Supporting Children's Web Search in School Environments

Carsten Eickhoff
Delft University of Technology
Delft, Netherlands
c.eickhoff@tudelft.nl

Pieter Dekker
Delft University of Technology
Delft, Netherlands
dekker.pieter@gmail.com

Arjen P. de Vries
Centrum Wiskunde & Informatica
Amsterdam, Netherlands
arjen@acm.org

ABSTRACT

Nowadays, the Internet represents a ubiquitous source of information and communication. Its central role in everyday life is reflected in the curricula of modern schools. Already in early grades, children are encouraged to search for information on-line. However, the way in which they interact with state-of-the-art search interfaces and how they explore and interpret the presented information, differs greatly from adult user behaviour.

This work describes a qualitative user study in which the Web search behaviour of Dutch elementary school children was observed and classified into roles motivated by prior research in cognitive science. Building on the findings of this survey, we propose an automatic method of identifying struggling searchers in order to enable teaching personnel to provide appropriate and targeted guidance where needed.

Categories and Subject Descriptors

H.5.4 [**Information Interfaces & Presentation**]: Hypertext / Hypermedia—*User Issues*; H.1.2 [**Models and Principles**]: User/Machine Systems —*Human Factors*

General Terms

Human Factors, Experimentation

Keywords

Children, Classification, Web Search, Search Success, Search Roles

1. INTRODUCTION

Over the last decade, children have been growing considerably more acquainted with technologies such as computers and the Internet. This trend can be observed across all age levels, starting with children as young as 3-5 years. As a consequence, the age of first contact with said technologies is decreasing while the overall time spent using Web-based information systems rises significantly. According to the EU Kids on-line report [25] 83% of European children in the age of 6 to 10 and 96% of the 11 to 14 year-olds regularly use the Internet. Similar figures are reported by national counterparts in countries such as the UK [29] or Germany [11].

This media-centric development is reflected in modern school curricula that encourage and support children, already at young ages, to use computers. In this way, schools aim to prepare children for the recent changes in skills demanded by both society and the labour market. More concretely, Web information search is commonly incorporated in the preparation phase of essay writing or the creation of classroom presentations.

Let us proceed to inspect this concrete scenario in more depth. When supervising and guiding classroom assignments involving Web information search, elementary school teachers assist groups of 20 - 25 children who are individually searching the web at the same time. A particular challenge lies in the fact that such groups tend to be heterogeneous in their information search capabilities. While some students are coping well with the task, others may struggle. Identifying those few students in a large group who need the most assistance at a given point in time, however, is not easy. Time that could have been invested into aiding struggling children may be wasted due to the problem of identifying them in the first place [34].

In this work, we propose a solution for this problem by devising an automatic scheme for determining search success based on a wide range of cognitive, information theoretic and empirical features of search sessions. We envision a teachers "dashboard" highlighting where help is needed most and perhaps even integration in electronic learning environments. The novel contributions of our work are three-fold: (1) We conduct a comparative study in a Dutch elementary school and provide detailed qualitative analyses of the findings. (2) Building on Allison Druin's work [13], we annotate the children's search sessions with the search roles that she observed among children's recreational web searching at home. (3) Based on a range of features grounded in literature, we design and evaluate an automatic classification scheme that enables us to estimate searcher roles and their likelihood of search success.

The remainder of this article is structured as follows: Section 2 gives an overview of related work in the areas of behavioural psychology, cognitive science, information retrieval and human computer interaction. Section 3 details the set-up of our elementary school user study. In Section 4, we explore the collected data by discussing a number of key

IIiX'12, August 21–24, 2012, Nijmegen, The Netherlands.

figures and statistics as well as the subsequent annotation process. Section 6 describes our search success classification scheme and its performance on previously unseen real-world data. Section 7 closes with concluding remarks and an outlook on future work in the domain of child-friendly information access and its potential for classroom use.

2. RELATED WORK

This section encompasses three major themes of related previous work relevant to our objective: (I) Formal models of information seeking. (II) Dedicated studies of children's web search behaviour. (III) General search success frameworks.

2.1 Information seeking models

In one of the early studies on human information seeking behaviour, Saracevic et al. [31] conducted a qualitative survey in which 40 participants were interviewed concerning their search habits and various manners of reasoning were tested for. In a different study, Kuhlthau revisits the information search process from a user perspective [19], arguing for the necessity of users understanding IR systems in order to interact with them faithfully. According to the author's proposed model, this hypothesis is confirmed in a user survey. In their survey article, Belkin and Croft [3] establish a close relationship between information retrieval and information filtering. The latter of the two is described to be necessarily user-aware in order to provide suitability of retrieved results given the user's context and information need. Marchionini [26] describes the information seeking process as constructed of 3 partially parallel phases. Each search starts with the user *understanding* her or his information need. Followed by *planning and executing* the query and a subsequent *evaluation and use* of the returned results. Choo et al. [10] incorporated the search session's motivation into the model, finding it to have clear influence on the observed search strategies such as undirected browsing or targeted searching. In 2005, Ingwersen and Järvelin [18] highlighted the importance of searcher context for models of information seeking. In a dedicated cognitive model, they establish various levels on which context crucially influences search processes.

