SEVENTH FRAMEWORK PROGRAMME
IST FP7 231507

D3.3 Report on information extraction and mining techniques

File name	PuppyIR-D3.3-ReportOnInformationExtractionAndMiningTechniques-v1.0.pdf
Author(s)	Karl Gyllstrom (KUL), Carsten Eickhoff (TUD), Jan De Belder (KUL), Sergio Duarte (UT)
Work package/task	D3.3
Document status	final
Version	1.0
Contractual delivery date	30 September 2010
Confidentiality	Public
Keywords	information extraction, Mechanical Turk, child-appropriateness, queries, language simplification
Abstract	Describes research within WP3 toward information extraction. Covers web page classification, query log analysis, and techniques for text-simplification and re-writing.

Table of Contents

Executive Summary

This report covers the work of the WP3 package toward the goal of automated extraction of information for use in children's search engines.

We cover 4 main bodies of work:

First, we explore two applications that enable the automated classification of child-appropriateness among children's web pages. In one approach, a feature-based classifier is applied to a number of features on children's web pages. In another approach, link-analysis is used to find children's web pages by following links among a set of known child-friendly pages to web pages for which no child-appropriateness assessment is known. Both of these approaches are validated against a large collection of human ratings of child appropriateness.

Second, we explore a method to extract queries – from an existing query log – that are deemed likely to be from or on behalf of children. This approach applies some heuristics such as whether or not a query results in the user clicking on a known or presumed children's page. It is validated against some query-session level behavioral statistics that have been produced by the community of researchers in children's information interaction.

Finally, we describe a technique to automatically simplify text to make it more appropriate to child audiences. This method is particularly useful to make web pages that are otherwise not child-friendly accessible to children.

In our conclusion, we describe ways in which these methods can be applied in a children's search engine. In fact, some of these techniques have already been integrated in a search engine prototype, whose implementation is ongoing, and whose release is forthcoming.

1 Introduction

We report the contributions among researchers within WP3 toward the end of information extraction. We separate discussion into 3 main areas:

First, two main research lines have covered the process of web page classification. The goal of classification is to determine the suitability of web pages for children. Though an obvious application in children's search is to identify inappropriate content (e.g., pornography), classification can also be used to distinguish among general web pages such that pages that are more interesting, attractive, readable, etc., for children are identified. For example, consider the Wikipedia article for *Superman*[1], a page for the query "superman" that is highly ranked by Google and Bing. Although pertaining to a subject of appeal to children, the article contains words and phrases that require a relatively advanced reading level – such as "distinctive and iconic", and "imbued with a strong moral compass" – and contains relatively complex topics such as "Copyright issues" and "Literary analysis". Conversely, the page from the *simple language* domain of Wikipedia[2] is much more appropriate in terms of both reading level and topic. The goal of our classification is to provide the ability for a search engine to promote the more child-friendly page to young users. This topic is covered in two sections. The first describes *feature-based classification* (Section 2). The second describes *link-based classification* (Section 3).

The next topic involves the extraction of children's queries from generic query logs. Children's queries are elusive because there is no existing data set and gathering naturalistic queries from children remains difficult at this stage. Though it is feasible that queries might be collected by the PuppyIR project at a later date, the current work created a virtual children's query log by extracting – from a generic query log – queries that, for various reasons, exhibit a high likelihood of being for children. This work is covered in Section 4.

Finally, we describe the work of text-simplification and rewriting. This work describes automated methods to convert text from its original form to a simpler form that may be more readable to young audiences. This work provides a complement to the classification work mentioned above, as it can be applied in cases where classification is unable to produce good results (e.g., for a very recent news event, there may not yet be child-friendly pages covering the event). This work is covered in Section 5.

In the original description of work, we mentioned interest in exploring other areas such as entity recognition and social network analysis. We maintain interest in these areas but focused on the work reported here for practical reasons. In particular, goals the query assistance deliverable (D3.2) matured into turning it into a complete search tool for children that may soon be publicly available. Being able to use this tool on publicly available – and unlabeled – web pages was seen as a core objective and this motivated the works described in Sections 2 and 3. The work in summarization and simplification described in Section 5 has an obvious role in a search engine as it has the potential to improve web page surrogates (i.e., the summary snippets viewed in web search results) as well as expand upon the available web pages discovered by automated processes, being able to translate less child-friendly pages to more accessible language.

The work described in this document has been published at numerous venues in the field. The following are directly described in this document:

Section 2

- C. Eickhoff, P. Serdyukov, and A. P. de Vries. Web Page Classification on Child Suitability. In *Proceedings of the 19th International Conference on Information and Knowledge Management (CIKM)*. New York, NY, USA, 2010. ACM.

[1] `http://en.wikipedia.org/wiki/Superman` (May 2010)

[2] `http://simple.wikipedia.org/wiki/Superman`

- C. Eickhoff, P. Serdyukov, and A. P. de Vries. A Combined Topical/Non-topical Approach to Identifying Web Sites for Children. In *Proceedings of the 4th International Conference on Web Search and Data Mining (WSDM), Hong Kong, China*. ACM, 2011.

Section 3

- K. Gyllstrom and M.-F. Moens. Wisdom of the ages: toward delivering the children's web with the link-based AgeRank algorithm. In *Proceedings of the 19th International Conference on Information and Knowledge Management (CIKM)*, New York, NY, USA, 2010. ACM.

Section 4

- S. Duarte Torres, D. Hiemstra, and P. Serdyukov. An analysis of queries intended to search information for children. In *IIiX '10*, pages 235–244, New York, NY, USA, 2010. ACM.
- S. Duarte Torres, D. Hiemstra, and P. Serdyukov. Query log analysis in the context of information retrieval for children. In *SIGIR '10: Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 847–848, New York, NY, USA, 2010. ACM.

Section 5

- J. De Belder and M.-F. Moens. Text simplification for children. In *Workshop on Accessible Search Systems*, Geneva, 2010.

The following papers address related work from WP3 that are not directly covered in this document:

- K. Gyllstrom and M.-F. Moens. A picture is worth a thousand search results: finding child-oriented multimedia results with collage. In *SIGIR '10: Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval*, pages 731–732, New York, NY, USA, 2010. ACM.
- C. Eickhoff and A. P. de Vries. Identifying Suitable YouTube Videos for Children. In *Proceedings of the 3rd Networked & Electronic Media Summit (NEM), Barcelona, Spain*, 2010.
- J. De Belder, K. Deschacht, and M.-F. Moens. Lexical simplification. In *1st International Conference on Interdisciplinary Research on Technology, Education and Communication (ITEC)*, Kortrijk, Belgium, 2010.
- J. De Belder and M.-F. Moens. Integer linear programming for dutch sentence compression. In A. Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, volume 6008 of *Lecture Notes in Computer Science*, pages 711–723. Springer Berlin - Heidelberg, 2010.

