Web Page Classification on Child Suitability

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

Children spend significant amounts of time on the Internet. Recent studies showed, that during these periods they are often not under adult supervision. This work presents an automatic approach to identifying suitable web pages for children based on topical and non-topical web page aspects. We discuss the characteristics of children's web sites with respect to recent findings in children's psychology and cognitive sciences. We finally evaluate our approach in a large-scale user study, finding, that it compares favourably to state of the art methods while approximating human performance.

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

Web Search, Children, Suitability, Filtering, Classification

1. INTRODUCTION

In recent years children's age of first contact with the Internet has decreased significantly [24]. At the same time their overall Internet consumption is growing. This tendency is accompanied by a growing number of children who, even at very young ages, search the Internet without any form of adult supervision. A recent UK study [21] found that up to 40% of British children aged 5-15 years regularly access the Internet without parental guidance. Regularly exposing children to potentially unsuitable web resources can become a significant danger to their well-being and development. Although popular web search engines such as Google and Yahoo! offer a wide range of "safe search" settings to protect users from confrontation with undesired content, most of them state in their terms of service that their search functionality should not be used by minors.[1]

State of the art search engines could greatly benefit from an automatic means of identifying suitable web pages. The current means of child-friendly information access are typically listings of manually selected resources. Examples of this type of hand-picked web directories are Yahoo! Kids [5] and Ask Kids [2]. Because of the high demand for manual labour in its maintenance, this approach is less flexible and has a lower coverage than an automatic method could achieve. It is important to note that ensuring suitability of web pages exceeds merely filtering out erotic material. Well-studied topical approaches are capable of identifying such web pages reliably. Notions of text difficulty and age-appropriate web site design are however largely independent of the page topic and should strongly contribute to the decision of showing a particular page to a child. In this work we will demonstrate how the use of non-topical web page aspects can enhance state of the art topical methods in making the suitability decision.

The contributions of this work are threefold: 1) We identify the criteria of good children's web sites and show how to encode them in features. 2) We conducted a large-scale user study to measure human performance for web page suitability assessment. The results are used to give a frame of reference to an automatic approach's performance evaluation. 3) We describe an automatic web site classification method, showing how purely topical models can be outperformed by models augmented by non-topical features.

The remainder of this work 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, that are combined in the classification scheme described in Section 4. Section 5 presents a comparison of our method against a state of the art approach as well as human performance. Finally, Section 6 concludes the paper with a discussion of our experimental results as well as their implications on future work in the field of accessible search.

2. RELATED WORK

To the best of our knowledge, there has not been any prior research on determining a web page's suitability for children. There are however various threads of work in closely related fields that should be mentioned. Since the early 1950's linguists have devised formal means of assessing a given text's

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

CIKM'10, October 26–30, 2010, Toronto, Ontario, Canada.

reading level [14]. These early readability measures are typically linear combinations of several shallow textual features. In 2004, Collins-Thompson and Callan described a language modelling approach to measure readability [9]. Subsequently the research community focused on more and more sophisticated ranges of features [7, 22, 10], using machine learning techniques such as SVMs [23].

Previous work on topical classification using the Open Directory Project's large-scale web taxonomy reported the "Kids and Teens" branch of the directory [4] to be among the most challenging categories with respect to classification [6, 20]. Some research projects even excluded the category from their scope [12], as they could not find a homogeneous underlying topical distribution. This tendency is already a first indicator that purely topical classification is not appropriate to distinguish web resources for children from those for adults.

Non-topical classification has been an active research field with advances in weblog identification [15], spam detection [8] and sentiment analysis [17]. In this work we will show that neither readability measures nor language modelling, when used in isolation, are strong predictors of a web page's suitability for children. We present a more appropriate method that considers a combination of topical and non-topical web page aspects.

3. FEATURE DESIGN

In this section we will discuss the cognitive specifics of children and how they can be encoded in features for classification. We investigate suitability along two dimensions, *child-friendliness* and *focus towards child audiences*.

Child-friendliness

The first dimension we will inspect with respect to suitability of web pages for children is child-friendliness. Its core concerns are ensuring children's safety while providing a frustration free search environment [19, 16]. Child-friendliness of web pages is expressed in the page's complexity of text, its presentation style as well as navigational aspects.

Textual complexity

Shallow features (12)[2]**.** Children's text is commonly syntactically simpler than text for adults. To measure this we use a range of shallow textual features such as the total number of words in the text or the average word length.