2.2 Search success

A particular aspect of information seeking models recently established is the notion of search success. Several studies define search success as the rate of search sessions that result in satisfying the user's information need [35, 27]. Stronge et al. relate a user's search strategies to their likelihood of search success in web search [33]. Bilal investigated the search success rates of children using Yahoo!'s child-oriented platform Yahooligans [4]. This work will employ search success as a key surrogate for determining children's need for assistance with information search in school settings.

2.3 Children's Web search

Several studies have investigated the possibilities of creating models of children's information seeking behaviour. Shenton and Dixon [32] reviewed models that have been developed based on results from research among children in the age of 4 to 18. The grounded model of information seeking via the Internet consists of 11 different actions or influences. Before the start of the actual search process, a multi-step framework accounts for factors such as the origin of the information need, the directness of use or the place of information access. Dania Bilal [6] presented a study of Arabic children's interaction with a digital library system. Based on their search behaviour on the International Children's Digital Library (ICDL), she formulated an alternative model of children's information seeking. The ICDL is a web interface that introduces children to various cultures with books. Her model is centred around three key concepts:

Browsing. A child scans the list of book thumbnails and moves to the next page with thumbnails.

Backtracking. A child uses the back arrows of ICDL or the back button of their browser to return to an earlier stage of their search.

Navigating. A child uses the ICDL's functionality for page-internal navigation such as zooming in on particular page aspects.

A number of particular challenges are frequently reported to frustrate young searchers and prevent them from achieving search success. (A) Query formulation is difficult due to insufficiently-developed writing skills and small active vocabularies [28]. (B) Identifying relevant organic search results often overwhelms children as they struggle to judge which results will satisfy their information needs [7]. (C) The overall number of results presented by a typical web search engine puts a high cognitive load on children that often leads to confusion [21].

A line of related work central to this research is led by Allison Druin. In a pilot study, the authors investigate how children of an age between 7 and 11 years old search the Internet using keyword interfaces at home [12]. The study highlights a number of barriers that hinder children from successfully searching the Web using technologies designed for adult users. The particular challenges include spelling, typing, query formulation and deciphering results. In a subsequent qualitative home study among 83 US children, the authors established a searcher categorization [13]. After an initial interview, the children were encouraged to search for information using a web search engine of their choice. Qualitative analysis revealed a number of characteristics that motivated a framework of the following 7 searcher roles:

Developing searchers tend to use natural language queries, "asking" the search engine questions. They are able to complete simple queries, but have trouble with complex ones.

Domain-specific searchers limit searches to finding content specific to a domain of personal interest. They repeatedly return to a small number of specific websites which they have accepted as authoritative or informative.

Power searchers display advanced search skills and are able to use keywords instead of natural language in the query formulation step. They do not suffer from breakdowns and are able to solve more complex search assignments.

Non-motivated searchers are not persistent when searching. They lack motivation to find alternative problem solutions or query reformulations and easily give up after set-backs.

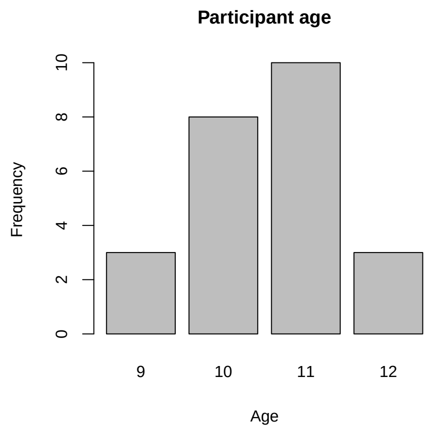

Figure 1: Age distribution of participants.

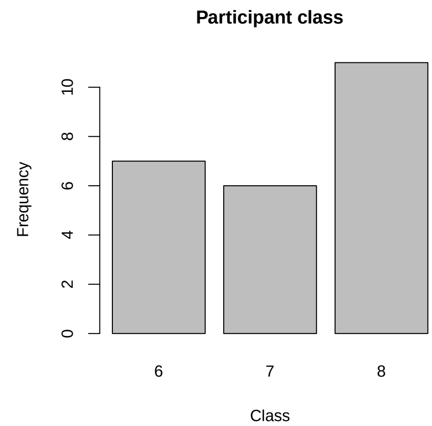

Figure 2: Class distribution of participants.

Distracted searchers have trouble staying focused on the current search task. They frequently side track into investigating other information needs and are easily distracted by external stimuli.

Visual searchers guide their information search along visual clues. They often start search sessions from image and video search engines, identifying visual concepts relevant to their assignment.

Rule-bound searchers follow a strict set of rules from which they are not easily able to deviate.

