Domain-Specific Modeling of User Knowledge in Informational Search Sessions

Rui Tang[1], Ran Yu[2], Markus Rokicki[3], Ralph Ewerth[3] and Stefan Dietze[4,5]

[1]Ping An Technology, China
[2]Data Science & Intelligent Systems Group, University of Bonn, Germany
[3]L3S Research Center, Leibniz Universiy Hannover, Germany
[4]GESIS - Leibniz Institute for the Social Sciences, Germany
[5]Heinrich Heine University Dusseldorf, Germany

Abstract

Users frequently search on the Web to fulfill information needs with learning intent. In this context, usefulness of the search results depends strongly on the knowledge state of the user. In order to satisfy learning needs effectively, it is necessary to take users' knowledge gain and knowledge state within learning-oriented Web search sessions into account. Previous works studied the use of supervised models to predict a user's knowledge gain and knowledge state. However, the impact of knowledge domains of the search topics on a user's learning process have not been adequately explored. In this paper, we suggest domain detection techniques for search sessions and build domain-specific knowledge prediction models accordingly. Experimental evaluation results demonstrate that our approach outperforms the state-of-the-art baseline.

Keywords

search as learning, knowledge gain, informational search

1. Introduction

Users frequently surf the Web to search for a variety of information and to satisfy a wide range of information needs. Web search sessions are commonly categorized into three classes: navigational, informational and transactional [1]. Informational search sessions involve an inherent learning intent, i.e. the desire of a user to acquire knowledge or information with respect to a particular topic, assumed to be present on one or more Web pages. In this context, the individual relevance of search results is strongly dependent on the current knowledge state of the corresponding user.

The importance of learning scopes has been recognized by recent work at the intersection of information retrieval and learning theory. Eickhoff et al. [2] investigated the relationship between query and Web search session-related metrics and learning progress. Collins-Thompson et al. [3] studied the effectiveness of user interaction with respect to certain learning outcomes. The correlation between Web search behaviors and a user's learning gain has been explored by prior work [4], while the importance of learning as an implicit element of Web search has been established. Using various features computed based on user interactions and Web resource content, Yu et al. [5, 6] proposed approaches and built models for the prediction of a user's knowledge gain (KG) and knowledge state (KS). Their work demonstrates that knowledge gain and state of users can be predicted from their behaviors in Web search sessions.

Through more in-depth analysis of the relation between user knowledge state and various features based on user study data published by [5], we observed that correlations between features and knowledge gain/state in different knowledge domains of Web search sessions are different. For example, the correlation between the ratio of words related to the concept of *health* in user browsed webpages and knowledge gain/state for search sessions on topics in the *health* domain, is stronger than the correlation between them in sessions on topics in the *history* domain. Similar observations have been reported by Yu et al. in [6], where they proposed a new feature selection method to remove domain dependent features and thereby improve the topic generalizability of the knowledge prediction models. However, we argue that, instead of eliminating such features, we could use them to build fine-grained domain-specific models.

In this paper, we detect the most relevant domain of a search session based on textual information extracted from queries and webpages accessed by the user. We then carry out feature selection and build prediction models for each domain. Experimental results demonstrate that our proposed model outperforms the state-of-the-art baseline.

Proceedings of the CIKM 2021 Workshops, November 1–5, Gold Coast, Queensland, Australia
tangrui213@pingan.com.cn (R. Tang); ran.yu@uni-bonn.de (R. Yu); rokicki@L3S.de (M. Rokicki); ralph.ewerth@tib.eu (R. Ewerth); stefan.dietze@gesis.org (S. Dietze)
0000-0002-1153-4898 (R. Tang); 0000-0002-1619-3164 (R. Yu); 0000-0003-0918-6297 (R. Ewerth)

2. Related Works

Many studies have been carried out for understanding the relationship between learning progress and observable features in a search session. By matching the learning tasks into different learning stages of Anderson and Krathwohl's taxonomy [7], Jansen et al. studied the correlation between search behaviors of 72 participants and their learning stage [8]. They showed that information searching is a learning process with unique searching characteristics corresponding to particular learning levels. Cole et al. [9] observed that behavioral patterns provide reliable indicators about the domain knowledge of a user, even if the actual content or topics of queries and documents are disregarded entirely. Collins-Thompson et al. [3] studied the influence of distinct query types on knowledge gain, finding that intrinsically diverse queries lead to increased knowledge gain. Moraes et al.'s [10] work compared the learning outcome of instructor designed learning videos against three instances of search ("single-user", "search as support tool", "collaborative search") in order to find the most efficient approach for their learning scenario. Vakkari [11] provided a structured survey of features indicating learning needs as well as user knowledge and knowledge gain throughout the search process. Gadiraju et al. [4] described the use of knowledge tests to calibrate the knowledge of users before and after their search sessions, quantifying their knowledge gain, and investigated the impact of search intent and search behavior on knowledge gain of users. Bhattacharya et al. [12] investigated the relationship between users' search and eye gaze behaviors and their learning performance. In a recent work, Roy et al. [13] investigated at which time during a search session learning occurred, and found that the learning curve is largely influenced by a user's prior knowledge on the searched topic. Kalyani et al. [14] explored this direction further by designing search tasks that fit into the different learning stages of the revised Bloom's taxonomy. Through knowledge tests before and after each search session, they found significant impact of the learning stage on a user's search behavior and knowledge gain.

