A Combined Topical/Non-topical Approach to Identifying Web Sites for Children

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

Pavel Serdyukov
Delft University of Technology
Delft, Netherlands
p.serdyukov@tudelft.nl

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

ABSTRACT

Today children interact more and more frequently with information services. Especially in on-line scenarios there is a great amount of content that is not suitable for their age group. Due to the growing importance and ubiquity of the Internet in today's world, denying children any unsupervised Web access is often not possible. This work presents an automatic way of distinguishing web pages for children from those for adults in order to improve child-appropriate web search engine performance. A range of 80 different features based on findings from cognitive sciences and children's psychology are discussed and evaluated. We conducted a large scale user study on the suitability of web sites and give detailed information about the insights gained. Finally a comparison to traditional web classification methods as well as human annotator performance reveals that our automatic classifier can reach a performance close to that of human agreement.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Information Filtering*; H.1.2 [**Models and Principles**]: User/Machine Systems —*Human Factors*

General Terms

Human Factors

Keywords

Children, Classification, Filtering, Suitability, Web Search

1. INTRODUCTION

The use of the Internet has become an integral part of most people's lives. This tendency also holds true for children who naturally adopt the behaviour that is displayed by parents and teachers. The age at which children have their first contact with the Internet has become consistently and significantly lower over the last years [31]. Even adults often struggle to cope with the amounts of information on the Web. Judging which content is relevant for their information need can be hard. Children face the same situation, but suffer even more from it as they tend to unquestioningly believe in any kind of information [29] and readily absorb new knowledge. While this in general is a desirable and essential characteristic that enables children to easily acquire large amounts of information such as new words, in the case of the Internet it imposes potential dangers. Several popular web search engines such as Google and Yahoo! even state in their terms of service[1][2] that their search should not be used by minors. It is hardly possible to supervise children's Internet usage continuously in person as would be necessary in order to make sure that children are not exposed to inappropriate content. Recent surveys [27] found that close to 40% of UK children aged 5-15 years access the Internet without parental guidance or supervision. Considering these numbers, an automatic means of determining child-appropriateness of web pages would be highly desirable.

State of the art children's web search engines could greatly benefit from a method of finding children's web sites. At the moment children's search resources are typically directories of manually selected web pages. As such they offer high quality children's pages, but due to the high amount of manual labour involved in their maintenance they are less flexible and have a much more limited coverage than an automatic approach could achieve. Examples of current web search facilities for children are Yahoo! Kids [5] or Ask Kids [1]. It is important to note that assuring appropriateness exceeds filtering offensive content. Most state of the art search engines offer safe search functionalities that reliably remove adult material. This kind of filtering usually takes a topical approach, which is already well-understood. Notions of text difficulty and age-appropriate web site design however are independent of the page topic and should strongly contribute to the decision of showing a certain page to a child. An automatic solution should take care of various aspects, assessing appropriateness in terms of topical relevance, textual content difficulty and presentation style. In this work, we will show how well-studied means of topical classification can be augmented by non-topical techniques in order to make the multi-faceted suitability decision.

The contributions of our work are threefold: 1) We identify the criteria of good children's web pages based on children's specific needs and abilities and show how to encode

[1]http://www.google.com/accounts/TOS
[2]http://info.yahoo.com/legal/us/yahoo/utos/

WSDM'11, February 9–12, 2011, Hong Kong, China.

these criteria with features. 2) We conducted a large-scale user survey to better understand the nature of children's web pages. The results of this survey and its implications for our work are discussed in detail. 3) We describe a classification method for web pages, demonstrating that purely topical models can be outperformed by models augmented by non-topical aspects.

The remainder of this paper is structured as follows. We begin with a brief review of related research in Section 2. In Section 3, we introduce a range of web page features and show their relevance to our task. Section 4 aims to find promising subsets of the full feature space. After presenting our performance baseline, we conduct a range of experiments to evaluate our classifier's performance in Sections 5 and 6. Finally, we close with a discussion of the insights gained from this study in Section 7.

2. RELATED WORK

While to the best of our knowledge there has not been prior research on finding suitable web content for users of different age groups, there are a number of related tasks that have been addressed recently. As early as in the 1950s advances have been made into designing readability measures that would capture the difficulty of a natural language text [24]. Collins-Thompson and Callan investigated the use of language models to estimate reading difficulty [11]. They employed various modified naive Bayes models to represent different reader age groups. Schwarm and Ostendorf applied support vector machines to determine a text's reading level [30], using a wider range of features such as for example linguistically-motivated information. The state of the art mainly focused on shallow features. Recent work on readability assessment now also employs grammar- and discourse-level features [8, 13].

Previous work using ODP [4], a large scale web taxonomy, reported the Kids & Teens category yielded particularly low result scores [7]. This type of classification typically benefits from exploiting topic-inherent similarities within each topical category. Those occur less frequently in a heterogeneous category such as Kids & Teens which can deal with an arbitrary range of topics as long as the pages satisfy the suitability requirements. For this reason some authors [15] altogether excluded this branch due to its different structure. This tendency suggests that classification based on purely topical web site aspects is not sufficient for the task of finding children's resources. The fact that children's search engines still heavily rely on manual classification strongly supports this hypothesis.

Non-topical web classification has been a thriving field for applications like the identification of weblogs [19], web spam detection [10] and sentiment analysis [22]. This work will show that neither readability measures nor language modelling approaches are satisfying predictors of suitability when considered in isolation. We will introduce a more appropriate way of identifying children's pages by combining topical and non-topical aspects taking into account the specific needs of children.

