives} = \frac{|\text{adjectives}|}{|\text{LW}|}$$

The group of *lexical sophistication* features is based on different frequency lists [18, 19]. All words of the Web page text are assigned to sets of all words AW, lexical words LW (as mentioned before consisting of nouns, verbs, adjectives and adverbs) and functional words FW (i.e., not LW). The logarithmic or absolute frequency in the frequency lists (per million words) of AW, LW and FW is consequently used as a feature. Furthermore, the Karlsruhe Childrens Text (KCT) [20] list is used to determine the average and minimum age of active use of AW, LW and FW.

The group of *syntactic constituents* consists of features that determine the number of different syntactic constituents, like noun phrases, relative clauses or T-units. Additionally, ratios to each other are calculated, e.g., noun phrases per T-unit, but also words per T-unit or noun phrases per sentence. Moreover, we consider the tenses in the text based on Kurdi [12]'s observation that there may be a connection between more difficult texts and more complex tenses. To extract the tenses, we use the tool of Dönicke [21].

The last group *Connectives* (according to Breindl et al. [22]) examines units of the German language that express semantic relations between sentences. The connectives form a class consisting of subsets of defined parts of speech like conjunctions (and, or, etc.) or adverbs (in contrast, therefore, etc.). The absolute number of connectives, as well as ratios, such as multi-word connectives divided by single-word connectives, are calculated as features.

The eight groups consist of a total of 248 features that are calculated for each Web page visited during the search sessions. Since the participants accessed a different number of Web pages, we compute the average, the minimum and the maximum for each feature for each participant. As a result, we obtain a total of $3 \cdot 248 = 744$ features for knowledge gain prediction.

3. Experimental Results

In this section, we report results for knowledge gain prediction using features for text complexity. For a fair comparison, we use the same evaluation setting including hyperparameter optimization for all experiments. In the same way, we replicate the results of Otto et al. [9][4] with our evaluation procedure.

3.1. Knowledge Gain Definition

To categorize the measured knowledge gain, we use the common approach [7, 8, 9] to assign each search session to one of three classes $C = \{Low, Moderate, High\}$ based on the *Standard Deviation Classification* approach. For this purpose, the knowledge gain X_i of participant i is z-normalized ($\hat{X}_i$) according to equation 1.

$$\hat{X}_i = \frac{X_i - \mu}{\sigma} \tag{1}$$

Here, μ is the mean and σ is the standard deviation of all knowledge gain measures X. Then, for every z-normalized knowledge gain $\hat{X}_i$ the class is assigned as follows:

$$C(X_i) := \begin{cases} \text{Low}, & \text{if } \hat{X}_i < -\frac{1}{2} \\ \text{Moderate}, & \text{if } -\frac{1}{2} \leq \hat{X}_i \leq \frac{1}{2} \\ \text{High}, & \text{if } \hat{X}_i > \frac{1}{2} \end{cases}$$

[4] Otto et al. [9] analyzed features for 113 participants. Technical issues with logging led to missing HTML data for nine participants which were crawled at a later date. We rely on the data crawled during the original experiment, leading to $N = 104$ records for our analysis.

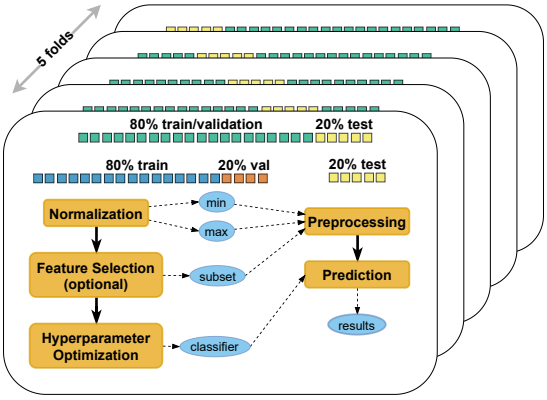

Figure 1: Overview of our evaluation method. A 5-fold cross-validation is performed and for each split the features are first normalized, optionally selected/reduced and the hyperparameters of the respective classifier are optimized on the 80% train/validation data. The test data are scaled with the minimum and maximum of the train/validation data and optionally the features are filtered. Finally, the classifier optimized on the train and validation data is used to predict the knowledge gain on the test data set.