For predicting user's knowledge state or change in a search session, Zhang et al. [15] explored using search behavior as an indicator for the domain knowledge level of a user. Through a small study ($n = 35$), they identified features such as the average query length or the rank of documents consumed from the search results as being predictive. Syed and Collins-Thompson [16] explored the possibility of using regression models and features extracted from user accessed document content to predict user knowledge change on vocabulary learning tasks [17]. Gwizdka et al. [18] proposed to assess learning outcomes to search environments by correlating individual search behaviors with corresponding eye-tracking measures. Yu et al. [5, 6] proposed to use features based on user interactions and Web resource content to build classification models to predict user knowledge state and knowledge gain in search sessions. Liu et al. [19] adopted mind maps to capture user's knowledge change process and hence identified four types of knowledge change styles.

Although previous works have studied the relation between various features and user knowledge state, and knowledge prediction models have been proposed, the impact of the knowledge domain on the effectiveness of features hasn't been explored. In this paper, we propose a novel approach for predicting user knowledge state and knowledge gain in informational search sessions by taking the knowledge domain into consideration.

3. Task Description & Approach Overview

As defined in [5]: an intentional learning-related *search session* comprises the sequence of a user's actions with respect to satisfying her learning intent in a Web search environment through informational queries. A user's sequence of actions begins with an initial Web query and includes browsing through the search results, click and scroll activity, navigation via hyperlinks, query reformulations, and so forth. We refer to such an intentional learning-related search session as "session" in the remainder of this paper for simplicity.

Let s be a search session starting at time t_i and ending at time t_j aimed at satisfying a particular information need, that is, a learning intent ι of user u. In this work, we study the knowledge indicators (KIs): pre-knowledge state (pre-KS) $k(t_i)$, post-knowledge state (post-KS) $k(t_j)$ and knowledge gain (KG) $\Delta k(t_i, t_j)$ during time period $[t_i, t_j]$. This work aims at building **domain-specific** models (with respect to users' learning intents), to predict the KIs.

Figure 1 gives an overview of the approach we propose for building domain-specific KI prediction models. Given a session, we first extract textual information from different fields (e.g. query terms, webpage contents, etc.) and use it to detect the relevant domain of the session. After domain detection, the sessions are assigned to their most relevant domains. In the next step, we conduct the feature selection and knowledge modeling process using sessions assigned to each domain. More specifically, we compute Web resource features and user behavior features of each session, and then select a subset of these features based on two feature selection strategies. Using the selected features, we build KI prediction models for each domain. The process labeled in blue in Figure 1 shows an example of the data flow when predicting KI for a new session using the trained models.

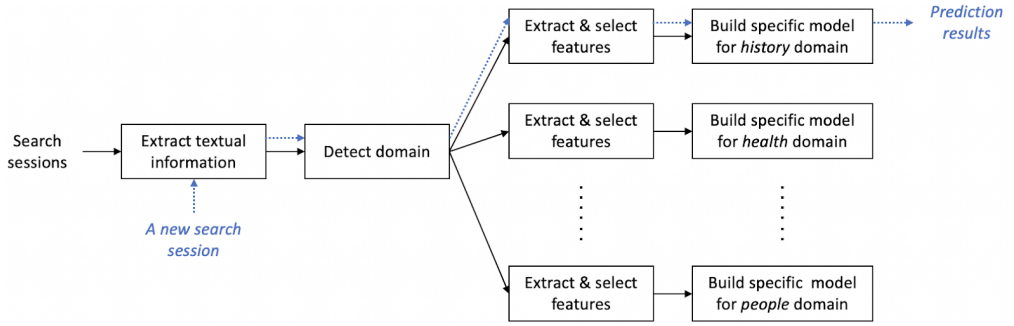


Figure 1: Overview of the modeling (in black) and example of the prediction (in blue) process.

We conclude the three main tasks of building domain-specific KI prediction models as follows:

1. **Domain detection of informational search sessions.** Each session s can be associated with one or more domains to a different extent. For the modeling purpose, we assign each session to a single domain that it has the strongest association with based on textual information involved in the session. As each session contains textual information in multiple fields, it is also our task to find the most suitable fields to be used for the domain detection.

2. **Feature extraction and domain-specific feature selection.** In this step, we first extract a set of features for each session s from the user behaviors and the related Web resource contents. For the sessions assigned to a specific domain, we select features reflecting the users' knowledge gain and state.