In this work, we start from Druin et al.'s established search roles and investigate their generalizability from home settings to a school scenario. In a first step, we collect search sessions in a Dutch elementary school and subsequently annotate them with role labels. In a following step, we automatically predict the role label based on low-level search session features.

Figure 3: Gender distribution of participants.

3. ELEMENTARY SCHOOL USER STUDY

In order to investigate children's search behaviour in a school setting, we conduct a user study with 29 children from different grade-levels at the Dutch elementary school "De Kroevendonk" in Roosendaal. The experiment was carried out during regular school hours with informed consent by the students' legal guardians and conforming to the Dutch and European Data Protection Acts [22, 23, 24]. To introduce the researcher as well as the research goal to the children and to make them comfortable with the experiment, we gave an explanatory presentation in all participating classes with the possibility for asking questions.

The sessions of 5 participants were collected in a pilot run used to refine the experiment set-up. This leaves 24 participants in the final data collection. Figures 1 - 5 show key statistics such as age, gender or class distribution of the participants. We could observe a generally high degree of computer skills, with the majority of participants reporting regular computer and Internet contacts. The age group of children ranging from 8-12 years of age has been previously confirmed to be interesting for Web search experiments [15] as children of this age have well-developed reading skills, while still displaying significantly different behaviour from adult searchers [1, 14].

3.1 Experimental set-up

To limit external distractions, experiments were conducted in a separate room in the school building and children were individually participating while their peers would continue with regular regular classroom activities. One researcher was always present during the experiment to take notes as well as to assist if the child had questions concerning the general experiment. The research did not interfere with or comment on the search processes. A browser-based survey system guided the participant through the experiment. After a brief introduction, we asked three initial questions about well-being and prior experience:

1. `How do you like participating in this study?`
2. `How often do you use a computer?`
3. `How often do you use the Internet?`

After this collection of personal background information, the actual search tasks started. We use three types of ques-

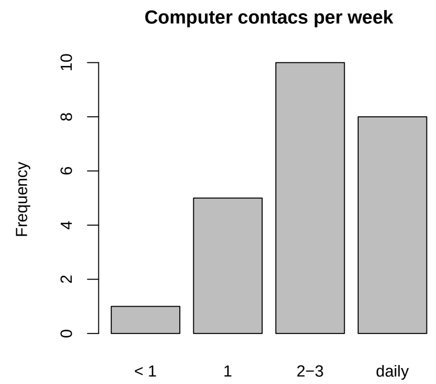

Figure 4: Computer experience of participants.

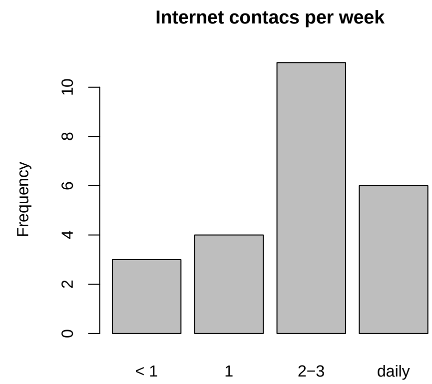

Figure 5: Internet experience of participants.

tions: (a) *Factual questions* can be answered with a single sentence. Tasks like this can typically be answered with a single query. (b) *Open-ended questions* express exploratory information needs that aim towards acquiring broad knowledge about a given topic. (c) *Multi-step questions* require advanced reasoning to combine information acquired over multiple queries in a session. To create an initial feeling of success and to enable the participant to adjust to the search interface, we started with a simple fact-based question, before moving on to the actual search assignments (one per question type).

`1. What do whales eat? (a)`

`2. How many brothers and sisters does Queen Beatrix have? (a)`

`3. What can you find out about the first car ever built? Write down some facts about it. (b)`

`4. Which day of the week is the birthday of the Dutch prime minister in 2011? (c)`

The questions were shown one by one. Only after answering the current assignment, the next one would be made available. After completing the experiment the researcher asked about the participant's opinion on the questions and how she or he liked the overall experience. To prevent frustration in the case of struggling searchers who could not find an answer to a question, we introduced a time limit to the tasks. For the first two questions the participants had 6 minutes each, for the second one 8 minutes and for the last one 10 minutes. After this time, the researcher ended the task and encouraged the participant to move on to the next step. The default search engine shown in the survey interface was Google which has been previously found to be popular among young searchers [13]. We did, however, not restrict the use of other search facilities. After they completed the final search task, participants were once more asked to indicate how much they enjoyed the experiment, before they were guided back to the classroom.

3.2 Data collection

Besides the qualitative observations made by the researcher who takes notes on physical signals of motivation, confidence and immersion, we exploit a range of additional data sources in order to accurately capture relevant session properties. To facilitate manual annotation of search sessions, we used CamStudio 2.0, an open source screen capturing software [8] to be able to revisit all screen activity in the form of video files. Additionally, to create a more machine-readable representation of search sessions, we employed a Firefox add-on, the HCI Browser [9]. This program can be used to log HTTP requests, mouse movements, keyboard input and click data. Instead of logging events on page level by injecting Javascript as is done by popular tools such as Usaproxy [2], this add-on makes it possible to log every action within the browser. Consequently, we can also capture signals as for example the use of the browser's back button that would otherwise have eluded recording. Figure 6 shows an example of the data recorded by the HCI Browser.