2 Finding Children's Web Pages I: Feature-based Classification

The PuppyIR DoW section 1.2.2 identified a need for means of moderated discovery targeted towards children's specific requirements in order to alleviate the effort of monitoring and selecting the subset of data sources that can be returned safely as search results for children. In this section we will summarise our method to using non-topical web page aspects to enhance state of the art topical classification methods in making the suitability decision.

2.1 Overview

Suitability was investigated along two distinct dimensions, a page's *child-friendliness* and its *focus towards a child audience*. Child-friendliness is essentially concerned with ensuring children's safety while providing a frustration free search environment [30, 33]. Child-friendliness of web pages can be encoded using three feature categories:
(1) The page's complexity of text. Following the hypothesis that children's texts are usually simpler than those for adults, we investigated various reading level scores, syntactical analyses and text's usage of difficult terms.
(2) The page's presentation style. 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 [30]. Presentation was considered in terms of the employed HTML elements and their distribution as well as the number and size of on-page pictures.
(3) Navigational page aspects. A good web portal for children will not link to adult pages. We used our classifier to determine each linked page's suitability and reported the share of pages that were classified as adult/kid with a given threshold confidence.

Our second dimension of suitability, focus on child audiences, was introduced to further differentiate between adult pages that are generally harmless for children (child-friendly ones) and actual children's pages. Three distinct feature categories were used to identify those web pages that were specifically designed for children.
(1) Language models. Language models are widely accepted predictors of topical affiliation. We built several models of different order (unigram, trigram and word-internal character trigram) trained on confirmed collections of children's text. Smoothing was applied in well-known Jelinek-Mercer fashion.
(2) Reference analysis. 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 teaching professionals. 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. The essential intuition behind this method was to differentiate between 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").
(3) URL features. Recent studies [30] 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.

Previous research [21] 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

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)

from 79 to 241 dimensions. An exhaustive overview of all considered features can be found in Table 6 in the appendix.

2.2 Evaluation and Results

Our research corpus consisted of 25000 web pages from the Open Directory Project [3]. Training and parameter tuning was exclusively done on this data set. A wide range of state of the art machine learning techniques was considered and trained on varying subsets of the feature space. The best-performing method was a multinomial logistic regression model (Ridge estimation parameter $\rho = 10^{-8}$) trained on the feature set detailed in Table 1.

A non-overlapping additional set of 1800 web pages was manually annotated through the crowdsourcing platform CrowdFlower [2] and was used for final evaluation. To measure our method’s performance we will compare it to a state of the art approach of topical classification. Our baseline is 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 as a metric 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 classifier’s 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.

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**

3 Finding Children's Web Pages II: Link-based Classification

The research described in this section addresses a similar goal to the classification approach mentioned in the previous section, though it adopts a complementary approach. The specific goal in this work – titled AgeRank – is an automated method to assign a child-appropriateness score to web pages such that pages can be compared to one another in ranking contexts, or filters can be applied such that pages below a certain threshold are withheld from child searchers [22].

In the design of AgeRank, we adopted a link-based label-propagation approach, where we exploit age-level locality among pages. Specifically, we hypothesize that a page designed for children is more likely to link to and be linked from other pages designed for children than to link to or be linked from pages designed for adults. As a link-based approach, it is simple, fast, and highly scalable in a manner akin to PageRank. It produces single scores for pages, meaning it is modular and can easily be integrated into existing ranking functions. We view AgeRank as a complement to other possible approaches, such as feature-based classification.

3.1 Overview

We hypothesize that web pages for children exhibit locality; specifically, we hypothesize that children's pages are more likely to link to, and be linked from, other children's pages, than pages for adults. This hypothesis motivates the AgeRank algorithm.

Let us present our terminology. A web graph is a directed graph $G = (P, L)$ where P are pages and L are the directed links among pages. The functions $outlinks(p)$ and $inlinks(p)$ define the sets of links pointing outwardly and inwardly from p, respectively.

For AgeRank, we adopt a label-propagation approach [44]. We have a set of positively labeled pages $L_+ \subset P$ and a set of negatively labeled pages $L_- \subset P \setminus L_+$. From L_+ and L_- we seek to propagate labels outward to pages that link to or are linked from them. In essence, this expands the available evidence of a page being for children (via proximity to L_+) or for adults (via proximity to L_-).

The AgeRank algorithm assigns four scores to pages $p \in G$, corresponding to two positive and two negative scores. Scores are separated into inward and outward, indicating whether the label was propagated from a page linking to p or from a page to which p links. We represent this as a 4-tuple $(P_{out} \times P_{in} \times N_{out} \times N_{in})$. P_{out} represents the amount of positive score that a page receives from its outgoing links. For example, if page p_i links to page p_j, p_i will receive some score from p_j by virtue of linking to it. P_{in} represents the amount of positive score that comes from incoming links. N_{out} and N_{in} represent the negative score that comes from incoming and outgoing links, respectively. These scores indicate the degree to which the page is related to positively and negatively labeled pages. Pages in L_+ have scores of $(1, 1, 0, 0)$, and pages in L_- have scores of $(0, 0, 1, 1)$. All other pages are initialized with scores $(0, 0, 0, 0)$. We separate the propagation of labels outwardly and inwardly. The outward propagation g_o (i.e., score received from outward links) is as follows, where pol indicates the polarity of the score (e.g., positive corresponds to P_{out}, and negative corresponds to N_{out}):

$$g_o(p, pol) = \frac{1}{|outlinks(p)|} \times \sum_{p_j}^{outlinks(p)} \frac{g_o(p_j, pol)}{|inlinks(p_j)|}$$

The propagation for inward scoring is g_i is:

$$g_i(p, pol) = \frac{1}{|inlinks(p)|} \times \sum_{p_j}^{inlinks(p)} \frac{g_i(p_j, pol)}{|outlinks(p_j)|}$$

As the above show, the amount of score that is transferred outwardly is divided by the number of outward links from the propagating pages times the number of inward links from the receiving page (and vice-versa for inward transfer). This is a simple similarity score between two pages across a directional link, indicating the exclusivity of the link relationship. Conceptually, the more outgoing links of p_i, the less likely any particular one is especially meaningful (this assumption has been made elsewhere, e.g., [23]). Since we are using links as measures of commonality in age-appropriateness, fewer links to p_i indicate that the pages linking to p_i are individually more revealing of the relationship.