3. **Domain-specific knowledge modeling.** We formulate the prediction of knowledge state/gain as classification tasks, i.e. we aim to classify a specific KI (e.g. knowledge gain) of the user corresponding to a search session into low, moderate, high classes, with respect to a particular information need. That is, for each domain, we conduct feature selection and train classifiers to build the prediction models.

4. Dataset

To address the aforementioned tasks, we adopt an existing dataset which has been used by previous works on understanding and predicting user knowledge state and gain [4, 5]. This dataset includes search sessions conducted by crowd workers spanning across 11 information needs for different topics randomly selected from the *TREC 2014 Web Track*[1] dataset. This includes knowledge assessment data before and after each of the search sessions per information need, they also crawled the webpages that were assessed by the users. The experimental setup for obtaining the data and KIs was described by the authors in [6].

Data Cleaning. We filtered out untrustworthy workers who meet any of the following conditions: 1) did not complete the post-session test, 2) did not issue at least 1 search query, 3) selected the same option; either 'YES', 'NO' or for all items in the calibration test or the post-session test. In the next step, we filter out sessions that are insufficient of computing features we need for building knowledge prediction models, that includes: 1) sessions with no click on any results on the SERPs, and 2) sessions that contain at least 1 non-English resource browsed by the user. After applying all the aforementioned filters, we retain 233 search sessions, with 1.361 queries and 2.622 clicks per session on average.

Knowledge Measures. Knowledge tests are scientifically formulated tests that measure the knowledge of a participant on a given topic. The authors of [4] created knowledge tests pertaining to each of the information needs. The pre (post)-knowledge score of a user in search sessions corresponding to a topic is measured as the percentage of the correct answers on the knowledge test that a given user has completed. Correspondingly, the knowledge gain is measured as the difference between a user's pre- and post-search session knowledge score.

For the classification tasks described in Section 3, we follow the same approach as used in [5], i.e. a *Standard Deviation Classification* approach to obtain three classes of learners with regard to their level of pre-KS. Assuming approximately normal distributions of the respective test scores (X) for the different topics, we transformed the test scores into Z-scores with a mean of 0 and a Standard Deviation (SD) of 1 (standardization). We then used sta-

[1] http://www.trec.nist.gov/act_part/tracks/web/web2014.topics.txt

tistically defined intervals (low: X < -0.5 SD; moderate: -0.5 SD < X < 0.5 SD; high: 0.5 SD < X) for the classification of the learners into roughly equal groups with low, moderate, or high pre-KS. The same procedure was repeated for post-KS and KG. Table 1 shows the resulting numbers of learners for the respective classes and underlying statistics.

Table 1
Knowledge state and knowledge gain classes created based on thresholds of *mean* ± *0.5SD*.

Task	Mean	SD	Low	Moderate	High
pre-knowledge state	0.36	0.255	87	52	94
post-knowledge state	0.66	0.174	61	95	77
knowledge gain	0.23	0.208	84	84	65

5. Domain Detection

The goal of this step is to assign each informational search session to a most relevant domain. More specifically, we extract textual information from queries and consumed Web resources of a session and apply two text classifiers on them to detect its domain.

5.1. Methods for Domain Detection

Domain detection in this paper is formulated as a text classification problem ("to which predefined class or category is this text most likely to belong?" [2]). This work aims at exploring the possibility of improving KI prediction performance by building more focused models, rather than developing novel domain detection techniques. We therefore utilize two existing domain detection tools, namely *TagTheWeb* and *uClassify*.

TagTheWeb [20] can automatically categorize a given text into Wikipedia categories with a probability. The category with the highest probability is considered to be the most relevant domain. The 19 top level Wikipedia categories adopted by *TagTheWeb* are: *arts*, *culture*, *games*, *geography*, *health*, *history*, *humanities*, *industry*, *law*, *life*, *mathematics*, *matter*, *nature*, *people*, *philosophy*, *reference works*, *religion*, *science and technology* and *society*. Moreover, *TagTheWeb* could also classify text into Wikipedia sub-categories, however, in this work, we focus only on the 19 top-level categories as the granularity fits better into the task scenario and the size of experimental dataset.

uClassify[3] is a free machine learning Web service that provides classifiers for different applications. A classifier called *Topics* from *uClassify* can classify a given textual document into 10 different top-level domains. Each domain has a score of probability, and the domain with the highest probability is considered as the most relevant domain. The 10 top-level domains we adopted in this work are *arts*, *business*, *computers*, *games*, *health*, *home*, *recreation*, *science*, *society* and *sports*. The classes are adopted from the Open Directory Project [4].

5.2. Textual Information Extraction

Table 2
Domain detection configurations and abbreviations based on extracted textual information.