3. FEATURE EXTRACTION

Our experimental dataset was acquired using the Open Directory Project [4]. In this Internet directory web pages are represented as leaves in a hierarchical topical tree. The Kids & Teens section of the ODP has human annotators categorize web pages and, if appropriate, set one or several of the age flags {kid ($\leq$12), teen (13-15), mature teen (16-18)} indicating that the page's content is suitable for those age groups. The ODP editors state in their content selection guidelines that a good children's web page should be:

- informative
- age-appropriate
- non-commercial
- **for** children, not **about** children

We identified parallel branches between the children tree and the general tree to select web pages that deal with the same topic from an adult/ child perspective. This way we hope to reduce the topic-specific noise in the pages and to detect age-specific patterns that can be used for classification. The specific information needs of teens and mature teens are in the middle ground between those of children and adults. We kept only pages suitable for children in order to get clearly disjoint categories for this study. Since many of our features are language-dependent we concentrate on English resources by excluding the *World* and *Regional* branches of the ODP as suggested by for example [7]. We extracted a training set of 20,778 web pages (6225 for children and 14553 for adults) from a range of 1350 distinct topics. Among the kids pages, a share of 26.71% were deemed suitable exclusively for children.

Expressing appropriateness in features

Determining the child appropriateness of web content is a novel task for which we explored a wide range of prospective features that reflect human notions of web site child-suitability. To better understand what this means for web pages we will discuss in detail the various criteria that qualify a "good" children's page. The two main dimensions of appropriateness are child-friendliness and focus towards child audiences.

Child-friendliness

The first and most important characteristic of any children's page is its child-friendliness. The decision of what qualifies as child-friendly should ultimately be made by children. Based on recent studies on children's preferences towards web sites [9, 21, 25, 26] we propose a range of child-friendliness criteria. We try to formalize their findings, discussing child-friendliness in terms of complexity of text, presentation style, navigational as well as ethical considerations.

Complexity of text

There are significant differences between texts suitable for children who are inexperienced readers at best and those for adults. We expect child-friendly web pages to rely on a language use and general design of textual resources that respect these specific aspects.

Shallow features (12)[3]**.** State of the art text readability assessment often uses shallow characteristics as a measure of syntactic text complexity. There are commonly assumed

[3]The numbers in brackets denote the number of distinct features within each category.

to be significant differences in complexity between texts for children and adults. Examples of features in this category are the average number of words per sentence, the number of complex (3+ syllables) words, or the average word length.

Readability scores (8). A more high level notion of syntactic complexity is delivered by automatic readability scores [18]. Most of them are linear combinations of several shallow features that result in an age group whose members should be able to understand the evaluated text. Examples are the well-known ARI or Coleman-Liau measures.

Part-of-speech features (5). The features in this category are based on the notion of syntactic differences between adult and children's texts on a higher linguistic level than could be captured by mere shallow features. Statistics of POS tag distribution are generated for the textual page content. They include various POS parse tree statistics as, for example, the average number of noun phrases per sentence or the token/type ratio of observed words.

Entity occurrences (6). Commonly, children's text is not only syntactically, but also semantically, simpler than general text. Children's cognitive abilities are not yet suited for understanding complex texts, as for example newspaper articles, which often contain several different entities per sentence. This is reflected by a smaller number of entities per article and per sentence in low reading level texts [14]. We use the LingPipe toolkit [6] to extract the entity types person, location and organization.

OOV rates (8). On child-friendly web sites we notice not only a simpler text structure in terms of shorter sentences containing less named entities, but also the use of more basic language. To reflect this difference, we constructed 7 distinct vocabularies of the most frequent/ most basic English words. We expect the out of vocabulary rates of adult texts for these vocabularies to be higher than those of children texts which commonly use a smaller and simpler range of words. We use an additional vocabulary of academic terms for which the opposite tendency is expected.

Wiktionary features (4). A possible way of capturing textual complexity makes use of the Wiktionary on-line dictionary. Ambiguous words which have a number of possible meanings (dictionary definitions) are supposedly harder to understand and use than unambiguous ones. The great coverage of Wiktionary should enable us to capture vocabulary difficulty and cognitive complexity of texts in a more universal way than mere OOV rates would allow. We use statistics on, for example, the average number of definitions or average definition length of a page's textual content to represent its cognitive complexity.

Presentation

Child-friendly web sites should be presented in a way that is appealing to children. They should make use of the types of media that children, even at younger ages, are familiar with. While longer textual resources often cause frustration, the use of videos, images and colours in general have been shown to appeal to them. Large et al. concluded that a page's content can be relevant and still children will not look at it unless its presentation is also attractive [21]. In this work we will measure child friendliness of presentation through HTML page characteristics as well as visual features.

HTML features (10). Many high quality children's pages run a great number of scripts and animations in order to make them interesting and usable for an audience with limited reading abilities. We incorporate various HTML features as for example the tag distribution or the number of scripts on a page in order to capture the page's specific presentation style. Especially scripts were expected to be a strong indicator of child-friendliness.

Visual features (8). Child-friendly web pages often rely on great numbers of images to convey their message. Especially for age groups that have not yet developed high literacy skills, visual resources are easier to understand. To further pursue this notion, we analyse the use of visual elements in terms of the number, type and size of pictures on the page.

Navigational

While the previous two aspects of child-friendliness were concerned with the page's content, navigational aspects target the embedding of the page into its link neighbourhood. Previous research in topical classification of web pages [28] has found that a page's neighbours are strong indicators of its topic. However the assumption that pages on a given topic contain links to or are linked from other pages on that topic does not always hold. If we however transfer the same principle onto age scale it may prove even more powerful as children's pages regardless of the actual topic should not link to non-child-friendly pages. Large et al. [20] found that children prefer browsing over searching and generally trust links without closer inspection of anchor text. This exploratory search behaviour of children requires that safe children's web sites do not contain links to web pages for adults.