As a recursive algorithm, AgeRank is run iteratively. Each iteration i represents the extent of label propagation after i hops. We employ clamping [44], meaning each page $p \in L_+ \cup L_-$ retains its original label scores after each iteration. We define the label for a page at iteration i as:

$$label(p) = \begin{cases} 1, 1, 0, 0 & p \in L_+ \\ 0, 0, 1, 1 & p \in L_- \\ g_i(p, +), g_o(p, +), g_i(p, -), g_o(p, -) & o.w. \end{cases}$$

Consider the web graph example (a) depicted in Figure 1, where page A is from L_+. At the initial state, A has values $(1, 1, 0, 0)$. After iteration 1, B's score is $(0, \frac{1}{2}, 0, 0)$, since the similarity score is $\frac{1}{2}$ since B has two incoming links. At iteration 2, D's score is $(0, \frac{1}{2}, 0, 0)$; since its similarity to B is 1, its incoming value is the same as B. Note that C does not change because P_{in} cannot travel out incoming links. At iteration 3, E and F both have the score $(0, \frac{1}{4}, 0, 0)$ because their similarity scores with D are both $\frac{1}{2}$ since D has two outgoing links.

We combine these scores into a single AgeRank score called Tot as follows:

$$Tot = (1 + \frac{(P_{out} + P_{in}) - (N_{out} + N_{in})}{P_{out} + P_{in} + N_{out} + N_{in}}) \times \frac{1}{2}$$

The Tot score captures the ratio of the positive scores to the negative scores, with higher relative positive values yielding a higher Tot score. The reason a simpler approach like $\frac{P_{out}+P_{in}}{N_{out}+N_{in}}$ is not used is to avoid division-by-zero, as any or all values can be 0. The AgeRank score, as defined by Tot, ranges from 0 (absolute adult) to 1 (absolute child).

Note that our approach separates propagation between outward and inward links. For example, if page A links to page B, the propagation will only allow P_{out} to transfer from B to A, and not P_{in} (although P_{in} would transfer from A to B). This is to prevent feedback. For example, consider page A, which links to page B, where the score of A is greater than B. After one iteration, B's score rises due to A's link to it.

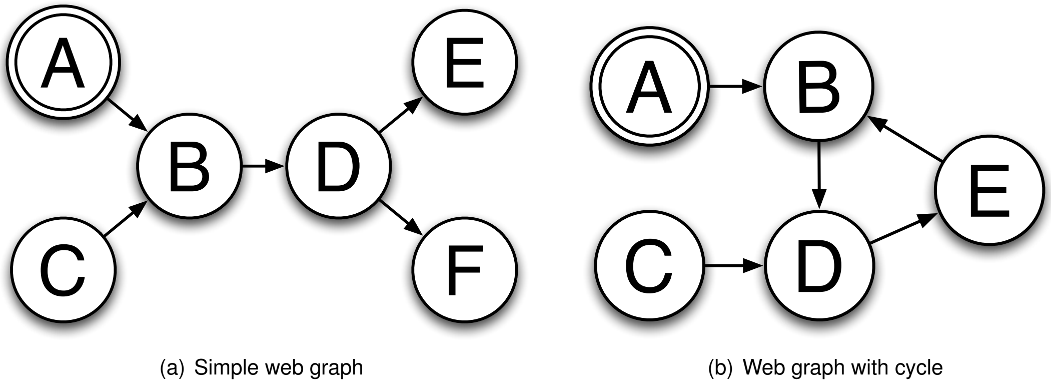

Figure 1: Two simple web graphs. For both graphs, page A is in L_+.

Since B's score rose, on the next iteration A will draw a higher incoming score from B, which causes A's score to rise; this problem will continue with each iteration.

Though feedback is not tolerated, cycles can occur; this is an inescapable property of web graphs. We define feedback as a continuous (and erroneous) push of a value toward 0 or 1. Cycles are not problematic due to the following. First, feedback toward 0 does not happen because, due to the clamping, a page's individual score (e.g., P_{out}) score can never decline after an iteration of AgeRank because L_+ and L_- retain and propagate their scores on every iteration. To understand the problem of feedback toward 1, consider the cycle $A \rightarrow B \rightarrow C \rightarrow A$, with A having P_{in} $P_{in}(A) > max(P_{in}(B), P_{in}(C))$. Feedback toward 1 means that P_{in}(A) would grow larger in successive iterations due to the cycle. However, our use of the label propagation is attenuating, in that the score transferred from a page can be at most the amount that was transferred to the page. Hence, the highest score that can circulate to A *from* A would be A's original score.

Consider the web graph example (b) in Figure 1, where page A is in L_+. After iteration 1, B acquires a P_{in} score of $\frac{1}{2}$, since it has two incoming links. In the next two iterations, D, then E acquire P_{in} scores of $\frac{1}{4}$. On the next iteration, B's score rises to $\frac{5}{8}$ and this addition circulates in successive iterations, although the relative change is lower with each iteration. B's score converges to $\frac{2}{3}$, while D's and E's converge to $\frac{1}{3}$. Note that a modification of the graph such that C were removed means that B, D, and E would converge toward 1. We consider this to be correct behavior, since the absence of any other links means they all share a similarity to A of 1.

3.2 Evaluation and Results

This work has been extensively evaluated [22]; here we provide a summary. AgeRank was run on a set of positively and negatively labeled pages that were drawn from DMOZ [1], producing over 11.5 million labeled pages from $\sim 30k$ positively labeled pages and $\sim 300k$ negatively labeled pages.

With respect to human assessments of the child-appropriateness of web pages (depicted in Table 3), we found significant correlations among P_{in} and P_{out} (positively correlated with AgeRank scores), and N_{out} (negatively correlated with AgeRank scores), while N_{in} had a negative though non-significant correlation. These provide assurances of the effectiveness of AgeRank through explicit ratings.

We further studied the coverage of pages for which our methods produced child-apprорriate ratings. First, the overlap with the 7 million most highly PageRanked pages in ClueWeb09 was as high as 91.4% among the top 100,000 PageRanked pages, and 19.9% at 7 million. This shows that our method

Score	Slope	P-value
P_{out}	0.3547	0.0008
P_{in}	0.4131	0.0060
N_{out}	-0.2064	0.0079
N_{in}	-0.0939	0.2204
Tot	0.4172	0.0000

Table 3: Linear regression of various AgeRank scores on human scores and DMOZ on human scores.

is producing labels for a large amount of the most important web pages.

Through some applied search experiments, we showed that AgeRank labels are also being found for search results across a diverse range of queries, visible in over 14% of the top 1000 web results across all queries, and appearing in 99% of all queries. These values were dramatically higher than those of the initially labeled pages from DMOZ, indicating that AgeRank is useful in finding many pages beyond an labeled set.

4 Extracting Children's Queries from Generic Query Logs

Query logs are valuable resources to explore the search behavior of users and have been found highly useful to improve their search experience. In this study, we employed the AOL query log and the ODP [3] *kids*&*teens* directory to identify differences in the queries and sessions employed by users to retrieve children and general-purpose information. This comparison also allows us to confirm on a large-scale environment previous findings in children physical search on Web search engines [6]. This work represents an important first step toward the automated extraction of children's queries from arbitrary query logs for which no such explicit indication exists.

For the analysis we rely on the Kids and Teens section of the DMOZ directory[3] to automatically identify the queries employed to retrieve content for children. The aim of this DMOZ section is to provide child friendly and safe content to cover the specific needs of people up to 12 (kids), 15 (teens) and 18 years old (mature teens).