This yields the following class distribution: $|X_{Low}|$=40, $|X_{Moderate}|$=39, $|X_{High}|$=25.

3.2. Metrics

To evaluate the classification results, we use precision, recall, F_1 score, and accuracy. These are defined as follows:

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{2}$$

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{3}$$

$$\text{F}_1\text{ score} = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \tag{4}$$

$$\text{accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{5}$$

where TP are the values correctly classified as positive, TN are the values correctly classified as negative, and FP are the values incorrectly classified as positive and FN are the values incorrectly classified as negative.

3.3. Experimental Setup

Cross-validation is a good way to evaluate the classification result, since every feature vector acts as a test sample in one fold. We thus choose a 5-fold cross-validation with 80% train/validation and 20% test set split. This results in five elements per class in each test set in each iteration of the cross-validation.

We use min-max normalization to normalize each feature of the 80% to the interval $[0, 1]$. This is an essential step for some of the classifiers, e.g., Support Vector Machine. The 20% test set is then normalized by the minimum and maximum of the 80% for evaluation. It is possible that the values lie outside the interval of $[0, 1]$. However, we decide against clipping in order to not lose any information due to normalization. Figure 1 provides an overview of our proposed evaluation. In our evaluation we use the implementation of Scikit-learn [23].

3.3.1. Hyperparameter Optimization

The performance of classification algorithms strongly depends on the chosen hyperparameters. However, since the training, validation and test data change in each iteration due to cross-validation, these cannot be determined once and used for the entire evaluation. Therefore, to obtain valid results, we perform an optimization of the hyperparameters in each of the five iterations. We utilize Optuna [24] for an Bayesian search to efficiently find a good configuration and limit the number of runs to 500 to reduce the computational cost. From the 80% of the data coming from the 80:20 split of the cross-validation, another 80:20 split is performed, where 80% is training data and 20% is validation data. We set the maximization of the weighted F_1 score as the optimization objective. This is to prevent the class imbalance from making the underrepresented class *High* less important, as it would be, for example, with overall accuracy.

3.3.2. Feature Selection

The classification results may also depend on the number of input features (more is not always better). For example, in the Random Forest algorithm, a subset of the features is selected several times to create weak classifiers and there is no guarantee that "good features" will prevail. For this reason, we want to reduce the number of features while trying to preserve valuable features. Again, it is important to separate the feature selection from the test data, which changes in each iteration. As with hyperparameter optimization, we use the further split into training and validation data to do this. It follows that the selected features may change in each iteration. For the selection of the features to be used for this evaluation, we rely on two strategies:

1. χ^2**-based Feature Selection:** This method examines whether a feature has a statistically significant relationship to knowledge gain. While one feature is analyzed for a relationship, all other features are ignored. The features with the N highest values based on the χ^2-test are selected.
2. **Tree-based Feature Selection:** Features without a direct correlation to the knowledge gain

Table 1
Results of the Knowledge gain classification for the classes *Low*, *Moderate* and *High* respectively for the classifiers (clf) Adaboost (Ada), Decision Tree (DT), K-Nearest Neighbors (KNN), Multi-layer Perceptron (MLP), Random Forest (RF) and Support Vector Machine (SVM) and for weighted guessing (WG). For the reported results of Otto et al. [9] (*Otto**) and reproduced (*Otto*), respectively, our results (*our*) and the combination of Otto et al. 's [9] and our features (*Otto+our*), precision (pre), recall (rec), F_1 score (f1) and overall accuracy (accu) are reported.