Link neighbourhood features (2). We use a basic version (without neighbourhood analysis) of our classifier to analyse a web page's outgoing links. The share of pages that were classified as for adults/ for kids with at least a given threshold confidence are incorporated as features. Link analysis approaches sometimes take into account not only the page's immediate neighbourhood but also pages with a distance of 2 or more links. Since using more than one level of neighbouring pages did not yield significantly better results for our application, we restrict the neighbourhood to web sites directly linked on the classified page to limit computational cost. Although previous work has analysed incoming links, we exclusively consider outgoing links. The reason for this is the hypothesis on which these features are built. While a children's page should not link to an adult page, there is no such limitation for an average page for adults.

Ethical

The final dimension of child-friendliness to be considered in this work is based on ethical considerations. The particular ethical concern lies in the presenting of advertisement to children who were found to be less resistant to marketing strategies than adults [26].

Commercial intent (1). As stated in the ODP's content guidelines child pages should not be of commercial nature even though their products (e.g., toys) may be targeted towards children. We use the Microsoft AdCenter on-line commercial intent detection [12] as an indicator of suitability. Pages with a high likelihood of commerciality are considered not child-friendly. It is important to note that there are two possible reasons for page commerciality. A given page can either offer products or services itself or display advertisements which lead to actual commercial pages. Both forms of commerciality are considered inappropriate for children as it would clearly try to exploit their inexperience.

Focus towards children

Sometimes, web pages are easy to understand and not harmful, although they do not convey an impression of being intended for children. These pages qualify as child-friendly according to all previously discussed dimensions but they are not targeted towards a child audience. To capture this second aspect of appropriateness we will discuss what makes a page focused on children. Often this focus is already expressed through the choice of topic as some topics such as *colouring books* or *The Sesame Street* are mainly interesting for children. Focus can also be visible in the way the reader is addressed on the page. Web sites focusing on child audiences often employ a distinct style of addressing them (e.g. using baby talk or diminutives).

LM scores (9). Language models have been widely shown to be strong representations of topical affiliation in web scenarios. Using a language modelling approach we hope to capture the language use specific to children's web pages. Textual resources from Simple Wiktionary (http://simple.wiktionary.org, 18,206 entries), Simple Wikipedia (http://simple.wikipedia.org, 108 articles) and web page content for children (A subset of DMOZ pages topically disjoint from our training and test sets) are used to build up character-n-gram, uni-gram and token-n-gram language models for this purpose. We used the simple Wikipedia/ Wiktionary versions because their more basic language use is easier to understand for young readers and thus closer to the language good children's pages will display. The language model score $P_{LM}(T|cat)$ is computed as the maximum likelihood estimate of the observed text T given the category's language model.

$$P_{LM}(T|cat) = \prod_{t \in T} P_{LM}(t|cat)$$
$$P_{LM}(t|cat) = \lambda \frac{count(t,cat)}{|cat|} + (1-\lambda) P_{backoff}(t)$$

For each term the number of occurrences within the category $count(t, cat)$, divided by the overall number of category terms $|cat|$ is computed. An interpolated character-n-gram model $P_{backoff}(t)$ serves for smoothing purposes in Jelinek-Mercer fashion with smoothing factor λ. Each page is scored against these models and the scores are used as features.

Reference features (2). In order to find children's pages we localized clue words (considering all affixations of the terms "child" and "kid" and manually rejecting those terms that had no relevance to the children's domain, e.g. "kidney"). We analysed text windows of variable size around these terms. N-gram counts for the windows are collected. The notion behind using this approach is that there should be a difference in the way children are referred to on general pages as opposed to on child pages. On a general page about education or childcare we expect to observe higher frequencies of strings like "your child" or "the average child". Here children are **talked about**. The reference style should be different on actual children's pages where phrases like "for kids" or "us kids" in which kids are **talked to** are assumed to be more dominant. Finally the share of about-references and to-references are reported as features.

$$p(kids|page) = \frac{1}{|M_{n,page}|} \sum_{w \in M_{n,page}} p(kids|w)$$
$$p(kids|w) = \begin{cases} 1 & if\ c_{rel}(w) > \delta_{threshold} \\ 0 & else \end{cases}$$
$$c_{rel}(w) = \frac{count(w,kids)}{count(w)}$$

Where $M_{n,page}$ denotes the set of text windows of size n around the page's clue word occurrences. $p(kids|w)$ expresses whether the term w is a to-reference. $c_{rel}(w)$ is the ratio of n-gram w's occurrences on children's pages versus its general frequency. $\delta_{threshold}$ is the threshold value of $c_{rel}(w)$. Only terms w that reach this threshold are considered relevant. Best results could be achieved for a window size of 2 words (one before and one after the actual clue word) and a $\delta_{threshold}$ of 0.66.

URL features (5). Well-designed web sites put considerable effort into the choice of domain name. Good examples are "www.pbskids.org" or "www.kidsdinos.com". Previous surveys even found that children preferred pages with funny URL names [21]. We inspect the occurrences of child terms within the URL by considering all its sub-strings that form valid terms contained in our simple Wikipedia vocabulary. The maximum likelihood estimate of these terms according to our children's text language model is incorporated as an additional feature.

Page segmentation

Previous research [16] has shown page segmentation to be beneficial for web page classification. We investigated its impact by splitting the web pages into title, headlines, anchor texts and main text. For each of these segments the above features were extracted and used as an extended feature space. Some features have to be considered on page level rather than on segment level. Therefore the HTML, URL and visual features, a page's commercial intent and its link neighbourhood are extracted per page. Using page segmentation the original set of 80 features is strongly enlarged yielding a 242-dimensional new feature space. A complete overview of all features used in this work can be found at http://blackboard.tudelft.nl/bbcswebdav/users/ceickhoff/features.xls.