We consider that using this directory to identify children-content queries is reasonable and realistic enough given that the content of the directory is frequently regulated and maintained by senior editorial staff, which guarantees that websites with harmful or unsuitable content for children are excluded.

Note that although it is not possible to establish if these queries were performed by children, we are still able to study the characteristics of the queries and sessions when the underlying information need is related to content for children.

In the following paragraphs we describe the data employed in this study, the methodology followed for its analysis and the main lessons[4].

4.1 Overview

The AOL query log employed contains approximately 36 million entries and it was collected during two months in 2006. The identification of the queries employed to retrieve content for children was performed automatically by matching the entries listed in the DMOZ kids & teens directory, which contained 45,635 entries, with the domains clicked in the entries of the query log. Given that the query log does not include the entire URL visited, matches were restricted to the cases in which only the domain is listed as DMOZ entry. Three data sets of query log entries were constructed by employing the matching procedure described above on the DMOZ entries tagged for kids, teens and mature teens. Sessions are constructed by grouping contiguous queries submitted with a time difference smaller than t_θ and that are from the same user. A formal definition of session is shown in Equation 1.

$$S = \langle \langle q_{i_1}, u_{i_1}, t_{i_1} \rangle, ..., \langle q_{i_k}, u_{i_k}, t_{i_k} \rangle \rangle \tag{1}$$

where $u_{i_1} = ... = u_{i_k}$, $t_{i_1} \leq ... \leq t_{i_k}$ and $t_{i_{j+1}} - t_{i_j} \leq t_\theta$ for all $j = 1, 2..., k-1$ The parameter t_θ was set to 30 minutes because it is the most common value employed in the literature [8, 24, 27]. We consider that this time window is also suitable for sessions expressing children information needs because it has been shown that the average time children spend to fulfill an information need varies between 10 and 16 minutes [7]. The sessions used to satisfy children's information needs are those that visit at least one DMOZ domain. Table 5 summarizes the characteristics of the set of sessions collected and Table 4 shows the 10 most frequent queries in the children data[5].

[3] http://www.dmoz.org/Kids_and_Teens/

[4] For a comprehensive detail of our evaluation, please see the work detailed in [18].

[5] Variations of the same domain were removed from this list (e.g nickjr and nickjr.dom).

Table 4: Most frequent queries

Kids	Teens	M_Teens
nickjr.com	the n.com	prom hairstyles
elmo	nasa	american idol
nick jr	hairstyles	cheats
coloring pages	kingdom hearts 2	cheat codes
postopia	claires	prom dresses
candystand	celebrity hairstyles	pussycat dolls
the wiggles	christina aguilera	ea sports
starfall.com	degrassi	bladder infection
dora the explorer	gurl.com	scholarships
primary games	homestarrunner	game cheats

Table 5: Size of the query sets

	Queries	Uniq. Q.	Sessions	Goals
Kids	485,861	10,252	21,009	32,292
Teens	411,474	4,169	7,930	14,503
M.Teens	516,570	10,057	15,519	26,600
All log	36,389,577	10,154,747	10,769,830	8,005,597

4.2 Evaluation and Results

In the following paragraphs we summarize our main results.

4.2.1 Query Length Analysis

Query length is an indicator of the complexity of the query and the difficulty of the user to express information needs using keywords. The average query length found for the kids, teens and mature teens datasets were 3.8, 3.4 and 3.2, respectively. These values differ from the mean of the entire query log, which is 2.5 words per query, with statistical significance using the Wilcoxon signed-rank test at the 95% confidence level. The average query length of the entire data is also in line with the average length reported for other large scale query logs [5, 40]. Interestingly, these results confirmed previous studies in the field of Human-Computer Interaction (HCI). Druin et al. [17] stated that kids of age 8 to 12 tend to formulate longer queries. We also found that 65% of the children retrieval queries contain either of these phrases compared to 56% of the queries in the whole query log. These results also suggest greater difficulty of the former users to formulate queries using keywords.

4.2.2 Query Intent Analysis

Queries were also analyzed using Broder [9] classification which captures three types of user intent: informational, navigational and transactional queries. The results were obtained by manually classifying the queries using the guidelines given by Jansen et al. [26] to classify query intent (Broder categories are also used in this study). The queries were classified by sampling randomly at 15% the unique queries of the kids, teens and mature teens data set. For the whole data set a random sample of 1400 queries was obtained. This size is comparable to the size of the sample employed in a previous study on a large query log [9].

We found that informational queries are preferred in the whole, teens and mature teens data set over transactional and navigational queries. Interestingly, this trend was not observed for the kids queries in which transactional queries are preferred (increase of 20% in respect to the average user of the query log). We found that transactional queries are mainly used in the kids and teens queries to interact with web applications (e.g. flash/java games, academic quizzes) or to obtain free on-line resources (e.g. poems, songs lyrics, coloring pages).

The lower use of informational search in the kids queries compared to the other query sets can be caused by the current lack of specialized IR applications to satisfy children's information needs or to the unsuitability of most of the content in the Web for these users. Given that these type of users are more familiar with the interaction of multimedia and on-line applications, the design of more interactive tools can highly improved the motivation and success of users searching for children-friendly

4.2.3 Click Analysis

The click information of the datasets was analyzed to compare the retrieval performance between children and general purpose content queries. For this analysis we collected the rank distribution and the click frequency of the queries in the datasets. Queries in which highly ranked domains are clicked often indicate that information needs could be more efficiently satisfied by the IR system. On average the retrieval performance of the queries to retrieve children information is poorer than the queries used to retrieve non-children oriented content (5.77 vs 3.58 average rank). This behavior can be explained by the fact that children tend to explore more results given their lack of focus during the search and their difficulty to specify the information needed, as is mentioned in [7]. Statistical significant differences in the rank between the data sets were found using the Wilcoxon test at 95% confidence level.

4.2.4 Sessions Length Analysis

Sessions used to retrieve information for children are longer than general-purpose sessions. The longer average length found for the children sessions (8.76 vs 3.3 query entries per session) suggests that these users were not certain of the relevance of the information found since they had to perform more queries and explore more documents. This result can also indicate that the documents retrieved by the search engine are not sufficient to satisfy the user's information need. These results were proven statistical significant using the Wilcoxon test at 95% confidence level.

4.2.5 Session Duration Analysis

Users require more time to explore and complete information needs associated to children content (20.38 minutes for the kids dataset vs 19.69 minutes), which suggests more difficulty to solve the information tasks associated to these sessions. This results is consistent with the greater amount of queries and clicks on lower ranked pages found in the children queries. Statistically significant differences were found in the average session duration between the kids/teenagers/mature teenagers and general-purpose sessions. No significant differences were found among the types of children sessions. The long duration found in the children's sessions can also suggests that in general users are less successful finding information for children.