4. DATA ANALYSIS

The previously described user study leaves us with a collection of 96 search sessions (4 per unique participant). For each session, we assigned 2 types of labels: (1) A role label, following Druin et al.'s categorization [13]. (2) A binary search success label, indicating whether the participant could find a valid answer to the task. The decisions were based on the qualitative notes taken during the search session as well as the screen recordings of the full sessions. We conduct our annotation on session-level rather than user-level to account for individual preferences and abilities for solving different task types. Each session was independently labelled by 2 researchers. As a measure of task feasibility and annotation reliability, we investigate inter-annotator agreement. An overall share of 82% of all sessions received identical labels by both annotators. Table 1 shows task-level agreement ratios and Cohen's κ scores. We can observe an interesting tendency of task 1 and 4 agreements being significantly higher than those for tasks 2 and 3. The tasks were designed and ordered by increasing difficulty. This initial overview suggests that very easy or difficult tasks are more

```
1292494008865 16-12-2010 11:06:48 Focus http://www.google.nl/ http://www.google.nl/ clientx=1366 clienty=605
1292494008929 16-12-2010 11:06:48 LoadCap http://www.google.nl/ http://www.google.nl/ clientx=1366 clienty=605
1292494009544 16-12-2010 11:06:49 MouseMove http://www.google.nl/ x=596 y=254
1292494010592 16-12-2010 11:06:50 MouseMove http://www.google.nl/ x=606 y=253
1292494010822 16-12-2010 11:06:50 LClick x=606 y=253 undefined http://www.google.nl/
1292494012166 16-12-2010 11:06:52 KeyPress key=D keycode=68 combi= http://www.google.nl/
1292494012503 16-12-2010 11:06:52 KeyPress key=E keycode=69 combi= http://www.google.nl/
1292494013273 16-12-2010 11:06:53 KeyPress key=Space keycode=32 combi= http://www.google.nl/
1292494013795 16-12-2010 11:06:53 KeyPress key=E keycode=69 combi= http://www.google.nl/
1292494014167 16-12-2010 11:06:54 KeyPress key=E keycode=69 combi= http://www.google.nl/
```

Figure 6: Data sample captured by the HCI Browser.

Table 1: Inter-annotator agreement per task.

Task	Agreement	κ
1	0.92	0.83
2	0.71	0.45
3	0.75	0.57
4	0.96	0.92

Table 2: Search role distribution as observed in our school study and Druin et al.'s home setting.

Role	School	Home
Developing searcher	48%	43%
Domain-specific searcher	0%	21%
Power searcher	47%	12%
Non-motivated searcher	2%	9%
Distracted searcher	0%	6%
Visual searcher	3%	5%
Rule-bound searcher	0%	4%

beneficial for determining role affiliations. We will take this intuition as one of our hypotheses for later design and evaluation of our automatic classification scheme. To obtain final judgements, the annotators discussed all instances of disagreement, arriving at consensus labels for each.

Table 2 shows the distribution of roles in our and Druin et al.'s work [13]. The developing and power searcher roles were found to be dominant in the present data set. All other roles could at most be observed sporadically. The developing role was already frequent in the 2010 study, but many other roles follow significantly different frequency distributions. We see the reason for this difference in the changed setting between information search at home and search assignments in a school setting. Due to the more formal environment, phenomena such as non-motivated searchers are intuitively less likely. Both our and Druin's studies find a strong correlation between the participants' age and their likelihood of being a power searcher. An even stronger connection could be found between the participants' school grade and their power searcher status. Despite the correlation between age and school grade, formal school education, rather than age, seems to be the cause for greater search proficiency.

To give further insight into the effect of prior experience in information search and general computing on search success and power searcher status, we analysed this relationship more deeply. There was, no substantial correlation between the participants' background credentials such as their gender or their self-reported computer and Internet experience

Table 3: Development of participant motivation before (rows) and after (columns) the experiment.

	1	2	3	4	5
1	-	-	-	-	-
2	-	-	-	-	-
3	-	-	1	1	2
4	-	-	4	8	3
5	-	-	-	3	2

and their role affiliations and search success. This finding supports our claim that dedicated support and training are valuable even for children who are practised computer users.

In addition to the previously-discussed questions on personal background, we asked each participant about their emotional state before and after participating in the experiment. The questions offered a 5-point scale ranging from "I really do/did not want to participate." to "I really like/liked to participate.". Based on the findings of Yusoff et al. [36], the answers were supported by a smiley-scale that visually underlined emotional states. The concrete scale used is depicted in Figure 7. Table 3 shows the emotional state before and after participating in the experiment.