4. FEATURE ANALYSIS & SELECTION

In the previous section, we introduced a number of promising features founded on very different assumptions, which we hope capture the essence of what makes a good children's web site. The aim of this paper is to gain deeper understanding of what makes a web site a suitable and valuable resource for children. Trying to comprehend the individual contributions of each of the features in a 242-dimensional space is however hardly possible for humans. Therefore, we

Table 1: Top 10 features by Information Gain ratio

Feature	IG
Child neighbourhood rate	0.050
Occurrences of "kid" in main text	0.026
Kid's 1-gram LM on title	0.021
Kid's 3-gram LM on title	0.016
Wiki 3-gram LM on title	0.016
Wiktionary 3-gram LM on title	0.016
To-references on title	0.014
Coleman-Liau on headlines	0.013
To-references on main text	0.012
Kid's character LM on title	0.010

employ different means of feature subset selection and feature/category evaluation in order to better understand how to automatically assess suitability.

Reducing the number of features is not just good for human interpretation of results, but often also beneficial for classification performance. Several state of the art machine learning methods tend to perform sub-optimally when facing a large number of redundant or even irrelevant features.

Information Theoretic feature evaluation

An accepted and well-known estimator of feature performance is the information gain or mutual information criterion. It measures the reduction in entropy when introducing an additional feature. To get a first notion of the importance of the various features described previously we ranked them by their information gain ratios. Table 1 shows the top 10 features according to this ranking.

The strongest overall feature according to information gain is the share of linked pages that were classified as suitable for children. This supports our intuition that web page neighbourhoods should be homogeneous for good children's pages. The high ranking of the number of occurrences of the term "kid" on the page as well as the child reference ratios for main text and title follow the requirement that children's web pages should be targeted towards them. As required good children's pages will mention kids (kid term occurrence) and will do it by addressing them (high child reference ratio) rather than talking about them. Finally among the top-rated dimensions we observe many features from the title segment and the category of language models. This suggests that already the title of a web page is a strong predictor of its suitability and that a lot of the suitability decision is captured within the page's vocabulary use.

To further inspect the predictive potential of certain web page aspects we compared average information gain scores per category in order to see which categories contribute most to the reduction of entropy. The results of this comparison can be found in Table 2. We can observe the same tendency that already emerged in the top 10 ranking of single features. Web page neighbourhoods remain by far the strongest overall category, followed by language models. Entity, Wiktionary and OOV features at this stage hardly add information that was not already expressed by the language models. At the bottom of the ranking we find the commercial intent score. This clearly surprised us as according to the DMOZ editors the children's section should not contain commercial web pages. We inspected the data further and found that there are no significant differences in commercial intent of children's (intent confidence $\mu = 0.31$ and $\sigma = 0.255$) and general pages ($\mu = 0.28$ and $\sigma = 0.27$). Since commercial intent detection has been shown to yield reliable results [12] we have to assume that in spite of their content guidelines the distribution of commerciality among pages for children and grown-ups in DMOZ is rather arbitrary. Manual assessment of web pages showed that the majority of commercial pages from the kids class published advertisement banners rather than actively offering products or services. This carelessness with respect to advertisements, however, is a key weakness of today's Internet for children.

Table 2: Avg. Information Gain per category

Category	Average IG
Neighbourhood	0.0280
LM	0.0050
URL	0.0049
Reference	0.0043
Visual	0.0034
Shallow	0.0022
POS	0.0022
HTML	0.0020
Readability	0.0020
Entity	0.0013
Wiktionary	0.0010
OOV	0.0009
Commercial	0.0007

Table 3: Avg. Information Gain per segment

Segment	Average IG
Title	0.0057
Body	0.0041
Headlines	0.0032
Anchor text	0.0030

Finally the same comparison was conducted per segment. The results can be found in Table 3. Again we see the tendency of a high importance of the title segment confirmed. The results presented in this section give an impression of the predictive strength of individual features, categories and segments. As we will show in the next section however, a high information gain score does not guarantee that a feature will end up in the best-performing subspace. Often the relevant information lies in the interplay of several features. Single feature information gain is not suitable to capture such synergies.

Feature subset selection

After the information theoretic inspection we will now evaluate feature subsets' actual prediction performance in order to find strong indicators of child suitability. We tried a number of different state of the art classification methods and found logistic regression to be the strongest overall method for this application. The results reported in this section will therefore always refer to a logistic regression classifier trained on the varying feature sets and evaluated using 10-fold stratified cross validation. We will report precision and recall for this task as well as the $F_{0.5}$-measure and ROC area under curve. We decided for the precision-biased F-measure as recall is desirable but precision is the crucial aspect for

Table 4: Classification performance per category

Category	P	R	$F_{0.5}$	ROC
All features	0.79	0.55	0.73	0.78
Neighbourhood	0.76	0.54	0.70	0.75
LM	0.69	0.67	0.69	0.76
Shallow	0.74	0.45	0.66	0.71
HTML	0.73	0.45	0.65	0.71
URL	0.72	0.47	0.65	0.72
POS	0.66	0.50	0.62	0.68
Readability	0.66	0.47	0.61	0.66
Wiktionary	0.67	0.45	0.61	0.65
OOV	0.65	0.45	0.60	0.64
Entity	0.63	0.47	0.59	0.63
Reference	0.79	0.12	0.37	0.55
Visual	0.29	0.11	0.22	0.56
Commercial	0.10	0.05	0.08	0.50

Table 5: Classification performance per segment

Segment	P	R	$F_{0.5}$	ROC
Full page	0.79	0.55	0.73	0.78
Title	**0.78**	**0.49**	**0.70**	**0.74**
Body	0.70	0.45	0.63	0.69
Anchor	0.65	0.50	0.61	0.66
Headlines	0.62	0.37	0.55	0.59

child-friendly web search that wants to promote as few as possible out-of-age-group results.

Category/Segment-based subsets

As a first step to understanding the predictive power of different feature sets we follow the natural division present in the data, namely feature categories and web page segments. We will inspect the individual performance of each division by excluding all other features for classification. The results of this analysis are shown in Table 4.