Abbreviation	Description
QW	Query words
WPT	Web page titles
WPC	Web page contents
QW & WPT	Query words and Web page titles
QW & WPC	Query words and Web page contents
WPT & WPC	Web page titles and Web page contents
QW & WPT & WPC	Query words, Web page titles, and Web page contents
all MV	Mjority vote based on QW, WPT and WPC result

During a session, a user enters query terms to communicate her information need on a topic related to certain domains, we extract and combine all query terms in a session and use it for domain detection. Titles of visited Web pages can be an indicator of the domain that a user choose to learn in a session. Therefore, we combined the titles of all the visited Web pages as the second source of textual information. Besides titles, we also analyze their content by combining all textual content of visited Web pages in a session. This result in three types of textual information: query words (QW), title of the visited Web pages (WPT) and textual contents of the visited Web pages(WPC). Moreover, we consider all the five combinations of these sources (as listed in Table 2) and a majority vote strategy based on results of using the three textual sources respectively (*all MV*). For the *all MV* strategy, when all three votes are different from each other, we assign the session to *other* domain.

5.3. Evaluation

We apply both text classification tools for all 8 configurations (Table 2) respectively. In this section, we present the evaluation results of domain detection, and choose the configuration that the next step relies on accordingly.

Ground Truth. Two authors of this paper manually assigned labels to the sessions according to the corresponding topics that were presented to the crowd workers when creating the dataset. As sessions corresponding

[2] https://www.uclassify.com/docs/intro
[3] https://www.uclassify.com/browse/uclassify/topics
[4] http://www.dmoz.org

Table 3
Ground truth labels used for the domain detection evaluation.

Topic Description	Labeled Domain	
	TagTheWeb	***uClassify***
Altitude Sickness	*health*	*health*
American Revolutionary War	*history*	*society*
Carpenter Bees	*nature*	*science*
USS *Cole* Bombing	*history*	*society*
Evolution	*nature, life, philosophy*	*science*
HIV	*health*	*health*
NASA Interplanetary Missions	*history, science and technology*	*science*
Orcas Island	*geography, history, nature*	*society, science*
Sangre de Cristo Mountains	*geography, nature, history*	*society, science*
Sun Tzu	*people, history, culture*	*arts, recreation, society*
Tornados	*nature*	*science*

to the same topic could have different domain focus, we decided to allow multiple correct domain labels when building the ground truth. Consequently, in the following evaluation, a domain classification outcome was treated as correct, if the predicted domain was among the assigned labels. The description of the pre-defined search topics and the domain labels assigned to them are shown in Table 3. The annotators agreed on all labels.

Evaluation Results. For each of the 16 configurations (2 classifiers X 8 textual information combinations), we compute the overall accuracy of the classification result. Based on the results shown in Table 4, we found that all accuracy scores are above 0.550 for *TagTheWeb*. The best performance of *TagTheWeb* is achieved when combining query words and Web resource titles (*QW & WPT*), as well as when combining all three fields (*QW & WPT & WPC*), 174 of 233 sessions are detected correctly (accuracy = 0.747). We choose the configuration *QW & WPT* for later steps, as it has higher efficiency compared to *QW & WPT & WPC*. Meanwhile, all accuracy scores of *uClassify* are below 0.25. Therefore, we decide not to pass the result of *uClassify* to later steps.

To better illustrate the domain detection result, we present a heatmap in Figure 2 showing the assignment of sessions corresponding to each topic to the target domains by *TagTheWeb* using *QW&WPT*. We found that 81.5% of sessions in our GT are assigned to 5 domains, namely, *history* (56 sessions), *health* (49 sessions), *nature* (32 sessions), *geography* (29 sessions) and *people* (24 session). As the next modeling steps require sufficient

Table 4
Evaluation of Domain Detection Results

Textual Information	Accuracy	
	TagTheWeb	***uClassify***
QW	0.665	0.224
WPT	0.605	0.236
WPC	0.592	0.236
QW & WPT	**0.747**	0.232
QW & WPC	0.712	0.236
WPT & WPC	0.665	0.236
QW & WPT & WPC	**0.747**	0.236
all MV	0.554	0.236

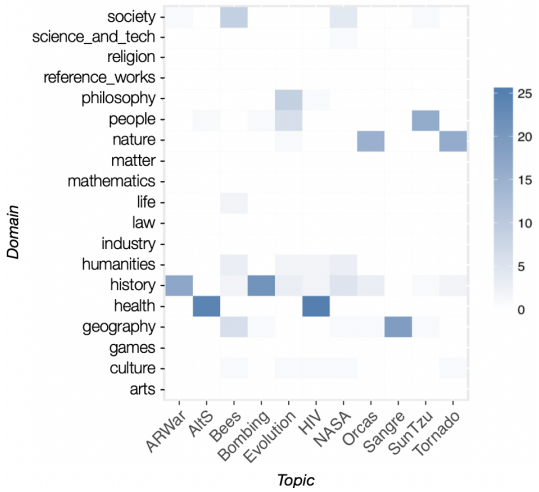


Figure 2: Heatmap of domain detection results by *TagTheWeb* X *QW&WPT*. The x-axis indicates the topics in the user study, and the y-axis represents the domains which sessions are assigned to.

amount of training data in order to build reliable models, we continue the experiment with the 190 sessions categorized into these 5 most frequent domains, and discard the rest 43 sessions which are categorized into *society* (15 sessions), *humanities* (10 sessions), *philosophy* (10 sessions), *culture* (5 sessions), *life* (2 sessions) or *science and technology* (1 sessions).