In the majority of sessions, the emotional state changed during the course of the experiment. In order to further understand this observation, we define δ_e as the number of categories by which a participant's emotional state changed before and after the experiment. A negative number indicates a drop in motivation while a positive number represents gains in well-being. We can find a mild correlation between δ_e and a participant's success rate ($\rho = 0.43$), and their likelihood of being power searcher ($\rho = 0.31$). This underlines the assumption, that search failures can have a frustrating effect on young searchers and may even prevent them from indulging in future searches. This emphasises the importance of appropriate search support at this stage of a child's development.

5. METHODOLOGY

In this section, we present a detailed outline of our automatic search role and success classification scheme. As a starting point, we will describe a wide range of features motivated by intuition as well as by literature in cognitive and behavioural science. We can identify 3 types of features accessible during a search session: (1) *task-independent features* are static properties of the participant such as age or gender. (2) *task-dependent direct features* are directly extractable from the interaction log but may vary across tasks for the same participant. Examples include the number of webpage visits or the number of mouse moves. (3) *task-dependent inferred features*, finally, cannot be directly read from the search log but require further processing steps that may involve external data. An example would be the percentage of natural-language queries issued in the session.

We identified a total of 37 individual features.

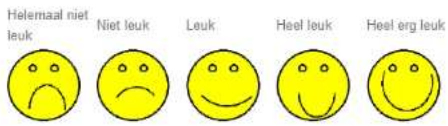

Figure 7: Smiley scale.

Task-independent features

Age. The age of the participant at the time he or she participated in the experiment. Druin already showed that for example older children are more likely to be power searchers than younger children. Bilal et al. investigated the differences between children and adults as web users[5].

Class. The participant's class according to the Dutch school system. While being correlated with participant age, this feature aims more at the amount of formal education the participant has received.

Gender. The participant's gender was included for completeness's sake. Previous research did not find significant gender-specific differences in children's search proficiency.

Computer Experience. The participant's self-reported previous experience with computers can be expected to give good indications of his or her likelihood of success.

Internet Experience. In analogy, we also include the self-reported Internet experience.

Task-dependent direct features

Total number of mouse movements. Qualitative analysis of our search sessions showed a good correlation between motor skills with the mouse and power searcher status. The ability to navigate the search interface with only the necessary user actions (e.g., a low number of mouse moves) is therefore seen as an indicator of operational competence.

Total number of page visits. Capable searchers are able to accurately decide on result relevance based on web page titles and snippets displayed on the search engine result list. High counts of visited and abandoned pages indicate headless browsing.

Total number of mouse clicks. In analogy to the previous features, click events are counted and employed as an additional indicator of operational confidence.

Total number of issued queries. Experienced searchers are expected to be able to phrase their information need more accurately in keywords than beginners who have to rely on subsequent rephrasings. We count the number of such reformulations per session.

Number of query expansions. We count the overall number of times a query is expanded by additional terms. An example of such an operation would be the step from "prime minister birthday" to "prime minister Netherlands birthday".

Number of query prunings. Query generalizations by means of dropping query terms are counted. E.g., from "prime minister Netherlands birthday" to "prime minister Netherlands".

Averaged query edit distance Drastic query reformulations are an indicator of low confidence in the original search terms and, more generally, in the participant's search skills. In this work, we measure the distance in terms of query term overlap between issued queries. We use the Jaccard Coefficient as distance metric. Finally, the computed distances are averaged.

Average query length. Long queries have previously been found to be problematic for modern search engines [20]. We count the number of terms per query and average across all queries within a session.

Query length standard deviation. To give another alternative measure of query reformulation activities, we include the query length standard deviation across a session's queries.

Typing speed. Interaction with keyboard interfaces has been previously reported to be one of the major sources of frustration for inexperienced searchers [17]. We measure typing speed for each sequence of keyboard inputs without any interruptions by mouse moves or clicks. Finally, the number of keystrokes per minute is averaged across all such sequences.

Time spent on search engine pages. We measure the absolute time per session the participant lingers on search engine pages. This represents the combined efforts of query formulation and result inspection.

Number of back button clicks. Inspired by Bilal's model model of children's information seeking [6], we inspect backtracking activities by means of counting the number of times the participant makes use of the browser's back button.

Session length. The total time a participant invests into solving a task is recorded and can be seen as a surrogate for search proficiency.

Number of backspace keystrokes. The number of backspace keystrokes during a session is summed up. Spelling has been frequently observed to be one of the specific challenges of children's query formulation steps [17]. The number of backspaces can give an indication of the participant's orthographic competence.

Number of scroll actions. During our manual inspection of search sessions, we saw that not every child knew how to use the mouse wheel for navigation. To capture the participant's ability to use this advanced control mechanism, we record the total number of mouse scroll actions per session.

Mouse movement patterns. Previous work employed mouse movement characteristics for user authentication [30]. Instead of identifying specific users, we try to generalize mouse movement patterns of groups of users by a number of additional mouse input features beyond the targeted ones that were previously introduced. Concretely, this encompasses: (1) The number of mouse moves per second (2) average mouse move distance (3) move distance standard deviation (4) average horizontal distance (5) average vertical distance (6) the ratio of vertical / horizontal distances.