		Low			Moderate			High			macro scores			
	clf	**pre**	**rec**	**f1**	**pre**	**rec**	**f1**	**pre**	**rec**	**f1**	**pre**	**rec**	**f1**	**accu**
	WG	**38.4**	**38.6**	**38.4**	**37.4**	**37.3**	**37.2**	**24.0**	**24.0**	**23.8**	**33.3**	**33.3**	**33.1**	**34.6**
Otto*	**RF**	**41.5**	**52.0**	**46.1**	**39.1**	**40.0**	**39.5**	**28.4**	**14.8**	**19.1**	**36.4**	**35.6**	**34.9**	**38.7**
Otto	**Ada**	**42.1**	40.0	41.0	35.4	43.6	39.1	16.7	12.0	14.0	31.4	31.9	31.4	34.6
	DT	40.0	**50.0**	**44.4**	40.5	38.5	39.5	23.5	16.0	19.0	34.7	34.8	34.3	37.5
	KNN	26.7	20.0	22.9	38.1	41.0	39.5	18.8	24.0	21.1	27.8	28.3	27.8	28.8
	MLP	41.2	35.0	37.8	**46.3**	**48.7**	**47.5**	**34.5**	**40.0**	**37.0**	**40.7**	**41.2**	**40.8**	**41.3**
	RF	30.2	40.0	34.4	32.6	35.9	34.1	12.5	4.0	6.1	25.1	26.6	24.9	29.8
	SVM	38.6	42.5	40.5	45.5	38.5	41.7	33.3	36.0	34.6	39.1	39.0	38.9	39.4
our	**Ada**	**42.1**	40.0	41.0	41.9	46.2	43.9	**34.8**	32.0	**33.3**	39.6	39.4	39.4	40.4
	DT	39.0	40.0	39.5	41.7	38.5	40.0	29.6	32.0	30.8	36.8	36.8	36.8	37.5
	KNN	21.7	12.5	15.9	45.0	46.2	45.6	26.8	**44.0**	**33.3**	31.2	34.2	31.6	32.7
	MLP	38.9	35.0	36.8	**55.9**	48.7	**52.1**	26.5	36.0	30.5	40.4	39.9	39.8	40.4
	RF	40.5	**42.5**	**41.5**	51.3	**51.3**	51.3	**34.8**	32.0	**33.3**	**42.2**	**41.9**	**42.0**	**43.3**
	SVM	40.6	32.5	36.1	54.3	48.7	51.4	27.0	40.0	32.3	40.6	40.4	39.9	40.4
Otto+our	**Ada**	42.3	**55.0**	**47.8**	**54.8**	43.6	48.6	**42.9**	36.0	39.1	**46.7**	**44.9**	**45.2**	**46.2**
	DT	**53.1**	42.5	47.2	42.9	38.5	40.5	27.0	40.0	32.3	41.0	40.3	40.0	40.4
	KNN	20.0	12.5	15.4	45.0	46.2	45.6	25.6	40.0	31.2	30.2	32.9	30.7	31.7
	MLP	25.0	12.5	16.7	41.3	**66.7**	**51.0**	28.6	24.0	26.1	31.6	34.4	31.2	35.6
	RF	37.5	45.0	40.9	45.0	46.2	45.6	37.5	24.0	29.3	40.0	38.4	38.6	40.4
	SVM	22.7	12.5	16.1	41.7	38.5	40.0	30.4	**56.0**	**39.4**	31.6	35.7	31.9	32.7

can be important predictors in combination with other features. For this reason, we employ a tree-based approach using a Random Forest classifier. This is fitted to the training data and then analyzed to see which features were most heavily used in the decision. The N values with the highest importance are selected. The goal is to select valuable features for the classification even without direct correlation.

3.3.3. Classifiers

Otto et al. [9] limit their evaluation to a *Random Forest* [25] classifier. In addition to that, we explore several alternative classifiers: *Adaboost* [26], *Decision Tree* [27], *K-Nearest Neighbors* [28], *Multi-layer Perceptron* [29], and *Support Vector Machine* [30]. The objective is to experimentally determine the best configuration in order to find the maximum potential for knowledge gain prediction, given the set of features.