4.2.6 Query Reformulations

Users constantly modify their queries in an attempt to get better results from the search engine. The analysis of these query refinements allow us to have a better understanding of the way user's interact with the search engine and the search strategies employed to satisfy their information needs. We consider the following query reformulations: word added to the query, word removed to the query, change of words in the query, spelling correction, new query and more results from the same query. As a matter of example, the following list are provided the definitions for some of these query reformulations.

- Words added to the query (w.a): The previous query is a strict suffix of the target query. e.g. $\{\text{dora}\}_{i-1}\underset{w.a}{\rightarrow}\{\text{dora the explorer}\}_{i}$
- Spelling correction (s.c): The Levenshtein distance between the target and previous query is ≤ 2. e.g. $\{\text{candysand}\}_{i-1}\underset{s.c}{\rightarrow}\{\text{candystand}\}_{i}$
- New query (n.q): Target query does not share any words with the previous query and the Levenshtein distance is greater than 2. e.g. $\{\text{sesame street}\}_{i-1}\underset{n.q}{\rightarrow}\{\text{elmo}\}_{i}$
- More results from the same query (m.r): Target query is identical to previous query and it is used to access a different website.

Although *m.r* is not a formal query reformulation (since no change is performed on the previous query), we included it in this analysis because this action is commonly use in the search process. We found that the *w.a.* and *w.r.* were the most frequent query reformulations in all data sets. In the whole data set *w.a.* correspond to 48.1% of the query reformulations while in the kids dataset 25.9%. The *m.r.* reformulation account to 38.8% of the query reformulations and 59.6% in the kids dataset. A salient difference is the average drop of 22.5% of new queries issued in the children sessions compare to the general-purpose sessions. Most of this drop is reflected in the greater use of the same queries to explore further results, which accounts on average for 90% of the new query reformulation type drop. This result suggest that the users of the children dataset are less skillful refining the query. These users would greatly benefit from query assistance functionalities.

4.2.7 Summary of Lessons Learned

The better understanding of users retrieving information for children on a large-scale achieved in this work allow us to discuss several ways to improve the search experience of these users. We discuss improvements on two IR dimensions: query assistance and aggregated search.

4.2.8 The Need for Query Assistance

The use of longer queries to retrieve children-friendly content can be one of the causes of the lower retrieval performance found for theses queries since it has been shown that longer queries have poorer performance in current search engines [29]. For this reason we consider that refining long queries is highly beneficial for these users as the studies presented in [4, 5]. Cue words are also a valuable resource to rewrite queries by using the cues terms associated to the contexts words or to the entities that occur in the query. This method can ease the exploration of information by providing different content types and dimensions of the topic being searched. The study of further query rewriting techniques as the one presented in [43] can also be beneficial to reduce the cognitive load of reformulating queries which are rarely used in the kids, teens and mature teens data sets.

4.2.9 Difficulty of Searching for Kids

The low percentage of informational queries found in the *children queries* suggests that although these users are familiar with interactive applications, they are not fully harvesting the information content available in the Web. This behavior may be due to the lack of expertise formulating information needs using keywords, to the difficulty of identifying relevant information from the web results or to the lack of more friendly methods to guide the search. This difficulty was also observed in the greater number of entries and longer duration of sessions. This finding may also be due to the preference of these users to use the internet for entertainment purposes as it is observed in the clusters presented in Table 3 and as it is reported in [34]. This user search behavior suggests that more efficient ways to gather the information and more efficient ways to present it to the user are required for these type of users. We consider that aggregated search represents a promising paradigm to address these difficulties. Aggregated search refers to the selection of results from diverse sources and content types and the integration of this information to aid the user to reach his/her information need more efficiently [32].

4.3 Implications for Automated Query Extraction

We believe this work can be extended to enable automated query extraction from arbitrary query logs. One simple approach that can be applied immediately would be to use queries whose landing pages are known children's sites, as described previously. We would then compare various features of sessions within which these queries appear to those known to be indicative of children's search behavior (e.g, longer duration of sessions, greater amount of clicks, less used of reformulations etc), keeping only the sessions whose user's behavior statistically aligns with children's search behavior.

5 Text Simplification and Re-writing for Children

The Internet contains a wealth of information, but only a small fraction of that information is suited for the reading level of children. Especially in the last decade, a lot of research has been put into automatically assigning a measure of readability to text, and retrieving documents that are suited for a predetermined reading level. This section of the report addresses a related issue, that arises when a document with the right reading level can't be found: rewrite the text so that it does become suited, according to an external readability measure. We introduce a method that takes complicated text as input, and generates a text that is simpler and easier to understand for children.

Text simplification may serve many purposes, and has been researched with very different objectives in mind. Originally, the purpose was to break down long sentences in order to improve the accuracy of parsers [12, 39]. Text simplification was also used to automatically make text more understandable by aphasic readers [11], or readers with low literacy skills [10]. Yet another application is the simplification of text as a preprocessing step for other NLP tasks, such as Relation Extraction [28], Semantic Role Labeling [41] and Machine Translation [35].

The goal of most research on text simplification is to make the text as simple as possible. Only [36] and [10] first train a classifier that decides whether or not a sentence is too difficult, and if it is the case then a rule based system is applied to attempt to simplify the sentence. The problem with training a classifier is that annotated training data is needed, and even then the decisions are made on the level of individual sentences, not on the level of the entire document. The problem with simplifying as much as possible is that the text might become too easy: we want the text to fit the reading level of a child as good as possible, rather than making it overly simple.

By casting the problem as an Integer Linear Programming (ILP) problem, it is possible to find a global solution (i.e. choice of simplifications) so that the entire text satisfies certain conditions regarding the reading difficulty. These conditions can be modeled through the objective function and constraints.

5.1 Overview

Our method consists of three components. The first two are the lexical and syntactic simplification of text. The third component concerns choosing the right set of simplifications that were generated by the previous components.

5.1.1 Lexical Simplification

In the lexical simplification step the aim is to replace difficult words and expressions with simpler ones. This task is closely related to paraphrasing and machine translation, with as source language English, and as target language 'simple' English. Unfortunately, whereas there are parallel corpora available for paraphrasing and machine translation, a similar parallel corpus to learn simplifying expressions from is not available. For this reason we focus our attention on an easier task, the lexical substitution of individual words.

Using the most frequent synonyms does not always generate the correct substitutions. Our approach uses a limited form of Word Sense Disambiguation to alleviate this problem. The main idea is that we not only generate alternative words from WordNet, but combine this with a language model [15]. The Latent Words Language model models both language in terms of consecutive words and the contextual meaning of the words as latent variables in a Bayesian network. In a training phase the model learns for every word a probabilistic set of synonyms and related words (i.e. the latent words) from a large, unlabeled training corpus. So rather than taking simply the synonyms from WordNet, we take the intersection with the words generated by the language model (see figure 2 for a graphical representation). Because of the one sense per context phenomenon [42], this gives reasonable grounds to assume the substitutions are correct.