Number of visits per unique page. Inexperienced young searchers have been found to seemingly arbitrarily revisit web pages multiple times [13]. We exploit this observation by measuring the average number of times each unique page was visited within a session.

Average display time per webpage. Query log analyses of children's interaction with a popular search engine showed, that young children often experience difficulties judging the relevance of search result snippets which manifests in a high number of very brief visits that are quickly abandoned once the participant realizes that the page was not what he or she was looking for [14]. We capture similar behaviour in terms of average display times per web page.

Task-dependent inferred features

Question words. Inexperienced searchers tend to "ask" the search engine for information. We check for the presence of question words such as "why", "when", "who" etc.

Stop words. Modern search engines are designed and optimized for accepting keyword queries. Excessive usage of stop words indicates low search experience. We report the averaged number of stop words per query.

Query-task distance. Query formulation is a crucial and cognitively-expensive step in the information search process. Inexperienced searchers have been found to take the "shortcut" of copying the assignment question as a query. We measure the Jaccard distance between tasks and observed queries. We expect this distance to be minimal for developing searchers and significantly larger for experienced users.

Average number of verbs|nouns|adjectives per query. Developing searchers tend to issue natural language queries [13]. We apply part-of-speech tagging to identify different token type distributions.

6. EVALUATION

The previous section described a wide range of features motivated by related literature, observations made during the study, as well as intuition. As a starting point, we trained a number of different machine learning techniques for the task of automatic role classification based on the collected session data and evaluate classification performance in a 10-fold cross-validation setting. Our experiments are based on the WEKA library's [16] implementation of the various machine learning techniques. Table 4 shows the best classification performance per method averaged across all classes. To set our results into context, we include a dominant class baseline that assigns the most frequent label to all sessions. All evaluated methods performed significantly and consistently better than the baseline intuition. The overall strongest approach was a support vector machine (SVM).

Table 4: Role classification performance by method.

Method	P	R	F_1
Dominant class baseline	0.23	0.48	0.31
Naive Bayes	0.65	0.68	0.66
Logistic Regression	0.59	0.57	0.58
MLP	0.74	0.75	0.74
SVM	**0.76**	**0.80**	**0.78**
Decision Table	0.61	0.65	0.63
Decision Tree	0.59	0.62	0.60
Random Forest	0.66	0.69	0.67

Table 5: Deficit detection performance by method.

Method	P	R	F_1
Dominant class baseline	0.28	0.53	0.37
Naive Bayes	0.66	0.82	0.73
Logistic Regression	0.67	0.71	0.69
MLP	0.79	0.73	0.76
SVM	**0.84**	**0.80**	**0.82**
Decision Table	0.67	0.67	0.67
Decision Tree	0.63	0.63	0.63
Random Forest	0.69	0.69	0.69

A reliable means of identifying individual search role affiliations makes an important contribution to educating children regarding their web search abilities. Knowledge about their specific deficits (e.g., those of a visual searcher) helps teachers and parents to give targeted advice on how to improve. For the task at hand, however, we can reformulate our task to finding those children in the classroom that fall into one of the defective search roles (i.e., all except for power searchers). We will refer to this lower-order classification problem as *Deficit prediction*. Adjusting to this new setting, we achieve significantly higher scores than for dedicated role prediction. The strongest models approximate the agreement ratio of our human annotators. Table 5 reports the resulting performance figures.

When working with search roles as defined by Druin et al., we noticed a conceptual disparity between some of the categories. While some roles are essentially performance oriented (power and developing searchers), others are based on the employed search strategy (visual, rule-bound and domain-specific searchers) and a third group is concerned with notions of attentiveness (non-motivated and distracted searchers). While a manual qualitative analysis of searcher behaviour may benefit from such a broad categorization scheme, it appears to be problematic for automatic methods designed for classroom teacher support. Given this use case, the search roles formed a surrogate for search success. At this point, we will investigate the feasibility of predicting search success directly from the session features. This appears intuitively sound based on the definition of power searcher status. A closer investigation of the data set, however, showed that power searcher status is only loosely correlated to the likelihood of search success ($\rho = 0.4$). It ap-

Table 6: Success prediction performance.

Method	P	R	F_1
Dominant class baseline	0.27	0.52	0.36
Naive Bayes	0.63	0.78	0.70
Logistic Regression	0.59	0.60	0.59
MLP	0.66	0.72	0.69
SVM	**0.77**	**0.75**	**0.76**
Decision Table	0.64	0.80	0.71
Decision Tree	0.58	0.60	0.59
Random Forest	0.66	0.66	0.66

Table 7: Best deficit / success prediction features.

Rank	Deficit	Success
1	# query shortenings	class
2	class	# query shortenings
3	# query nouns	# query nouns
4	horiz. mouse distance	# back buttons uses
5	mouse move interval	avg. query length
6	# back button uses	# query adjectives
7	avg. query length	# of visited pages

pears as if Druin's roles cannot be seen as direct surrogates for search proficiency. Further evidence was given in Section 4, where we observed a relationship between searcher motivation and search success.