3.4. Classifier Performance

In Table 1, we compare the performance for all classifiers. As baselines, we list the results for weighted guessing (*WG*), which is the mean of each metric for 10,000 randomly generated vectors consisting of class labels with respect to the class distribution, and the original reported results from Otto et al. [9] (*Otto**). For a fair comparison with *our* features, we reproduced the results using the features from Otto et al. [9] with our pipeline (*Otto*). Furthermore, to analyze the performance for a feature set as diverse as possible, we combined the features of *Otto* et al. [9], and *our* proposed feature set for evaluation (*Otto+our*). For the cumulative predictions for all five iterations of cross-validation, the precision, recall, and F_1 score are calculated for each class (*Low*, *Moderate*, and *High*), as well as the average of these metrics over all classes, and the overall accuracy.

First, it is notable that the reproduced results of Otto et al. [9] (*Otto*) are better compared to their reported result (*Otto**). The results of the Multi-layer Perceptron (MLP) provide a 5.9% higher F_1 score (34.9% compared to 40.8%). However, in direct comparison to the repro-

duced result with a Random Forest (RF), the original results are better. It is striking, that the improved outcome stems mainly from better predictions from the class *High*. A closer look reveals that the recall scores for the tree-based classifiers Adaboost (Ada), Decision Tree (DT) and Random Forest (RF) are comparatively low. These algorithms seem to preferentially predict the more represented classes for the features of Otto et al. [9] and accept a worse result for the underrepresented class *High*. This impression is enforced by the fact that for all feature sets the F_1 score (f1) for the three classifiers is significantly worse for the class *High* than for the classes *Low* and *Moderate*. This is not the case for any of the other classifiers.

Nevertheless, Random Forest (RF) and Adaboost (Ada) perform best for the other feature sets (*our* and *Otto+our*). The RF using the features of textual complexity (*our*) yields a slightly better macro F_1 score (42.0%) than the MLP using the features of *Otto* (41.2%). In addition, the RF achieves an overall accuracy of 43.3% while the MLP only achieves 40.6%. The best result is obtained by the Adaboost classifier for *Otto+our* with 45.2% macro F_1 score and 46.2% overall accuracy. Examining the results for the Random Forest algorithm for all three feature sets, we notice that the F_1 scores of all three classes for the combination of features are strictly between the F_1 scores of the individual feature sets. At the same time, the F_1 scores for the combination of features are all better than for the individual sets for Adaboost. We assume that the Random Forest algorithm is affected by too many (diverse) features. Adaboost can weight the features differently and thus utilize the strengths of both feature sets.

Another observation is that the F_1 scores of all feature sets for the K-Nearest Neighbors (KNN) algorithm are significantly higher for the class *Moderate* than for the classes *Low* and *High*. Therefore, we suspect that search strategies with *Low* (or *High*) knowledge gain differ much more. Furthermore, we can observe that the F_1 score for the class *Moderate* of *our* features is high compared to the classes *Low* and *High*, independent of the classifier. On closer inspection, we found that often instances of the class *Low* are classified as *High* and vice versa. If we put the classification result for the classes *Low* and *High* together, i.e., a new class *Not Moderate*, we would get 74.1%, 70.8% and 73.1% F_1 score for the classifiers MLP, RF and SVM, respectively, for this new class. It seems like the complexity features are useful to detect if someone does not have a *Moderate* increase in knowledge gain. We plan to investigate this interesting aspect in the future.

For our textual complexity features, the best result was obtained with the Random Forest classifier. In each iteration of the 5-fold cross-validation, an independent hyperparameter optimization was performed. The optimized hyperparameters for each fold $F_1, \ldots, F_5$ are shown in Table 2. No pattern can be discovered in the parameters, they are very different in shape. This could possibly be related to the heterogeneity of the data and the weakness of the features for prediction.

Table 2
The optimized hyperparameters per fold $F_1, \ldots F_5$ for the Random Forest classifier for *our* features.