Alternatively, another approach could be to use a standard trigram language model, and ignore the synonyms that have a language model probability below a certain threshold.

What remains is the problem of ranking the different candidates in the intersection of WordNet and the language model, in order to select the easiest. An indication of how easy a word is, could be obtained by looking at the *Age of Acquisition* rating, available from the Oxford psycholinguistic database [38]. Unfortunately, many words lack this rating, so like in previous work we use the Kucera-Francis frequency. The word with the highest frequency is chosen to replace the original word, if it has a higher frequency than the original word.

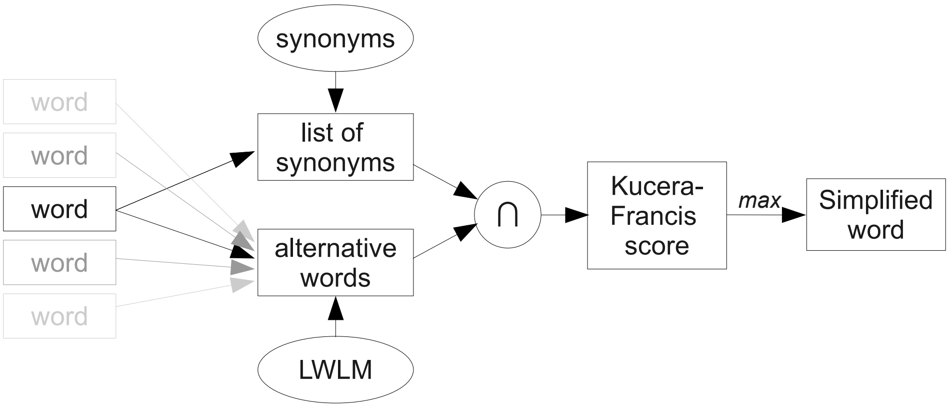

Figure 2: Schematic representation of the lexical simplification

5.1.2 Syntactic Simplification

Previous work has relied on rule based systems to simplify a certain number of syntactic constructions. This is also the approach we follow in this work. Constructions that are typically simplified are relative clauses, appositions, passive voice, and conjunctions [11], but also constructions such as subordinate clauses and if-then structures [39], which are inspired by Rhetorical Structure Theory (RST).

We use the Stanford parser [14] to perform a syntactic analysis of the input sentences. It has a rich annotation scheme that marks several structures that we aim to simplify. We selected the following set of operations to simplify the sentences:[6]

- Appositions: when an apposition is encountered, it is converted into a new sentence, by introducing an auxiliary verb. The clause it is attached to is copied and made the subject of the new sentence.
  Example: John Smith, a New York taxi driver, won the lottery.
  Becomes: John Smith is a New York taxi driver. John Smith won the lottery.
- Relative clauses: the wh-word is replaced with the word it refers to, and the clause is turned into a new sentence.

[6] Those that we did not choose to simplify did not occur in the data (if-then constructions), or did not have a significant effect on the readability measure used in the evaluation (activation of passive voice.)

Example: The mayor, who recently got a divorce, is getting married again.
Becomes: The mayor recently got a divorce. The mayor is getting married again.

- Prefix subordination: this simplification also involves the introduction of new words, slightly based on RST.
  Example: Although it is raining, the sun is shining. Becomes: It is raining. But the sun is shining.
- Infix coordination and subordination: trivially, two parts of a sentence connected by 'and' are split into two sentences. If the subject of the first sentence is also the subject of the second, the Stanford parser detects this, and the subject is duplicated. Next to *and*, two sentences conjoined by words such as *although*, *but*, *because*, ... are also split.

If a sentence can be simplified, and is split into two sentences, then we try to apply the rules again to both of the new sentences. We maintain a list of all possible combinations of rules that can be applied. Thus in this phase, we simply generate all possible simplifications of every input sentence. The actual decision of which rules to apply to which sentences is made by the method described in the next section.

5.1.3 Optimizing the Choice of Simplifications

Before starting the section on the Integer Linear Programming formulation, we will first motivate our choice of variable to optimize in order to make the text fit for a reader of a certain age. Afterwards we will extend this to a more general scenario, to incorporate more features.

Numerous features have been used in assessing the difficulty of text. One that recurs many times is the average sentence length. This feature has often been used in the traditional readability measures; the easiness with which it could be calculated probably played an important role. Still, in current research average sentence length is an important feature when training a classifier for readability assessment. It must be noted that in [37], the average sentence length feature is not significantly correlated with the readability, whereas in [19] this feature was found to be significantly different between original and simplified texts.

Integer Linear Programming A Linear Programming problem consists of decision variables and an objective function, that is a linear combination of the decision variables. Solving the problem means finding an assignment for these variables, so that the objective function is maximized (or minimized). The decision variables can be bounded by linear constraints. In the case of Integer Linear Programming, the decision variables are also constrained to take only integer values. ILP has often been used to find a global solution, for example for dependency parsing [31] and multi-document summarization [20]. One of the first applications of ILP in Natural Language Processing was in the work of [16], whose goal is somehow similar to ours. His goal was to apply paraphrases to sentences in a text, so that the text as a whole conforms to a set of guidelines (e.g., a conference paper that can be no longer than 8 pages). The paraphrases are defined over a Synchronous Tree Adjoining Grammar (STAG). Each paraphrase has a *cost* to apply, and the goal is to make the text conform to the guidelines with a minimal cost. In contrast, in our research the objective function serves to make the text fit a certain age as good as possible. We also take it a step further, by seeing how this is related to research in readability assessment.

Finding a global solution At the end of the previous step (see section 5.1.2), we have for every sentence a list of alternative formulations, that can replace the original sentence. For each of these alternatives, we can calculate the influence this will have on the text as a whole. Focussing on the average sentence length, the relevant features that will be influenced by each alternative are the number of sentences and the number of words.

Suppose the original text has S sentences and W words, and sentence $i, \forall i = 1 \ldots S$ has n_i possible alternatives, indicated by $a_{i1} \ldots a_{in_i}$, and a_{i0} the original sentence[7]. The a_{ij} variables can only be zero or one (a value of one meaning the corresponding alternative should be used), and for a fixed i exactly one of the a_{ij} variables must be one (there can only be one alternative chosen). We can calculate for each a_{ij} the influence this will have on the average sentence length, by calculating the difference in number of sentences, Δs_{ij}, and the difference in number of words, Δw_{ij}, compared to the original sentence. To illustrate with the example of the first rule in section 5.1.2: the application of this rule ($a_{10} = 0, a_{11} = 1$) would result in an increase of 1 in the number of sentences ($\Delta s_{11} = +1$), and an increase in number of words by 3 ($\Delta s_{11} = +3$).