In our final classification experiment, we turn to directly predicting search success without first determining search roles. Table 6 compares the performance of a number of classifiers for this task. The best overall performance could be achieved using an SVM approach with polynomial kernel ($\epsilon = 10^{-12}$, $c = 0.6$ and $e = 1$). Based on these performance figures, a classroom teacher could prioritise the order in which she or he visits students, based on their likelihood of search success as determined by an automatic classifier running in the background of the school's computers. Given the substantial performance gains over baseline intuitions, our method can be expected to result in less time being invested into identifying struggling students. This, in turn, frees up resources for actual assistance and teaching.

In order to gain a deeper understanding of the domain, we identified the best-performing features according to our SVM model. Table 7 shows the top 7 features for the tasks of deficit and success prediction. We find a high overlap between both sets, confirming the central role of those notions. In both scenarios, being in a higher school grade, phrasing short queries with only few nouns and refraining from substantial query shortenings, are indicators of successful searches and power searcher status. The ranking of features is consistent across task types with only minor differences in relative contribution weights. The relative contribution of features decreases rapidly with rising rank. Models based on the 3 highest-ranking features were able to approximate the performance of those incorporating the full feature space, showing no significant differences in performance.

7. CONCLUSION

In this work, we described a user study in a Dutch elementary school investigating 9-12-year-old children's web search behaviour. Based on the research of Druin et al. [13], we annotated the captured search sessions with role labels and investigated a wide range of features to enable automatic role classification to aid teachers in the classroom. Initial analyses of the collected data suggested a significantly different role distribution than was observed in Druin's original work. We attribute this difference to the domain change from home information search to a more formal school setting, as well as to the different age distribution of participants. In this new domain, most of the roles were only observed very infrequently.

The two dominant roles, *developing searchers* and *power searchers* can be seen as surrogates for the generally-accepted notion of search success. To account for this observation, we switched to predicting search success rather than roles. For both predicting concrete roles as well as predicting search success, we demonstrated significant improvements over baseline heuristics and were able to closely approximate human annotator performance. Based on the automatic models, we could form a number of key intuitions that best identified experienced and successful searchers. Most prominently, they issue short queries, limited to a low number of keywords. Additionally, our study showed the important role that formal school education plays in acquiring information search skills.

The implications this work has on the educational sector are two-fold: (1) Search success prediction can be reliably used to aid teachers to more efficiently identify those children that struggle with a search assignment and that would therefore benefit from assistance. (2) Role prediction is a valuable method for identifying children's search strategies. Some of these strategies are better suited for use in web search scenarios than others. Gaining knowledge about children's search strategies enables teachers and educators to provide targeted guidance, highlighting difficult aspects of the search process and how to best address them. The concrete roles drawn from previous work may, however, need to be revised for application in the classroom setting.

Future work in the domain of automatic classroom assistance based on search behaviour shows two especially promising alleyways: (1) In this study, we aimed for classification of complete sessions to confirm the general feasibility of the task. For practical application, it would be interesting to investigate intermediate classification to give assistance *during* the task. Most notably, we will establish estimates of classification confidence as a function of session length. (2) The insights gained in this work motivate interactive information retrieval experiments in which an automatic help feature assists the searcher. Once a problematic search strategy is detected, the system could hint at alternative, more promising solutions. In this way, we foresee a direct positive impact on the current session's likelihood of success as well as an educational effect of the searcher learning how to perform future searches better in the first place.

Acknowledgements

We would like to thank Allison Druin and Elizabeth Foss for their support and for sharing their original search role coding scheme. This research is part of the PuppyIR project[1]. It is funded by the European Community's Seventh Frame-

[1] http://www.puppyir.eu

work Programme FP7/2007-2013 under grant agreement no. 231507.

8. REFERENCES

[1] D.E. Agosto and S. Hughes-Hassell. Toward a model of the everyday life information needs of urban teenagers, part 1: Theoretical model. *JASIST*, 57(10), 2006.

[2] R. Atterer, M. Wnuk, and A. Schmidt. Knowing the user's every move - user activity tracking for website usability evaluation and implicit interaction. In *WWW*, 2006.

[3] N.J. Belkin and W.B. Croft. Information filtering and information retrieval: two sides of the same coin? *Communications of the ACM*, 35(12), 1992.

[4] D. Bilal. Children's use of the yahooligans! web search engine: I. cognitive, physical, and affective behaviors on fact-based search tasks. *JASIS*, 51(7), 2000.

[5] D. Bilal and J. Kirby. Differences and similarities in information seeking: children and adults as web users. *Information processing & management*, 38(5), 2002.

[6] D. Bilal, S. Sarangthem, and I. Bachir. Toward a model of children's information seeking behavior in using digital libraries. In *IIiX*. ACM, 2008.