	F_1	F_2	F_3	F_4	F_5
estimators	242	299	154	150	223
max_depth	22	17	8	17	17
max_features	sqrt	log2	sqrt	log2	sqrt
criterion	entr.	gini	sqrt	entr.	gini
min_n_split	6	3	7	7	4
min_n_leaf	5	8	3	8	7

3.5. Feature Selection

In Table 1, it is observable that the classification result for the Random Forest classifier (RF) performs worse for the combination of features (*Otto+our*) than for the complexity-only features (*our*). It seems that considering more features does not necessarily improve the classification quality. The result for the Random Forest classifier (RF) for the textual complexity (*our*) features for $N \in \{1, 3, 5, \ldots, 99\}$ is shown in Figure 2. It can be seen, that the classification result is achieved with fewer features, regardless of the feature selection strategy. With the χ^2-based selection method, the result is also achieved with fewer features, but later than with the tree-based method. This makes sense in so far as the

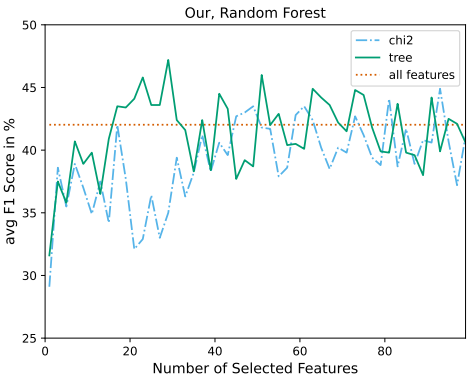

Figure 2: Average F_1 scores of the Random Forest classifier using $N \in \{1, 3, 5, \ldots, 99\}$ of *our* features for the χ^2-based (chi2) and the tree-based (tree) Feature Selection strategy. The result for all features is indicated with the dotted line.

Table 3
Features selected at least three out of five times during cross-validation by the tree-based selection strategy.

type	feature	aggregation	count
POS Density Feature	Subordinating Conjunction	min	4
Lexical Sophistication Feature	SUBTLEX Word Frequency (LW Token)	min	4
Syntactic Complexity Feature	Mean Length of Verb Cluster	min	3

χ^2-based method considers the features independently of each other, and only measures the individual correlation of a feature with knowledge gain. In contrast, the tree-based strategy selects features based on their importance for an upstream Random Forest. Thus, the baseline level can already be reached with $N = 19$ features.

Cross-validation is used for evaluation as described above (Section 3.3.1). Similarly, feature selection is performed five times. However, this implies that the features chosen in each iteration of the cross-validation may differ, which complicates the analysis of which features most influence the classification result. We therefore propose to highlight the features that were selected in at least three out of five iterations. Since the classification result of the Random Forest was already achieved with $N = 17$ features, we report the features based on this configuration. The features and their frequencies are shown in Table 3. Three features were selected at least three times, but none were selected in every iteration of the cross-validation. All three were aggregated by the minimum, indicating that the Web page with the lowest textual complexity is most important for the classification result. This strengthens the impression that the features or the aggregations (Minimum, Maximum and Average) are too weak to provide a strong prediction of the knowledge gain. In the future, we aim to include more features and find aggregations that are more suitable to reflect search patterns.

In the last section, it was observed that the F_1 score for the class *High* is significantly below the values for the classes *Low* and *Moderate*, regardless of the feature set. We performed feature selection before hyperparameter optimization and repeated the evaluation with $N \in \{1, 3, 5, ..., 79\}$ features. Figure 3 shows how the F_1 score for the class *High* changes with a subset of the features of Otto et al. [9]. The green curve describes the F_1 scores based on the tree-based feature selection strategy, which tries to select the most important features for classification. It is noticeable that almost any tested subset would have been more suitable than using the full feature set. Moreover, the curve does not change from $N = 65$ onward (same observation for the classes *Low* and *Moderate*), which suggests that the tree-based feature selection strategy does not consider many features at all.

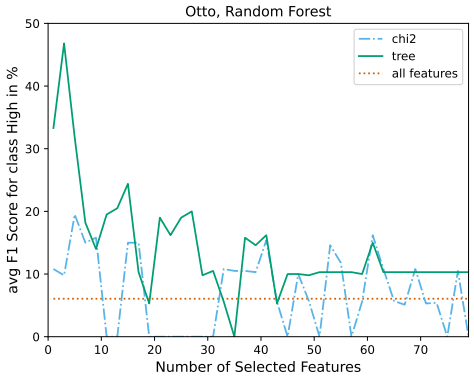

Figure 3: F_1 scores for the class *High* for the features of Otto et al. [9] for $N \in \{1, 3, 5, ..., 79\}$ features for the χ^2-based (chi2) and the tree-based (tree) Feature Selection strategy. The result for all features is indicated with the dotted line.