The tendency that was observed for information gain ratios in the previous section is repeated on classification scores. Link neighbourhood is still the strongest feature aspect, very closely followed by the language models. Most categories, regardless of their singular predictive power, add a small contribution to the overall score. The relatively weak performance of reference features is due to the low share of pages actually mentioning children. While child mentions prove to be a valuable page characteristic when they occur, for most pages it cannot be applied. When considering these results, it should be taken into account that the feature categories differ in size. Some of them contain 10 or 12 features while the commercial intent likelihood even has its own category. If we however assume that the features presented in this work capture the majority of the information a certain feature category holds, we can treat them as atomic. Under that hypothesis we can use the insights gained to discriminate those web page aspects that contain most information of age-suitability.

Table 5 shows evaluation results per page segment. As in the IG comparison using one segment at a time, the page's title proved to be the most predictive one in general. This finding is very promising with respect to the on-line scenario. A computationally cheap classification based on web

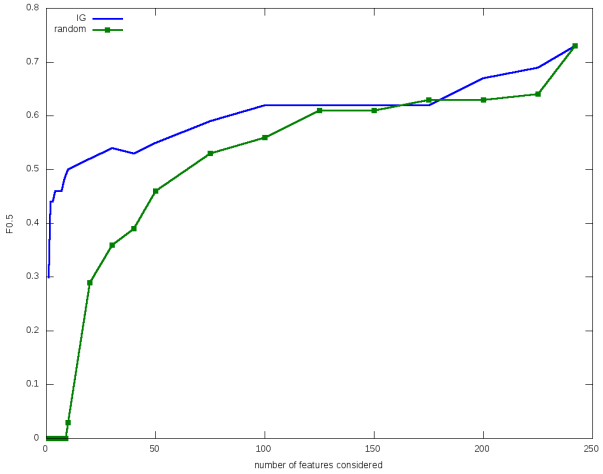

Figure 1: Performance of top n features ranked by Information Gain for increasing n

page titles and in case of low confidences backing off to using actual page content would constitute an efficient way of satisfying response time requirements.

Information Gain-based subsets

Previously we inspected the information theoretic importance per feature. To evaluate information theoretically motivated feature subsets we ranked features by their information Gain score constructing a feature subset out of the top n features for consecutively increasing values of n.

Figure 1 shows the performance of these subsets in terms of $F_{0.5}$ scores in comparison to a randomly drawn ranking of features. It can be observed that the early subsets constructed only of those features with the highest Information Gain scores significantly outperform randomly drawn subsets. As the number of features grows however the relative performance of the random sets approaches and even locally surpasses that of information theoretically selected features. This behaviour can be attributed to synergies between features not discovered by the current IG ordering.

Performance-based subset selection

Having shown that single feature information gain and subset selection by segment are capable of conserving most of the discriminative information while greatly reducing feature set size, we will now inspect the performance of automatically selected sets. Subset selection was done using the Weka library's Genetic Search approach [17].

In difference to the natural subsets along the boundaries of categories or segments and those constructed on information gain rankings this approach yields feature spaces which actually outperform the full set of features. This again supports the hypothesis that the suitability decision can hardly be made based on either only a single page aspect (category) or a set of strong individual features (IG). Evidently a well-balanced combination of different features yields best results. The overall strongest set of features reached an $F_{0.5}$-score of 0.8 using cross-validation and is described in Table 6. This 16-dimensional feature space can be extracted and clas-

Table 6: Best-performing feature subset

Kid link ratio	Number of words
Domain length	Total entities/unique entities
URL kid term score	Simple Wiktionary 1-gram LM
Script/word ratio	term freq "child" (headline)
Term frequency "kid"	number of words (title)
to-reference ratio	average word length (title)
Kid's pages 3-gram LM	OOV Academic (title)
Average word length	kid's 1-gram LM (title)

Table 7: Experiments with 10-fold cross-validation

Method	P	R	$F_{0.5}$	ROC
Random	0.50	0.50	0.50	0.50
Intuitive	0.82	0.05	0.21	0.56
LM	0.69	0.67	0.69	0.76
SVM	0.71	0.66	0.70	0.77
Classifier	0.85	0.64	0.80	0.83

sified far more quickly than the full range of features could while additionally yielding better performance through elimination of redundant and irrelevant dimensions.

When inspecting the resulting feature set we can notice that the general diversity of features is preserved. Most aspects of child-friendliness and focus towards children are present in form of members of the relevant feature categories. For all further experiments and evaluations in this work we will assume the use of this feature set.

5. CLASSIFIER EVALUATION

After having selected a promising classification method and the feature space in which to represent web pages, this section will measure our classifier's performance in comparison with several baseline methods.

5.1 Performance baseline

There has not been any specific research (known to us) on the classification of web pages specifically according to their suitability for different age groups. In order to evaluate our classifier's performance this section will introduce two baseline approaches from similar web tasks.

Intuitive approach

In the initial stage of our research we manually looked at different ways of how a human would tackle the task of finding child-friendly web content with the common means of information retrieval at hand. An intuitive approach is to use a conventional web search engine and expand the informational query by "kids", "children" or similar clue words. This expansion method obviously does not work satisfyingly all the time because web pages for children are by no means bound to actually mention them. On the other hand there will also be general pages talking about children while not being meant for a child audience. This method does however shift the focus of the results into the child-friendly domain to a certain degree. In order to exploit this intuitive notion our first baseline method will be to interpret the presence of the terms "kid", "child" and affixed forms thereof as an indicator of child suitability.

Table 8: Experiments on unseen sample

Method	P	R	$F_{0.5}$	ROC
Intuitive baseline	0.80	0.04	0.17	0.53
LM baseline	0.61	0.57	0.60	0.69
SVM baseline	0.63	0.60	0.62	0.70
Classifier	0.72	0.71	0.72	0.76
Human performance	**0.76**	**0.72**	**0.75**	**0.79**

Text classification approach

To construct a stronger baseline we apply an approach from text categorization as suggested by [23]. For this method we use unique terms as feature dimensions for an SVM classifier. Each dimension's value is the tf/idf-weighted term frequency. Stop words and terms that occur in less than 3 distinct documents within the collection are removed in order to keep the model from becoming too large. Our earlier findings on feature evaluation suggested language models to be powerful features. Because of term distribution statistics we expect this second approach to be a strong performance baseline.