6. Modeling User Knowledge

6.1. Approach

6.1.1. Model

As described in Section 3, we follow the same approach as in [5, 6] and cast the problem of predicting user KIs

as classification tasks. More specifically, each session s is represented as a feature vector, $\vec{v} = (f_1, f_2, ..., f_n)$, where the features considered are introduced later in this section. We apply a range of standard classification models, namely, Naive Bayes (*nb*), Logistic Regression (*lr*), Support Vector Machine (*svm*) and Random Forest (*rf*). For our experiments, we used the *scikit-learn* library for Python[5]. We tune hyperparameters of the algorithms using grid search.

6.1.2. Feature Extraction

As the focus of this work is to explore the performance of domain-specific knowledge prediction models, we make use of the same set of features as described in [6]. The features consist of two categories according to the data source: *Web resource* features and *user behavior* features. The 109 Web resource features are extracted based on the content of the webpages which users visited during a session, including features computed based on document complexity (e.g. average number of words per sentence, Gunning Fog Grade[6]), HTML structure (e.g. Number of<script>elements) and linguistic characteristics (based on the 2015 LIWC dictionaries[7]) of the Web resource content. The 66 user behavior features are extracted from the user interaction with the search engine during a session, namely features related to the session (e.g. session duration), queries (e.g. average query length), SERP (e.g. the lowest rank of click), browsing behavior (e.g. ratio of revisited pages) and mouse movements (e.g. total scroll distance). As the features have been introduced and investigated in details by previous works [6], we will not go into details in this paper.

6.1.3. Metrics for Feature Selection

Due to the difficulty in obtaining ground truth data with user knowledge assessment, the scale of training and testing data is limited. Hence, feature selection is important for building reliable models, and in particular, to avoid overfitting. For sessions assigned to each domain, our goal is to select a set of features $F\prime \subseteq F$ that produce the most reliable model for the KI prediction tasks. We introduce 2 metrics that are adapted from previous work [5].

Ensure feature effectiveness. We compute the Pearson correlation coefficient between each feature f_i and KI, i.e. $Corr(f_i, KI)$, across all sessions in a specific domain. To ensure effectiveness of features, we select features fulfilling the condition $|Corr(f_i, KI)| \geq \alpha$ for building the classification models.

[5]http://scikit-learn.org
[6]http://gunning-fog-index.com/
[7]http://liwc.wpengine.com/

Reduce feature Redundancy. We also compute the Pearson correlation coefficient $Corr(f_i, f_j)$ between each pair of features across all sessions in a specific domain. If $|Corr(f_i, f_j)| \geq \beta$, i.e. features are highly similar to each other, we remove the one which has a lower $Corr(f_i, KI)$ from the pair.

6.2. Evaluation

The generation of class labels of the sessions in our experimental dataset is described in Section 4. We evaluate model performances by means of 10-fold cross-validation. Further, classification performance is measured in terms of the following metrics:

- Accuracy (Accu): percentage of search sessions that were classified with the correct class label.
- Precision (P), Recall (R), F1 (F1) score of class i: the standard precision, recall and F1 score on the prediction result of each class i.
- Macro average of P, R and F1: the average of the corresponding score across 3 classes.

Baselines. We compare our approach against [5], who proposed to build classifiers to predict KG and post-KS using user interaction and session features only. Their approach considered feature selection based on the feature-KI-correlation (α) and the between-feature-correlation (β). Using their approach, we make use of all the 190 sessions which are relevant to the aforementioned 5 domains (*history*, *health*, *nature*, *geography* and *people*) to build classifiers for the knowledge prediction tasks. We also compare our approach against an improved baseline (denoted as *baseline'*) for which we apply these 190 sessions to build non-domain-specific classifiers using both user interaction features and Web resource features. In the experiment, we tuned the hyper-parameters of these models again using grid search to ensure a fair comparison.

6.2.1. Overall Performance

Using our approach, the overall accuracy scores are above 0.610 for all 3 prediction tasks and the overall average F1 scores are above 0.609 (see Table 5). Compared to the state-of-the-art baseline (*baseline*), we observed improvements for all 3 prediction tasks, with the improvements by 18.1%, 13.6% and 17.1% (average F1 score) as well as 16.3%, 12.2% and 15.8% (accuracy score) for pre-KS, post-KS and KG prediction tasks respectively.