Readability (8). An alternative approach towards measuring the syntactical complexity of a text computes several well-known readability scores such as SMOG or ARI [14] and uses them as suitability features.

Parts of speech (5). A more high-level notion of the difference in syntactical complexity between web sites for adults and those for children is expressed through this feature category. POS parse statistics are employed in the form of e.g., the average number of proper nouns or adjectives per sentence.

Named entities (6). Typically, children's texts are also semantically simpler than those for adults. This is for example expressed through the lower number of named entities per sentence in children's stories [11]. We detect the entity types *person*, *location* and *organization* and use their counts as features.

Out of vocabulary rates (8). In addition to the previously mentioned aspects of text simplicity, children's texts typically also use a more basic choice of words. To reflect this, we built 7 dictionaries of basic English words to compute the OOV rates of web pages. Adult pages are assumed to yield higher shares of unknown terms than children's pages. An additional dictionary of academic terms is created for which the opposite tendency is expected.

Wiktionary features (4). An alternative way of measuring textual complexity makes use of the Wiktionary free on-line dictionary. Basic terms are assumed to have short, unambiguous definitions. For each word the number and size of the definitions in Wiktionary are looked up. The page-wide averages are reported as features.

Presentation

Child-friendly web sites do not only display simpler content, they should also present it accordingly. Recent studies pointed out how important appealing presentation and general fun were for children's web portals [16].

HTML features (10). In this category, we inspect the distribution of various HTML elements such as scripts and animations. They are assumed to occur with significantly different frequencies on children's and adult pages.

Visual features (8). Due to their comparably low lexical skills children often prefer visual content over textual resources. This category collects counts and size distributions of web page images and reports them as features.

Navigation

Neighbourhood analysis (2). Topical web classification tasks have been widely shown to benefit from an analysis of a page's link neighbourhood. A page is obviously not bound to contain only topically related links. However, we expect this tendency to hold true for suitability. A good web portal for children will not link to adult pages. We use our classifier to determine each linked page's suitability and report the share of pages that were classified as adult/kid with a given threshold confidence $c_{threshold}$.

Focus towards child audiences

During our manual assessment of web pages in the course of this work we frequently encountered pages that would classify as child-friendly according to the previously discussed features but that did not convey the impression of having been designed for children specifically. To account for these pages we introduce the second dimension of suitability, focus towards child audiences.

Language models (9). Language models are widely accepted predictors of topical affiliation. We built several models of different order (word-internal character trigram, unigram and trigram) of ODP children's web sites (ensuring disjointness with training and test sets), as well as textual resources from simple Wikipedia (http://simple.wikipedia.org)

[2]The numbers in brackets refer to the number of features in each category.

and simple Wiktionary (http://simple.wiktionary.org). 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 web page is scored against these models and the resulting language model scores are reported as features. This method is able to detect topical patterns that occur frequently on children's web pages.

Reference analysis (2). To identify children's web sites we investigated the occurrences of clue words such as "kid" or "child". We found that many pages mentioning children were actually meant for child audiences. There was however a significant number of educational pages, targeting parents and educators. Those pages deal with children as a subject rather than as an audience. In order to distinguish these two types of pages, we analysed text windows around clue word occurrences and created n-gram distribution statistics. We expect to find a different way of referring to children on adult pages where children are **talked about** (e.g., "your child" or "the average child") and children's pages, where they are **talked to** (e.g., "us kids"). We use the relative share of *about-mentions* and *to-mentions* on a page 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). Recent studies [16] found, that a page's URL and domain name play a significant role in how well the page will be received by child audiences. Children typically struggle with remembering long and complex domain names. To account for this fact we measure features such as URL length or the maximum likelihood estimate of possible URL terms according to Wiktionary.

Page segmentation

Previous research [13] found splitting web pages into segments beneficial for classification. We follow this idea by segmenting each page into title, headlines, anchor texts and main textual content. The previously introduced features are computed for each of these segments, (HTML, visual and neighbourhood features are still only considered globally per page.) thus greatly expanding the feature space from 79 to 241 dimensions. An exhaustive overview of all considered features can be found at http://blackboard.tudelft.nl/ bbcswebdav/users/ceickhoff/ features.xls.