Table 7 shows a comparison of the baseline methods and our classification approach. As expected the fairly naive intuitive method achieves high precision (most general pages will not mention children) at low recall (Children's pages are not bound to explicitly mention them). This coverage problem results in a worse than random F-score. The SVM-based text classification approach performs solidly. Our classifier that combines topical aspects expressed by language modelling approaches with non-topical notions of suitability achieves best performance for cross-validation.

5.2 Evaluation on unseen data

Our test collection is a set of 1800 web pages listed in the ODP containing 900 instances for children and 900 for adult audiences. The pages were randomly sampled from the English part of the directory (again excluding the *World* and *Regional* branches) and ensuring disjointness with the training data. Aside the ODP annotation, we had the test set additionally judged by external human annotators. The suitability decision is a highly subjective one. Using the overlap of several independent judgements will help us to reduce this degree of subjectivity. Furthermore we were able to collect information beyond the page's mere suitability, that we will use for a more fine-grained analysis. For each of the 1800 pages at least 5 independent human judgements were collected through the crowdsourcing platform CrowdFlower [2]. All further results reported refer to the performance on predicting the label assigned by the majority of CrowdFlower judges. 53% of the participants stated they were helping children with web search on a regular basis. An additional 33 % do so less frequently. 49.57% said to have children themselves. Based on these numbers we are confident that their judgements represent sensible notions of suitability.

Table 8 shows the performance of our classifier in comparison with the baseline methods as well as human judges for unseen test pages. The naive baseline approach that simply considers pages mentioning kids as suitable achieves high precision but due to the number of suitable pages not explicitly mentioning children its low coverage however makes it the weakest overall method. The SVM classifier consistently proved to be the next better approach. Our clas-

Table 9: Testing on pages exclusively for kids

Method	P	R	$F_{0.5}$	ROC
Intuitive	0.81	0.05	0.19	0.58
LM	0.66	0.58	0.63	0.71
SVM	0.68	0.61	0.66	0.73
Classifier	0.77	0.68	0.75	0.78
Human performance	**0.79**	**0.75**	**0.78**	**0.81**

Table 10: Testing on pages for young kids

Method	P	R	$F_{0.5}$	ROC
Intuitive	0.84	0.08	0.29	0.60
LM	0.70	0.63	0.68	0.75
SVM	0.73	0.67	0.72	0.77
Classifier	0.80	0.76	0.79	0.82
Human performance	**0.86**	**0.84**	**0.86**	**0.84**

sification method was able to perform significantly better than both baseline methods (a relative improvement of 14% over the strongest baseline) and came close to human performance. We were able to outperform both baseline methods at $\alpha < 0.05$ significance level. (Determined using Wilcoxon Signed Rank Test.) Inspecting human judgement behaviour gives further evidence for the task's high degree of complexity. We observed an agreement ratio of 68% for the suitability decision among independent CrowdFlower workers. The decision whether a web site deals with sports appears to be far easier to make than the one whether the page is suitable for a child of a given age. While the topical classification task is mainly an objective one, the suitability decision involves subjective factors such as personal understanding of children's needs and ethical aspects.

6. DETAILED PERFORMANCE ANALYSIS

Besides determining the actual classifier performance the goal of our research is to answer the following four research questions: 1) Is it easier to distinguish pages for age groups that are divided by a broad age margin? 2) Does the performance of our method depend on the web page's topic? 3) Does the page's quality have an impact on classification performance? 4) Can we make the suitability decision for pages independently of their language?

Age margin analysis

For our previous experiments we considered all pages that are suitable for children but at the same time might be interesting for teenagers. Now we will alter the test set by using only those pages that ODP considered exclusively for children. For this category we observed inter-judge agreement ratios of 71%. This finding supports the hypothesis that a bigger age margin between the classes makes the decision easier for humans. Table 9 shows the automatic approaches' performances for distinguishing pages exclusively for children from those for adults. Notice that only the target set for evaluation is restricted, and not the training process.

While this change in experiment set-up does not affect the ranking of methods for the task, it consistently raises performance of all approaches. To further increase the age margin between the classes we asked our CrowdFlower judges to additionally judge every page's suitability for very young children (aged 3-6). For this final experiment we rejected all kids' pages that were not deemed appropriate for young children. Table 10 shows that the task of distinguishing pages for very young audiences from general ones experiences another boost in performance. As an answer to our first research question, both experiments show how the information needs of different age groups become easier to discern for broad age group margins.

Table 11: Inter-annotator agreement by topic

Topic	Agreement
Kids_and_Teens/.../English/Grammar	1.00
Kids_and_Teens/Health/Safety/Fireworks	1.00
Society/Genealogy/Royalty	0.83
Kids_and_Teens/...Cartoons/Hello_Kitty	0.83
...	...
Kids_and_Teens/School_Time/.../Geography	0.59
Home/Family	0.55
Arts/Television/History	0.50
Kids_and_Teens/People/Biography	0.50
Kids_and_Teens/.../Math/Geometry	0.50

Topical analysis

An error analysis in the early stages of this work showed us not only that the suitability decision is often even difficult for humans to make, but also that there are topics that are in general harder to classify than others. We found that our classifier had particular problems with scientific and biographical content. Pages from this area even if they are deemed suitable for children often use a rather complex vocabulary and also focus more on pure information than nice presentation. One might argue that such pages will only be relevant to a small group of children who are really interested in such topics, but the challenge of correctly dealing with these pages remains. To get a more precise notion of our observation we analysed topical difficulty for human judges and our classifier. Page topics on which the inter-annotator agreement is very low imply a hard decision. Table 11 shows examples of topics that proved to be especially hard or easy. We can see that among the branches with high agreement scores we find typical children's / grown-ups' topics such as "English grammar" or "genealogy". Those branches with low agreement rates often belong to complex topics. Even if they were written and presented in a child-friendly way many annotators rejected them because they doubted the topic's general suitability for young audiences. Examples are "history", "geography" or "mathematics".