Our approach and *baseline'* make use of the same feature set which includes user behavior features and Web resource features. Our models outperform *baseline'* by 14.5%, 10.4% and 12.7% (average F1 score) as well as 12.1%, 9.5% and 12.1% (accuracy score) in the tasks of pre-KS, post-KS and KG prediction respectively. This demonstrates that our domain-specific knowledge modeling ap-

Table 5
Best performing results of different approaches according to average F1 score. Columns **#l, #m**, and **#h** present number of sessions with predicted label *low*, *moderate* and *high* respectively.

					low			*moderate*			*high*			average			
KI	**#l**	**#m**	**#h**	**Approach**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**Accu**
pre-KS	68	43	79	new	**0.727**	**0.824**	**0.772**	**0.532**	**0.581**	**0.556**	**0.773**	**0.646**	**0.703**	**0.677**	**0.683**	**0.677**	**0.695**
				baseline	0.529	0.662	0.588	0.324	0.279	0.300	0.647	0.557	0.599	0.500	0.499	0.496	0.532
				baseline'	0.654	0.750	0.699	0.308	0.279	0.293	0.630	0.582	0.605	0.531	0.537	0.532	0.574
post-KS	47	83	60	new	**0.583**	**0.596**	**0.589**	**0.644**	0.566	**0.603**	**0.594**	**0.683**	**0.636**	**0.607**	**0.615**	**0.609**	**0.611**
				baseline	0.487	0.404	0.442	0.481	0.614	0.540	0.511	0.383	0.438	0.493	0.467	0.473	0.489
				baseline'	0.525	0.447	0.483	0.486	**0.627**	0.547	0.581	0.417	0.485	0.531	0.497	0.505	0.516
KG	67	75	48	new	**0.710**	**0.657**	**0.682**	**0.590**	**0.653**	**0.620**	**0.689**	**0.646**	**0.667**	**0.663**	**0.652**	**0.656**	**0.653**
				baseline	0.543	0.567	0.555	0.453	0.520	0.484	0.500	0.354	0.415	0.499	0.480	0.485	0.495
				baseline'	0.544	0.552	0.548	0.513	0.547	0.529	0.548	0.479	0.511	0.535	0.526	0.529	0.532

Table 6
Knowledge gain (KG) prediction results by domain. Models were selected according to average F1 score. ***Geog*** = ***Geography***, **clf** = selected classifier, **#s.f** = number of selected features, α = *feature effectiveness* threshold, β = *feature redundancy* threshold, **#d.s** = number of sessions in a domain, {**#l,#m,#h**} = number of sessions with predicted label {*low,moderate,high*} respectively.

									low			*moderate*			*high*			average			
Domain	**clf**	**#s.f**	α	β	**#d.s**	**#l**	**#m**	**#h**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**Accu**
History	*rf*	2	0.4	0.2	56	9	28	19	0.375	0.333	0.353	0.679	0.679	0.679	0.600	0.632	0.615	0.551	0.548	0.549	0.607
Health	*rf*	3	0	0.3	49	31	14	4	0.852	0.742	0.793	0.500	0.714	0.588	0.500	0.250	0.333	0.617	0.569	0.572	**0.694**
Nature	*lr*	9	0.3	0.5	32	9	9	14	0.714	0.556	0.625	0.400	0.444	0.421	0.733	0.786	0.759	0.616	0.595	0.602	0.625
Geog	*nb*	19	0.2	0.7	29	10	13	6	0.667	0.600	0.632	0.625	0.769	0.690	1.000	0.667	0.800	0.764	0.679	**0.707**	0.690
People	*rf*	9	0.4	1	24	8	11	5	0.636	0.875	0.737	0.667	0.545	0.600	0.750	0.600	0.667	0.684	0.673	0.668	0.667

proach can improve the performance of the knowledge prediction tasks.

6.2.2. Domain Analysis

To give more insights on the performance of models built for different domains, we present the evaluation results for knowledge gain (KG) prediction in Table 6. For KG prediction, the accuracy scores are at least 0.607 and the F1 scores are at least 0.549. The best accuracy score (0.694) is achieved by the model focused on the *health* domain. Only 3 features, i.e. *average number of authentic words*[8] *in visited web resources*, *avg time to first click* and *average click interval*, are used to train the model. The best F1 score (0.707) is achieved by the *geography*-specific model, when using 19 features. Only 2 features, i.e. l_shehe_avg and c_char_avg, are used for the *history*-specific model.