Table 1: 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)

4. CLASSIFICATION SCHEME

Training collection

Our training data was acquired from the ODP [4]. The directory's "Kids and Teens" branch contains non-exclusive age labels that identify a web page's suitability for kids (up to 12 years), teens (13-15 years), mature teens (16-18 years) and adults. In order to keep a broad age margin between children's web pages and those for general audiences we excluded the pages for teenagers and mature teenagers. Since several of the previously introduced features are language-dependent, we also exclude the "World" and "Regional" branches of the directory as these cover mostly non-English resources. The resulting training corpus contains 20,778 web pages. (6,225 for children and 14,553 for general audiences.)

Classifier design

After an initial performance comparison of several state of the art learning methods on the training set we decided for a Logistic Regression model to combine our features. To reduce computational complexity we used an evolutionary method to evaluate feature subsets that would retain the predictive power and feature diversity while projecting the data into a lower-dimensional space. Due to the high dimensionality of the original feature space an exhaustive search was not feasible. The heuristically determined optimal set is described in table 1.

We can note, that while features from most categories are included, the majority of them originate from either the full page or its title, making these two segments the most important sources of information of page suitability.

5. PERFORMANCE EVALUATION

Test collection

Our test collection is a set of 1800 real web sites (900 for children and 900 for adults) from the ODP. They were randomly sampled from all categories ensuring disjointness with the training data. All non-English web sites were excluded. We asked human judges to annotate the collection through the crowdsourcing platform CrowdFlower [3]. In order to overcome subjectivity in these labels we collected 5 independent judgements for each page and assigned the major-

Table 2: Experiments on unseen sample

Method	P	R	$F_{0.5}$	ROC
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**

ity decision. The results of this user study give a frame of reference to the previously unstudied task of identifying web sites that are suitable for children.

Performance baseline

To measure our method's performance we will compare it to a state of the art approach of topical classification. We follow Liu et al. [18] in using a Support Vector Machine classifier (C-SVM, radial basis kernel, cost = 1, $\epsilon = 0.001$, $\gamma = 0.01$) with unique terms as dimensions and their tf/idf-weighted counts as values. In order to limit computational complexity and reduce data noise we only considered those terms that occurred in at least 3 distinct training documents.

Performance comparison

The final performance of our classification method was determined by a single run against the previously unseen test set. The evaluation was done in terms of precision, recall as well as their harmonic combination in the $F_{0.5}$-measure. We decided for the precision-biased version of the F-measure since in a filtering scenario an unsuitable page being shown to a child should be penalized more strongly than a missed children's page. In order to give an impression of the classifiers judgement confidence we additionally report the area under the ROC curve for each method. Table 2 shows our results in comparison with both the text classification baseline as well as the majority labels assigned by human annotators. The exclusively topical SVM performs solidly and provides correct predictions most of the time. Our combined topical/ non-topical method however was able to outperform this baseline at $\alpha < 0.05$ significance level. (Determined using a paired two-sided Wilcoxon Signed Rank Test.) We could achieve an improvement of 14% over the SVM baseline while approximating human performance.

6. CONCLUSION

In this work we presented a combined topical/ non-topical approach to determining the suitability of web pages for children. Previous work on topical classification along web taxonomies either excluded age-dependent categories or reported low result scores. We show that with careful consideration a topical approach augmented by non-topical features is able to reliably predict web page suitability. Our method achieved significant improvements over a state of the art topical classification model while approximating human performance. An especially challenging task for our classifier was to decide upon the suitability of pages that relied heavily on graphical elements such as Flash animations while not providing accessible textual content. Future work on web site suitability should pay close attention to analysing actual image content in terms of image suitability or even image "cuteness" in order to overcome this problem.

Acknowledgements

This research is part of the PuppyIR project [1]. It is funded by the European Community's Seventh Framework Programme FP7/2007-2013 under grant agreement no. 231507.

7. REFERENCES

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

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

[3] CrowdFlower. http://www.crowdflower.com, 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] P.N. Bennett and N. Nguyen. Refined experts: improving classification in large taxonomies. In *SIGIR 2009*.

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

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

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

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

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

[12] 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.

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

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

[15] 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.

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

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

[18] 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.

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

[20] A. Ntoulas, G. Chao, and J. Cho. The infocious web search engine: Improving web searching through linguistic analysis. In *WWW 2005*, pages 840–849.

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

[22] E. Pitler and A. Nenkova. Revisiting readability: A unified framework for predicting text quality. In *EMNLP 2008*.

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

[24] 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.