4. Conclusions

In this paper, we have investigated the impact of textual complexity of Web pages on knowledge gain during a Web search. The experimental results demonstrated that the state of the art can be improved by only considering the textual complexity of Web pages. The results also showed that a systematic assessment of different hyperparameter settings, feature selection, and several classifiers is important – in particular, since the correlations between features and the target outcome are relatively weak. During the evaluation, it became apparent that as little as 17 features per iteration of cross-validation would have been sufficient to achieve the result. Furthermore, we found that a moderate knowledge gain can be predicted relatively well, but, interestingly, the distinction between successful and unsuccessful Web search does not work well (in terms of knowledge gain). The reasons for this effect have to be investigated in more detail.

Although we have obtained state-of-the-art results, there are some limitations. In this case study, we analyzed only the data of a study on knowledge acquisition about a specific science topic, the formation of thunder-

storms. Consequently, limited conclusions can be drawn about general Web searches and the results need to be confirmed or extended by future studies. In this sense, the reported results need to be reproduced for (a) different types of learning tasks (e.g., procedural knowledge) and (b) conceptual learning tasks in other domains (e.g., non-science topics).

In the future, we would like to deepen our understanding of what behavioral patterns characterize effective Web searches, for instance, by examining how the sequence of Web pages (and their characteristics) influence learning success. An intuitive assumption is, for example, that a successful learning session consists of Web pages of increasing complexity. Furthermore, we have considered the textual complexity of the entire Web page, but not in every case is the Web page content read in its entirety. In future work we would like to focus more on the actual seen during Web search.

Lastly, we focused on text-based Web pages in this case study. However, many of the Web searches were not unimodal but multimodal. Consequently, further investigations will need to include further complexity measures such as visual complexity of the Web pages or videos.

Acknowledgments

Part of this work is financially supported by the Leibniz Association, Germany (Leibniz Competition 2018, funding line "Collaborative Excellence", project SALIENT [K68/2017]).

References

[1] A. Hoppe, P. Holtz, Y. Kammerer, R. Yu, S. Dietze, R. Ewerth, Current challenges for studying search as learning processes, in: 7th Workshop on Learning & Education with Web Data (LILE2018), in conjunction with ACM Web Science, 2018.

[2] M. Machado, P. A. Gimenez, S. Siqueira, Raising the dimensions and variables for searching as a learning process: A systematic mapping of the literature, in: Anais do XXXI Simpósio Brasileiro de Informática na Educação, SBC, 2020, pp. 1393–1402.

[3] P. Vakkari, Searching as learning: A systematization based on literature, J. Inf. Sci. 42 (2016) 7–18. URL: https://doi.org/10.1177/0165551515615833. doi:10.1177/0165551515615833.

[4] R. Syed, K. Collins-Thompson, Exploring document retrieval features associated with improved short- and long-term vocabulary learning outcomes, in: C. Shah, N. J. Belkin, K. Byström, J. Huang, F. Scholer (Eds.), Proceedings of the 2018 Conference on Human Information Interaction and Retrieval, CHIIR 2018, New Brunswick, NJ, USA, March 11-15, 2018, ACM, 2018, pp. 191–200. URL: https://doi.org/10.1145/3176349.3176397. doi:10.1145/3176349.3176397.

[5] K. Collins-Thompson, S. Y. Rieh, C. C. Haynes, R. Syed, Assessing learning outcomes in web search: A comparison of tasks and query strategies, in: D. Kelly, R. Capra, N. J. Belkin, J. Teevan, P. Vakkari (Eds.), Proceedings of the 2016 ACM Conference on Human Information Interaction and Retrieval, CHIIR 2016, Carrboro, North Carolina, USA, March 13-17, 2016, ACM, 2016, pp. 163–172. URL: https://doi.org/10.1145/2854946.2854972. doi:10.1145/2854946.2854972.