Stating that the average sentence length should be at most m words per sentence can then be written with the formula:

$$\frac{W + \sum_{ij} a_{ij}\Delta w_{ij}}{S + \sum_{ij} a_{ij}\Delta s_{ij}} \leq m \tag{2}$$

By rearranging, this equation can be rewritten to the following form:

$$\sum_{ij}(\Delta w_{ij} - m\Delta s_{ij})a_{ij} \leq Sm - W \tag{3}$$

With the following constraints:

$$a_{ij} \in \{0, 1\} \tag{4}$$

$$\sum_{j=0}^{n_i} a_{ij} = 1, \forall i \tag{5}$$

The left hand side of equation 3 can be minimized by using it as the objective function in the ILP formulation, with the constraints from equations 4 and 5. Defining a lower bound on the average sentence length can be done trivially by using equation 3 with a $\geq$ sign instead of the $\leq$ sign, in the form of another constraint. This way the average sentence length isn't made too small, and the text overly simple.

Extension to more general features A limitation of this method is that it is not possible to minimize a linear combination of averages, what would be needed for optimization towards e.g. the Flesh-Kincaid score. Because of the two averages in this formula (average sentence length and average syllables per word), the optimization problem becomes a Quadratic Programming problem, which is harder to solve.[8]

It is possible to optimize towards features that are not averages. For example, suppose that we can measure the difficulty of a text by a linear combination of the total number of sentences and the total number of words:

$$\text{difficulty} = \alpha W + \beta S$$

We can then use a similar ILP formulation as in equation 2, so that the difficulty can be minimized by choosing optimal assignments for the variables a_{ij}:

$$\alpha(W + \sum_{ij} \Delta w_{ij} a_{ij}) + \beta(S + \sum_{ij} \Delta s_{ij} a_{ij}) \leq \text{difficulty}$$

Which can be rewritten to:

$$\sum_{ij}(\alpha\Delta w_{ij} + \beta\Delta s_{ij})a_{ij} \leq \text{difficulty} - \alpha W - \beta S$$

[7] Note that a_{ij} can consist of more than one sentence for $j > 0$.

[8] See [16] for details.

with α and β the model parameters, originating from, for example, a linear regression model. Linear regression has been used often in predicting the reading difficulty (e.g. [19, 25]). As long as the features are defined as a total, rather than an average, it is possible to write this in the ILP formulation, and optimize for a certain difficulty. Also the statistical language modeling approach from [13] can be formulated in this way.

In the case that averages are still needed, an alternative solution would be to define upper and lower bounds on each of these features separately, e.g. by taking the average $\mu \pm$ the standard deviation σ, estimated from training data. If the resulting ILP is infeasible, i.e. it is impossible to solve, then the constraints can iteratively be relaxed to fall between $\mu \pm \gamma\sigma$, with $\gamma \geq 1$, until the ILP problem becomes feasible.

5.2 Evaluation and Results

We evaluated these methods on two types of text: wikipedia articles, and news articles. Using the language model in the lexical simplification showed a 12% increase in accuracy over the baseline method, unfortunately still limited to 61%. The syntactic simplification worked significantely better for some operations then others. On average, the accuracy on the wikipedia text is 63.8%, and 58.5% on the newswire text.

When trying to simplify the text for a given age (as measured with the Flesch-Kincaid grade level), it appears that these simplification operations by themselves are not sufficient. For Wikipedia, that grade level went roughly from 16 to 14, when trying to simplify as much as possible. For the news articles, the grade level went from 10.8 to 9.3 (grade 9 corresponds to age 14-15). There are two possible solutions. First, the Flesch-Kincaid grade level might be inaccurate, and better ways to determine the grade level are possible. This is further indicated by the average grade level of the used Wikipedia articles (i.c. 16). The other solution is to use more simplification operations. It appears that only one sentence in five can be split into multiple sentences using the current method. One promising operation is sentence compression (the summarization of single sentences), which will be included in further research. This could also solve a problem with the lexical simplification: often a difficult word was not replaced by a simpler one, because there was no simpler way of describing it with one word. “Summarizing” the word out of the sentence could offer a solution.

6 Conclusions

In this report we have covered a broad range of research toward the end of information extraction. All of this research offers key contributions to the PuppyIR project, much of which we intend to directly integrate into an existing demonstration search engine (see D3.2) whose evolution remains ongoing. The D3.2 provides the starting point and vehicle for a complete search engine with which children can directly interact.

First, both classification approaches offer complementary methods to determine the suitability of a web page. Currently we are exploring ways to integrate the ideas and technology from both toward a unified ranking function. This ranking function is intended to be directly integrated into our search engine to allow children to issue arbitrary web queries, and receive answers that are more suited to their age level.

Next, the text-simplification approach offers support for cases where the above ranking is insufficient. We envision that its application would be in taking existing results and simplifying them to the young reader. This could create useful opportunities to make the larger web more accessible to child users.

The query extraction approach discussed in Section 4 has more subtle applicability to our search engine. One possibility is that its ability to identify child users (based on landing pages and other session characteristics) could be used for our search engine to guess the age level of the user and adapt accordingly (e.g., by more aggressively promoting child-friendly pages via the ranking function, or by executing text-simplification toward an even younger target audience).

Regarding requirements for applying these techniques today, we can speak informally about what is needed. Both classification approaches are likely to be useful already. Both are tunable, with the threshold allowing greater confidence in results at the expense of fewer pages, allowing a practitioner to configure a system accordingly. Our anecdotal experience is that neither of these systems are risky in terms of extreme false positives (i.e., promoting pages that are extremely inappropriate for kids). We believe that the techniques already perform better than what a child would be exposed to using normal search.

Since the query work in Section 4 has not yet manifested in an applied system, we cannot speak to its implementation requirements.

Finally, the text-simplification work described in Section 5 has more substantial challenges in practical adoption. One limitation is the state of the art in sentence parsing, a technology upon which simplification relies as core functionality. Sentence parsers present a limitation in that they often make mistakes, and simplification cannot work despite parser errors. Nonetheless, the simplification research is expected to explore alternative approaches that may mitigate the importance of sentence parsers and hence enable higher potential success rates in simplification.

Regarding the use of natural language processing in preprocessing (NLP), it has taken a minimal role in this work with the exception of the text simplification research. The feature-based classification work employed extremely simple NLP approaches such as stemming and stop-word removal (two common and well-established techniques). In AgeRank, the use of link information alone obviates any need of NLP as language is not used. The query-extraction work concentrated more heavily on the destination page of queries so processing was unnecessary. Finally, the text-simplification work made heavy use of NLP techniques.

7 References

[1] ODP – Open Directory Project. `http://www.dmoz.org/`.

[2] CrowdFlower. `http://www.crowdflower.com`, 2010.

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

[4] M. Bendersky and W. B. Croft. Discovering key concepts in verbose queries. In *SIGIR '08*, pages 491–498, New York, NY, USA, 2008. ACM.