[7] C.L. Borgman, S.G. Hirsh, V.A. Walter, and A.L. Gallagher. Children's searching behavior on browsing and keyword online catalogs: the Science Library Catalog project. *JASIS*, 46(9), 1995.

[8] CamStudio. Free streaming video software. `http://camstudio.org`, 2012.

[9] R. Capra. Hci browser: A tool for studying web search behavior. *JASIST*, 47(1), 2010.

[10] C.W. Choo, B. Detlor, and D. Turnbull. Information seeking on the web: An integrated model of browsing and searching. *First Monday*, 5(2), 2000.

[11] D.H.S. Czajka and S. Mohr. Internetnutzung in privaten haushalten in deutschland. *Ergebnisse der Erhebung*, 2008.

[12] A. Druin, E. Foss, L. Hatley, E. Golub, M.L. Guha, J. Fails, and H. Hutchinson. How children search the internet with keyword interfaces. In *Proceedings of the 8th International Conference on Interaction Design and Children*. ACM, 2009.

[13] A. Druin, E. Foss, H. Hutchinson, E. Golub, and L. Hatley. Children's roles using keyword search interfaces at home. In *SIGCHI*. ACM, 2010.

[14] S. Duarte Torres and I. Weber. What and how children search on the web. In *CIKM*. ACM, 2011.

[15] C. Eickhoff, P. Serdyukov, and A.P. de Vries. A combined topical/non-topical approach to identifying web sites for children. In *WSDM*. ACM, 2011.

[16] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I.H. Witten. The weka data mining software: an update. *ACM SIGKDD Explorations Newsletter*, 11(1), 2009.

[17] H. Hutchinson, A. Druin, B.B. Bederson, K. Reuter, A. Rose, and A.C. Weeks. How do i find blue books about dogs? the errors and frustrations of young digital library users. *Proceedings of HCII 2005*, 2005.

[18] P. Ingwersen and K. Järvelin. *The turn: Integration of information seeking and retrieval in context*, volume 18. Kluwer Academic Pub, 2005.

[19] C.C. Kuhlthau. Inside the search process: Information seeking from the user's perspective. *JASIS*, 42(5), 1991.

[20] G. Kumaran and V.R. Carvalho. Reducing long queries using query quality predictors. In *SIGIR*. ACM, 2009.

[21] A. Large, J. Beheshti, and T. Rahman. Design criteria for children's Web portals: The users speak out. *JASIST*, 53(2):79–94, 2002.

[22] Dutch Law. Wet medisch-wetenschappelijk onderzoek met mensen. `http://wetten.overheid.nl/BWBR0009408`, 1998.

[23] Dutch Law. Wet bescherming persoonsgegevens. `http://wetten.overheid.nl/BWBR0011468`, 2000.

[24] European Law. European directive on privacy and electronic communications. `http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2002:201:0037:0047:EN:PDF`, 2002.

[25] S. Livingstone and L. Haddon. Eu kids online: Final report. *LSE, London: EU Kinds Online (EC Safer Internet Plus Programme Deliverable D6. 5)*, 2009.

[26] G. Marchionini. *Information seeking in electronic environments*. Number 9. Cambridge Univ Pr, 1997.

[27] G. Marchionini. Exploratory search: from finding to understanding. *Communications of the ACM*, 49(4), 2006.

[28] P.A. Moore and A. St George. Children as Information Seekers: The Cognitive Demands of Books and Library Systems. *School Library Media Quarterly*, 19(3):161–68, 1991.

[29] Ofcom. Uk children's media literacy: Research document. `http://www.ofcom.org.uk/advice/medialiteracy/medlitpub/medlitpubrss/ukchildrensml/ukchildrensml1.pdf`, March 2010.

[30] M. Pusara and C.E. Brodley. User re-authentication via mouse movements. In *Proceedings of the 2004 ACM workshop on Visualization and data mining for computer security*. ACM, 2004.

[31] T. Saracevic, P. Kantor, A.Y. Chamis, and D. Trivison. A study of information seeking and retrieving. iii searchers, searches, and overlap. *JASIS*, 1988.

[32] A.K. Shenton and P. Dixon. Models of young people's information seeking. *Journal of Librarianship and Information Science*, 35(1), 2003.

[33] A.J. Stronge, W.A. Rogers, and A.D. Fisk. Web-based information search and retrieval: Effects of strategy use and age on search success. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 48(3), 2006.

[34] R.M.C. Wallace, J. Kupperman, J. Krajcik, and E. Soloway. Science on the web: Students online in a sixth-grade classroom. *The Journal of the Learning Sciences*, 9(1):75–104, 2000.

[35] K.P. Yee, K. Swearingen, K. Li, and M. Hearst. Faceted metadata for image search and browsing. In *SIGCHI*. ACM, 2003.

[36] Y.M. Yusoff, I. Ruthven, and M. Landoni. The fun semantic differential scales. In *Proceedings of the 10th International Conference on Interaction Design and Children*. ACM, 2011.