[6] G. Pardi, J. von Hoyer, P. Holtz, Y. Kammerer, The role of cognitive abilities and time spent on texts and videos in a multimodal searching as learning task, in: H. L. O'Brien, L. Freund, I. Arapakis, O. Hoeber, I. Lopatovska (Eds.), CHIIR '20: Conference on Human Information Interaction and Retrieval, Vancouver, BC, Canada, March 14-18, 2020, ACM, 2020, pp. 378–382. URL: https://doi.org/10.1145/3343413.3378001. doi:10.1145/3343413.3378001.

[7] U. Gadiraju, R. Yu, S. Dietze, P. Holtz, Analyzing knowledge gain of users in informational search sessions on the web, in: C. Shah, N. J. Belkin, K. Byström, J. Huang, F. Scholer (Eds.), Proceedings of the 2018 Conference on Human Information Interaction and Retrieval, CHIIR 2018, New Brunswick, NJ, USA, March 11-15, 2018, ACM, 2018, pp. 2–11. URL: https://doi.org/10.1145/3176349.3176381. doi:10.1145/3176349.3176381.

[8] R. Yu, R. Tang, M. Rokicki, U. Gadiraju, S. Dietze, Topic-independent modeling of user knowledge in informational search sessions, Inf. Retr. J. 24 (2021) 240–268. URL: https://doi.org/10.1007/s10791-021-09391-7. doi:10.1007/s10791-021-09391-7.

[9] C. Otto, R. Yu, G. Pardi, J. von Hoyer, M. Rokicki, A. Hoppe, P. Holtz, Y. Kammerer, S. Dietze, R. Ewerth, Predicting knowledge gain during web search based on multimedia resource consumption, in: I. Roll, D. S. McNamara, S. A. Sosnovsky, R. Luckin, V. Dimitrova (Eds.), Artificial Intelligence in Education - 22nd International Conference, AIED 2021, Utrecht, The Netherlands, June 14-18, 2021, Proceedings, Part I, volume 12748 of *Lecture Notes in Computer Science*, Springer, 2021, pp. 318–330. URL: https://doi.org/10.1007/978-3-030-78292-4_26. doi:10.1007/978-3-030-78292-4_26.

[10] K. Collins-Thompson, Computational assessment of text readability: A survey of current and future research, ITL - International Journal of Applied Linguistics 165 (2014) 97–

135. URL: https://www.jbe-platform.com/content/journals/10.1075/itl.165.2.01col. doi:https://doi.org/10.1075/itl.165.2.01col.

[11] J. Hancke, S. Vajjala, D. Meurers, Readability classification for german using lexical, syntactic, and morphological features, in: M. Kay, C. Boitet (Eds.), COLING 2012, 24th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, 8-15 December 2012, Mumbai, India, Indian Institute of Technology Bombay, 2012, pp. 1063–1080. URL: https://aclanthology.org/C12-1065/.

[12] M. Kurdi, Lexical and syntactic features selection for an adaptive reading recommendation system based on text complexity, in: ICISDM '17, 2017.

[13] J. von Hoyer, G. Pardi, Y. Kammerer, P. Holtz, Metacognitive judgments in searching as learning (sal) tasks: Insights on (mis-) calibration, multimedia usage, and confidence, in: Proceedings of the 1st International Workshop on Search as Learning with Multimedia Information, SALMM '19, Association for Computing Machinery, New York, NY, USA, 2019, p. 3–10. doi:10.1145/3347451.3356730.

[14] R. Mayer, R. Moreno, A split-attention effect in multimedia learning: Evidence for dual processing systems in working memory, Journal of Educational Psychology 90 (1998) 312–320.

[15] F. Schmidt-Weigand, K. Scheiter, The role of spatial descriptions in learning from multimedia, Comput. Hum. Behav. 27 (2011) 22–28. URL: https://doi.org/10.1016/j.chb.2010.05.007. doi:10.1016/j.chb.2010.05.007.

[16] X. Chen, D. Meurers, CTAP: A web-based tool supporting automatic complexity analysis, in: D. Brunato, F. Dell'Orletta, G. Venturi, T. François, P. Blache (Eds.), Proceedings of the Workshop on Computational Linguistics for Linguistic Complexity, CL4LC@COLING 2016, Osaka, Japan, December 11, 2016, The COLING 2016 Organizing Committee, 2016, pp. 113–119. URL: https://www.aclweb.org/anthology/W16-4113/.