Since the same notion of topical difficulty that humans face might also apply for automatic approaches we determined the correlation between average inter-annotator agreement and our classifier's decision confidence. We found a Spearman Rank Correlation Coefficient $\rho = 0.58$ between the two dimensions. Figure 2 shows the distribution of classifier confidence scores and human agreement per page. To answer our second research question, we note that although it is a rather weak correlation, there is apparently an underlying topic-dependent difficulty that applies to humans as well as automatic approaches.

Inspecting the figure one can note a cluster of pages with high human agreement ratios (0.8-0.9) and at the same time

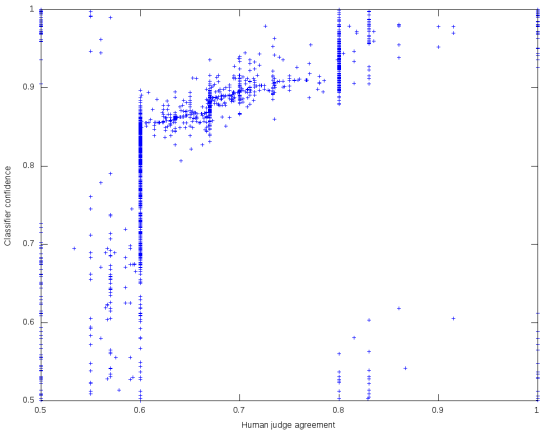

Figure 2: Correlation between human agreement on page suitability and classifier confidence

low classifier confidence (0.5-0.6). Manual analysis showed that the majority of these pages relied heavily on flash objects and images. Such pages are easy to classify for humans while our classifier finds hardly any accessible content. Dealing with pages with little to no textual content is clearly one of the future challenges in this domain.

Qualitative analysis

Previously we inspected the page topic's influence on prediction performance and compared to human judgements with respect to this. Another interesting idea to pursue is the notion of web page quality. During our manual study of ODP pages we found examples whose textual content might have been generally suitable for children but we hesitated because we did not find the page's layout appealing. The reason for this is a feeling that children should not be exposed to low quality resources. Although a web site's content might have been suitable for children many of the CrowdFlower workers rejected it stating that they disliked the confusing interface or the general impression the page made.

In order to quantify this intuition we asked the CrowdFlower workers to rate each page's quality on a scale ranging from 1 (lowest) to 4 highest quality. While quality of web pages is a highly subjective notion humans tend to have similar ideas of "good" and "bad" pages. The annotators' low standard deviation of quality judgements per page (0.79 quality grades) results in a coherent overall impression of page quality.

The first issue we will address is the connection of page quality and annotator agreement for that page. Intuitively we would assume that high quality pages are clear-cut specimen of their specific types and should therefore be easier to assign an age group to. Table 12 shows the agreement scores of human judges for web pages satisfying at least a given quality threshold. For very low quality threshold values the whole set of pages is considered. As we subsequently raise the threshold we note substantial rises in agreement levels.

Based on the previously shown correlation between annotator agreement and classification confidence we were interested in the influence of web page quality on classification performance. Considering the above results with respect to agreement scores we expected greater classification performance for higher quality pages. We can note that the classifier performance also increases for higher quality page sets. Both agreement and classifier performance rise steadily with quality. As an answer to our third research question we can see that while our classification method is robust to the quality of a page, it performs best on high quality pages. The same tendency holds true for human judgements.

Table 12: Quality-dependent performance

		Human	Classifier			
Qual	Share	Agreement	P	R	$F_{0.5}$	ROC
1	100%	0.68	0.72	0.71	0.72	0.76
2	95%	0.68	0.77	0.73	0.76	0.80
2.25	69.3%	0.70	0.77	0.75	0.77	0.80
2.5	51.7%	0.71	0.78	0.75	0.77	0.81
2.75	28.4%	0.73	0.79	0.75	0.78	0.81
3	11.4%	0.75	0.81	0.70	0.79	0.83
3.25	0.9%	0.82	0.90	0.82	0.88	0.90

Table 13: Language-independent performance

Language	# pages	P	R	$F_{0.5}$	ROC
English	2000	0.64	0.51	0.61	0.71
Chinese	1200	0.59	0.50	0.57	0.69
Dutch	1100	0.65	0.47	0.60	0.72
French	2000	0.64	0.46	0.59	0.70
German	2000	0.65	0.48	0.61	0.70
Russian	500	0.62	0.45	0.58	0.69
Spanish	2000	0.63	0.50	0.60	0.69
All	8800	0.63	0.48	0.58	0.70

Language-independent analysis

Finally it should be mentioned that one of the major limitations of our approach still is its heavy focus on textual features. While these have been shown to perform well, they force us to operate within the domain of a given language. Even shallow features like sentence length or general text length can easily become meaningless when crossing language boundaries between training data and on-line examples. Although we do not explicitly address this issue we were still interested in the performance level we could achieve on non-English resources with our current feature set. For assessment we relied solely on those features which are language independent as for example the visual, commercial or HTML features. We extracted a number of non-English web pages from the ODP and ran the reduced classifier trained on the original English set on it. Table 13 shows the distribution of languages amongst the web pages and our classifier's performance. For each language 50% of the pages originate from the ODP children's set and 50% from the ODP general set.

Regarding our fourth research question we conclude that even without any language-specific training we were able to reliably reach a minimum score of $F_{0.5} = 0.57$ over a set of very different languages. Further research in this direction might be dedicated to applying well-known language-specific scaling factors (e.g., a language's average word length) to use a wider range of features without having to retrain.