6.2.3. Impact of Features

To better understand the correlation between features and KIs in a specific domain, we evaluated the usefulness of feature selection strategies based on *feature effectiveness* ($|Corr(f_i, \text{KI})| \geq \alpha$) and *feature redundancy* ($|Corr(f_i, f_j)| \geq \beta$), i.e. we assess the impact of feature selection thresholds on domain focused model performance. The results are shown in Figure 3. We chose the highest average F1 score and the corresponding accuracy score for each feature selection configuration in different classifiers. We illustrate with the result on the domain with most number of sessions in our dataset. i.e. history.

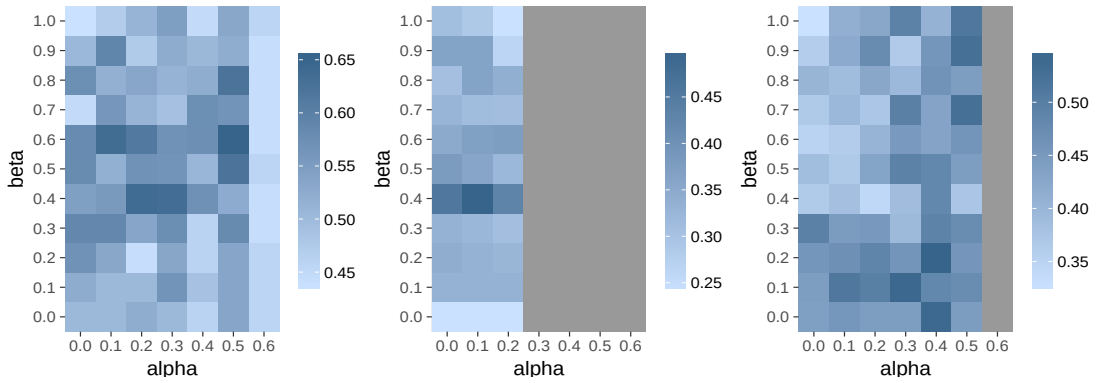
In the pre-KS prediction task of *history* domain, for a relatively restrictive *feature redundancy* setting ($\beta = 0.4$), performance is maximized for $\alpha = 0.2$, whereas less restrictive requirements on *feature effectiveness* (a lower α) and *feature redundancy* (a higher β) resulting in lower performance in terms of both F1 and accuracy score. The best average F1 score is achieved with the combination with a high *feature effective* threshold of $\alpha = 0.5$ and a slightly restrictive *feature redundancy* threshold of $\beta = 0.6$.

In terms of post-KS prediction in *history* domain, we observed that the F1 scores are higher than others when $\beta = 0.4$. The best F1 score is achieved when with a low $\alpha = 0.1$. The reason for this are the comparatively low correlations of features with post-KS in this domain.

For the KG prediction task in the *history* domain, we observed a general positive trend in prediction performance when increasing α. Moreover, relatively restrictive *feature redundancy* settings can further improve prediction performance. The best F1 score is achieved when $\alpha = 0.4$ and $\beta = 0.2$.

Overall, the approach of filtering features based on ef-

[8] https://liwc.wpengine.com/interpreting-liwc-output/: the algorithm for Authenticity was derived from a series of studies where people were induced to be honest or deceptive.



(a) average F1, pre-KS (b) average F1, post-KS (c) average F1, KG

Figure 3: Model performance of different feature selection configurations on the *history* domain. The x-axis represents the threshold α (*feature effectiveness*) and the y-axis represents the threshold β (*feature redundancy*). The color of the cell represents the F1 score of the evaluation result . The darker a cell is, the higher the corresponding value is. The grey cells represent no performance because no features are selected when applying certain thresholds.

fectiveness and redundancy results in markedly improved performance. On the other hand, the overall restrictive settings – resulting in only two features used for KG prediction – highlight further room for improvement. While these general models worked best in our experiments, refining the domain detection step (e.g. using a more fine-grained taxonomy) could result in more coherent sets of training data, allowing for the use of more (specific) features.

7. Conclusions

In this paper, we investigated the influence of the domain on learning-oriented informational Web search sessions, and proposed to improve the performance of knowledge prediction models by extending them to several domain-specific models. We evaluated two text classifiers, i.e. *TagTheWeb* and *uClassify*, using 8 types of textual information respectively to categorize a session into a most relevant domain. We observed the best domain detection accuracy when using *TagTheWeb* based on query words and web page titles. Based on this, we built domain-specific models for knowledge prediction tasks. In our experiments, the approach outperformed the state-of-the-art baseline by at least 12.2% in terms of accuracy and at least 13.6% in terms of F1-Score. Thus, our work contributes to the understanding and prediction of user knowledge in learning-oriented informational Web search sessions.

Due to the limited availability of Web search session data as well as the corresponding user knowledge assessment data, there are limitations in our current experimental dataset. Therefore, observations made herein should be validated on a large scale dataset in future work. Further, in the domain detection step, only top level categories (domains) of the taxonomies were used when applying *TagTheWeb*. Given sufficient data, accuracy could be improved by adopting more subcategories, i.e. more specific domains. Moreover, other than the two exemplary solutions investigated in this work, other domain detection techniques could be applied as well.