[17] M. Ziefle, Effects of display resolution on visual performance, Hum. Factors 40 (1998) 554–568. URL: https://doi.org/10.1518/001872098779649355. doi:10.1518/001872098779649355.

[18] M. Brysbaert, M. Buchmeier, M. Conrad, A. Jacobs, J. Bölte, A. Böhl, The word frequency effect: a review of recent developments and implications for the choice of frequency estimates in german., Experimental psychology 58 5 (2011) 412–24.

[19] E. L. Aiden, J. Michel, Culturomics: Quantitative analysis of culture using millions of digitized books, in: 6th Annual International Conference of the Alliance of Digital Humanities Organizations, DH 2011, Stanford, CA, USA, June 19-22, 2011, Conference Abstracts, Stanford University Library, 2011, p. 8. URL: http://xtf-prod.stanford.edu/xtf/view?docId=tei/ab-003.xml.

[20] R. Lavalley, K. Berkling, S. Stüker, Preparing children's writing database for automated processing, in: K. M. Berkling (Ed.), Language Teaching, Learning and Technology, Satellite Workshop of SLaTE-2015, LTLT@SLaTE 2015, Leipzig, Germany, September 4, 2015, ISCA, 2015, pp. 9–15. URL: http://www.isca-speech.org/archive/ltlt_2015/lt15_009.html.

[21] T. Dönicke, Clause-level tense, mood, voice and modality tagging for german, in: K. Evang, L. Kallmeyer, R. Ehren, S. Petitjean, E. Seyffarth, D. Seddah (Eds.), Proceedings of the 19th International Workshop on Treebanks and Linguistic Theories, TLT 2020, Düsseldorf, Germany, October 27-28, 2020, Association for Computational Linguistics, 2020, pp. 1–17. URL: https://doi.org/10.18653/v1/2020.tlt-1.1. doi:10.18653/v1/2020.tlt-1.1.

[22] E. Breindl, A. Volodina, U. H. Waßner, Handbuch der deutschen Konnektoren 2, De Gruyter, 2014. URL: https://doi.org/10.1515/9783110341447. doi:doi:10.1515/9783110341447.

[23] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, Scikit-learn: Machine learning in Python, Journal of Machine Learning Research 12 (2011) 2825–2830.

[24] T. Akiba, S. Sano, T. Yanase, T. Ohta, M. Koyama, Optuna: A next-generation hyperparameter optimization framework, in: A. Teredesai, V. Kumar, Y. Li, R. Rosales, E. Terzi, G. Karypis (Eds.), Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2019, Anchorage, AK, USA, August 4-8, 2019, ACM, 2019, pp. 2623–2631. URL: https://doi.org/10.1145/3292500.3330701. doi:10.1145/3292500.3330701.

[25] L. Breiman, Random forests, Mach. Learn. 45 (2001) 5–32. URL: https://doi.org/10.1023/A:1010933404324. doi:10.1023/A:1010933404324.

[26] Y. Freund, R. E. Schapire, A decision-theoretic generalization of on-line learning and an application to boosting, J. Comput. Syst. Sci. 55 (1997) 119–139. URL: https://doi.org/10.1006/jcss.1997.1504. doi:10.1006/jcss.1997.1504.

[27] L. Breiman, J. H. Friedman, R. A. Olshen, C. J. Stone, Classification and Regression Trees, Wadsworth, 1984.

[28] E. Fix, J. L. Hodges, Discriminatory analysis - nonparametric discrimination: Consistency properties,

International Statistical Review 57 (1989) 238.
[29] F. Rosenblatt, Principles of neurodynamics. perceptrons and the theory of brain mechanisms, American Journal of Psychology 76 (1963) 705.
[30] C. Cortes, V. Vapnik, Support-vector networks, Mach. Learn. 20 (1995) 273–297. URL: https://doi.org/10.1007/BF00994018. doi:10.1007/BF00994018.