7. CONCLUSION

In this work we investigated the potential of a combination of topical and non-topical aspects for determining age suitability. Previous work on classification along web taxonomies either excluded age-dependent categories or reported comparably low result scores. We argue however that with appropriate consideration classifying pages according to their suitability for different age groups is not only possible but also highly desirable. Given the ubiquity of the Internet we should investigate how to make web search less frustrating for all users whose needs differ from those of core users.

We investigated aspects of child-friendliness and focus towards child audiences of web pages and introduced a wide range of features to capture them. After an inspection of single feature and feature subset performance we evaluated the most promising feature set against a hand-annotated test set of web pages. Our approach compares favourably to state of the art web classification techniques over which we could achieve significant improvements. Following our four research questions we explored the influence of age margin, topic and web page quality on classification performance. For all of these dimensions parallels between human agreement and classifier performance could be noted. As we initially suspected broadening the age margin between the categories eases classification. This emphasizes that age-categories are not just artificially motivated constructs but actual underlying properties of the Internet. Our fourth research dimension highlighted the feasibility of language-independent classification.

Future research on making web search more accessible to specific user groups might rely more strongly on visual content such as images, videos and flash animations. It becomes especially relevant when the user's cognitive and lexical abilities are taken into consideration. While adult users routinely draw information from textual resources, children who might not yet be able to read are far more attracted to the less demanding visual media. Being able to determine the suitability or perhaps even the "cuteness" of visual content would greatly improve the quality of web search for children.

Acknowledgements

We would like to thank Martha Larson for her comments and suggestions. This research is part of the PuppyIR project [3]. It is funded by the European Community's Seventh Framework Programme FP7/2007-2013 under grant agreement no. 231507.

8. REFERENCES

[1] Ask Kids. http://www.askkids.com, 2010.

[2] CrowdFlower - Harness the advantages of crowdsourcing. http://www.crowdflower.com, 2010.

[3] PuppyIR: An Open Source Environment to Construct Information Services for Children. http://www.puppyir.eu, 2010.

[4] The Open Directory Project - Kids & Teens. http://www.dmoz.org/kids_and_teens/, 2010.

[5] Yahoo! Kids. http://kids.yahoo.com/, 2010.

[6] Alias-i. LingPipe. http://alias-i.com/lingpipe, 2010.

[7] P.N. Bennett and N. Nguyen. Refined experts: improving classification in large taxonomies. In *SIGIR 2009*.

[8] J. Callan and M. Eskenazi. Combining lexical and grammatical features to improve readability measures for first and second language texts. In *Proceedings of NAACL HLT*, 2007.

[9] S.L. Calvert. Children as consumers: Advertising and marketing. *The Future of Children*, 2008.

[10] C. Castillo, D. Donato, A. Gionis, V. Murdock, and F. Silvestri. Know your neighbors: Web spam detection using the web topology. In *SIGIR 2007*.

[11] K. Collins-Thompson and J. Callan. A language modeling approach to predicting reading difficulty. In *Proceedings of HLT/NAACL*, volume 4, 2004.

[12] H.K. Dai, L. Zhao, Z. Nie, J.R. Wen, L. Wang, and Y. Li. Detecting online commercial intention (OCI). In *WWW 2006*, page 837. ACM.

[13] L. Feng. Automatic readability assessment for people with intellectual disabilities. *ACM SIGACCESS*, (93), 2009.

[14] L. Feng, N. Elhadad, and M. Huenerfauth. Cognitively motivated features for readability assessment. In *EACL*, pages 229–237. ACL, 2009.

[15] E. Gabrilovich and S. Markovitch. Harnessing the expertise of 70,000 human editors: Knowledge-based feature generation for text categorization. *Journal of Machine Learning Research*, 8:2297–2345, 2007.

[16] K. Golub and A. Ardo. Importance of HTML structural elements and metadata in automated subject classification. *ECDL 2005*, pages 368–378.

[17] 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):10–18, 2009.

[18] G.R. Klare. The measurement of readability: useful information for communicators. *ACM Journal of Computer Documentation (JCD)*, 24(3):121, 2000.

[19] P. Kolari, T. Finin, and A. Joshi. SVMs for the blogosphere: Blog identification and splog detection. In *AAAI Spring Symposium on Computational Approaches to Analyzing Weblogs*, 2006.

[20] A. Large, J. Beheshti, and A. Breuleux. Information seeking in a multimedia environment by primary school students* 1. *Library & Information Science Research*, 20(4):343–376, 1998.

[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] B. Liu, M. Hu, and J. Cheng. Opinion observer: Analyzing and comparing opinions on the web. In *WWW 2005*.

[23] T.Y. Liu, Y. Yang, H. Wan, H.J. Zeng, Z. Chen, and W.Y. Ma. Support vector machines classification with a very large-scale taxonomy. *ACM SIGKDD Explorations Newsletter*, 7(1):43, 2005.

[24] G.H. McLaughlin. SMOG grading: A new readability formula. *Journal of reading*, 12(8):639–646, 1969.

[25] S. Naidu. Evaluating the usability of educational websites for children. *Usability News*, 7(2), 2005.

[26] Jakob Nielsen. Kids' corner: Website usability for children. http://www.useit.com/alertbox /children.html, May 2010.

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

[28] X. Qi and B.D. Davison. Web page classification: Features and algorithms. *ACM CSUR II 2009*.

[29] J. Schacter, G.K.W.K. Chung, and A. Dorr. Children's Internet searching on complex problems: performance and process analyses. *JASIST*, 49(9):840–849.

[30] S. Schwarm and M. Ostendorf. Reading level assessment using support vector machines and statistical language models. In *ACL 2005*, volume 43.

[31] E.A. Wartella, E.A. Vandewater, and V.J. Rideout. Introduction: electronic media use in the lives of infants, toddlers, and preschoolers. *American Behavioral Scientist*, 48(5):501, 2005.