Acknowledgments

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

References

[1] A. Z. Broder, A taxonomy of web search, SIGIR Forum 36 (2002) 3–10. URL: https://doi.org/10.1145/792550.792552. doi:10.1145/792550.792552.

[2] C. Eickhoff, J. Teevan, R. White, S. Dumais, Lessons from the journey: a query log analysis of within-session learning, in: Proceedings of the 7th ACM international conference on Web search and data mining, ACM, 2014, pp. 223–232.

[3] 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: Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval, ACM, 2016, pp. 163–172.

[4] U. Gadiraju, R. Yu, S. Dietze, P. Holtz, Analyzing knowledge gain of users in informational search

sessions on the web, in: 2018 ACM on Conference on Human Information Interaction and Retrieval (CHIIR), ACM, 2018.

[5] R. Yu, U. Gadiraju, P. Holtz, M. Rokicki, P. Kemkes, S. Dietze, Predicting user knowledge gain in informational search sessions, in: Proceedings of the 41st International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, 2018.

[6] R. Yu, R. Tang, M. Rokicki, U. Gadiraju, S. Dietze, Topic-independent modeling of user knowledge in informational search sessions, Information Retrieval Journal 24 (2021) 240–268.

[7] L. W. Anderson, D. R. Krathwohl, P. Airasian, K. Cruikshank, R. Mayer, P. Pintrich, J. Raths, M. Wittrock, A taxonomy for learning, teaching and assessing: A revision of bloom's taxonomy, New York. Longman Publishing. Artz, AF, & Armour-Thomas, E.(1992). Development of a cognitive-metacognitive framework for protocol analysis of mathematical problem solving in small groups. Cognition and Instruction 9 (2001) 137–175.

[8] B. J. Jansen, D. Booth, B. Smith, Using the taxonomy of cognitive learning to model online searching, Information Processing & Management 45 (2009) 643–663.

[9] M. J. Cole, J. Gwizdka, C. Liu, N. J. Belkin, X. Zhang, Inferring user knowledge level from eye movement patterns, Information Processing & Management 49 (2013) 1075–1091.

[10] F. Moraes, S. R. Putra, C. Hauff, Contrasting search as a learning activity with instructor-designed learning, in: A. Cuzzocrea, J. Allan, N. W. Paton, D. Srivastava, R. Agrawal, A. Z. Broder, M. J. Zaki, K. S. Candan, A. Labrinidis, A. Schuster, H. Wang (Eds.), Proceedings of the 27th ACM International Conference on Information and Knowledge Management, CIKM 2018, Torino, Italy, October 22-26, 2018, ACM, 2018, pp. 167–176. URL: https://doi.org/10.1145/3269206.3271676. doi:10.1145/3269206.3271676.

[11] P. Vakkari, Searching as learning: A systematization based on literature, Journal of Information Science 42 (2016) 7–18.

[12] N. Bhattacharya, J. Gwizdka, Measuring learning during search: Differences in interactions, eye-gaze, and semantic similarity to expert knowledge, in: Proceedings of the 2019 Conference on Human Information Interaction and Retrieval, ACM, 2019, pp. 63–71.

[13] N. Roy, F. Moraes, C. Hauff, Exploring users' learning gains within search sessions, in: Proceedings of the 2020 Conference on Human Information Interaction and Retrieval, 2020, pp. 432–436.

[14] R. Kalyani, U. Gadiraju, Understanding user search behavior across varying cognitive levels, in: Proceedings of the 30th ACM Conference on Hypertext and Social Media, 2019, pp. 123–132.

[15] X. Zhang, M. Cole, N. Belkin, Predicting users' domain knowledge from search behaviors, in: Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval, ACM, 2011, pp. 1225–1226.

[16] R. Syed, K. Collins-Thompson, Retrieval algorithms optimized for human learning, in: Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, ACM, 2017, pp. 555–564.

[17] R. Syed, K. Collins-Thompson, Exploring document retrieval features associated with improved short- and long-term vocabulary learning outcomes, in: Proceedings of the 2018 Conference on Human Information Interaction&Retrieval, ACM, 2018, pp. 191–200.

[18] J. Gwizdka, X. Chen, Towards observable indicators of learning on search., in: SAL@ SIGIR, 2016.

[19] H. Liu, C. Liu, N. J. Belkin, Investigation of users' knowledge change process in learning-related search tasks, Proceedings of the Association for Information Science and Technology 56 (2019) 166–175.

[20] J. F. Medeiros, B. P. Nunes, S. W. M. Siqueira, L. A. P. P. Leme, Tagtheweb: Using wikipedia categories to automatically categorize resources on the web, in: European Semantic Web Conference, Springer, 2018, pp. 153–157.