[5] M. Bendersky and W. B. Croft. Analysis of long queries in a large scale search log. In *WSCD '09: Proceedings of the 2009 workshop on Web Search Click Data*, pages 8–14, New York, NY, USA, 2009. ACM.

[6] D. Bilal. Children's use of the yahooligans! web search engine ii. cognitive and physical behaviors on research tasks. *J. Am. Soc. Inf. Sci. Technol.*, 52(2):118–136, 2001.

[7] D. Bilal. Children's use of the yahooligans! web search engine iii. cognitive and physical behaviors on fully self-generated search tasks. *J. Am. Soc. Inf. Sci. Technol.*, 53(13):1170–1183, 2002.

[8] P. Boldi, F. Bonchi, C. Castillo, D. Donato, A. Gionis, and S. Vigna. The query-flow graph: model and applications. In *CIKM '08*, pages 609–618, New York, NY, USA, 2008. ACM.

[9] A. Broder. A taxonomy of web search. *SIGIR Forum*, 36(2):3–10, 2002.

[10] A. Candido Jr, E. Maziero, C. Gasperin, T. Pardo, L. Specia, and S. Aluisio. Supporting the adaptation of texts for poor literacy readers: a text simplification editor for Brazilian Portuguese. In *Proceedings of the Fourth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 34–42. Association for Computational Linguistics, 2009.

[11] J. Carroll, G. Minnen, Y. Canning, S. Devlin, and J. Tait. Practical simplification of English newspaper text to assist aphasic readers. In *Proceedings of the AAAI-98 Workshop on Integrating Artificial Intelligence and Assistive Technology*, pages 7–10. Citeseer, 1998.

[12] R. Chandrasekar and B. Srinivas. Automatic induction of rules for text simplification. *Knowledge Based Systems*, 10(3):183–190, 1997.

[13] K. Collins-Thompson and J. Callan. Predicting reading difficulty with statistical language models. *Journal of the American Society for Information Science and Technology*, 56(13):1448–1462, 2005.

[14] M.-C. de Marneffe, B. MacCartney, and C. D. Manning. Generating typed dependency parses from phrase structure parses. In *In Proceedings of LREC-06*, pages 449–454, 2006.

[15] K. Deschacht and M.-F. Moens. The Latent Words Language Model. In *Proceedings of the 18th Annual Belgian-Dutch Conference on Machine Learning*, 2009.

[16] M. Dras. *Tree adjoining grammar and the reluctant paraphrasing of text*. PhD thesis, Macquarie University NSW 2109 Australia, 1999.

[17] 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 *IDC '09: Proceedings of the 8th International Conference on Interaction Design and Children*, pages 89–96, New York, NY, USA, 2009. ACM.

[18] S. Duarte Torres, D. Hiemstra, and P. Serdyukov. An analysis of queries intended to search information for children. In *IIiX '10*, pages 235–244, New York, NY, USA, 2010. ACM.

[19] L. Feng, N. Elhadad, and M. Huenerfauth. Cognitively motivated features for readability assessment. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 229–237. Association for Computational Linguistics, 2009.

[20] D. Gillick and B. Favre. A scalable global model for summarization. In *Proceedings of the Workshop on Integer Linear Programming for Natural Langauge Processing*, pages 10–18. Association for Computational Linguistics, 2009.

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

[22] K. Gyllstrom and M.-F. Moens. Wisdom of the ages: toward delivering the children's web with the link-based AgeRank algorithm. In *CIKM '10*, New York, NY, USA, 2010. ACM.

[23] Z. Gyöngyi, H. Garcia-Molina, and J. Pedersen. Combating Web spam with Trustrank. In *VLDB '04*, pages 576–587. VLDB Endowment, 2004.

[24] D. He and A. Goker. Detecting session boundaries from web user logs. In *In Proceedings of the BCS-IRSG 22nd Annual Colloquium on Information Retrieval Research*, pages 57–66, 2000.

[25] M. Heilman, K. Collins-Thompson, and M. Eskenazi. An analysis of statistical models and features for reading difficulty prediction. In *Proceedings of the Third Workshop on Innovative Use of NLP for Building Educational Applications*, pages 71–79. Association for Computational Linguistics, 2008.

[26] B. J. Jansen, D. L. Booth, and A. Spink. Determining the informational, navigational, and transactional intent of web queries. *Inf. Process. Manage.*, 44(3):1251–1266, 2008.

[27] R. Jones and K. L. Klinkner. Beyond the session timeout: automatic hierarchical segmentation of search topics in query logs. In *CIKM '08*, pages 699–708, New York, NY, USA, 2008. ACM.

[28] S. Jonnalagadda, L. Tari, J. Hakenberg, C. Baral, and G. Gonzalez. Towards effective sentence simplification for automatic processing of biomedical text. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, pages 177–180. Association for Computational Linguistics, 2009.

[29] G. Kumaran and V. R. Carvalho. Reducing long queries using query quality predictors. In *SIGIR '09*, pages 564–571, New York, NY, USA, 2009. ACM.

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

[31] A. Martins, N. Smith, and E. Xing. Concise integer linear programming formulations for dependency parsing. In *Proceedings of Annual Meeting of the Association for Computational Linguistics (ACLIJCNLP 09), Singapore*, 2009.

[32] V. Murdock and M. Lalmas. Workshop on aggregated search. *SIGIR Forum*, 42(2):80–83, 2008.

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

[34] Ofcom. Uk children's media literacy: Research document, March 2010.

[35] F. Oliveira, F. Wong, and I. Hong. Systematic processing of long sentences in rule based Portuguese-Chinese machine translation. *Computational Linguistics and Intelligent Text Processing*, pages 417–426, 2010.

[36] S. Petersen and M. Ostendorf. A machine learning approach to reading level assessment. *Computer Speech & Language*, 23(1):89–106, 2009.

[37] E. Pitler and A. Nenkova. Revisiting readability: A unified framework for predicting text quality. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 186–195. Association for Computational Linguistics, 2008.

[38] P. Quinlan. *The Oxford psycholinguistic database*. Oxford University Press Oxford, 1992.

[39] A. Siddharthan. Syntactic simplification and text cohesion. *Research on Language & Computation*, 4(1):77–109, 2006.

[40] C. Silverstein, H. Marais, M. Henzinger, and M. Moricz. Analysis of a very large web search engine query log. *SIGIR Forum*, 33(1):6–12, 1999.

[41] D. Vickrey and D. Koller. Sentence simplification for semantic role labeling. In *Proceedings of ACL-08: HLT*, pages 344–352, Columbus, Ohio, June 2008. Association for Computational Linguistics.

[42] D. Yarowsky. One sense per collocation. In *Proceedings of the Workshop on Human Language Technology*, page 271. Association for Computational Linguistics, 1993.

[43] W. V. Zhang, X. He, B. Rey, and R. Jones. Query rewriting using active learning for sponsored search. In *SIGIR '07*, pages 853–854, New York, NY, USA, 2007. ACM.

[44] X. Zhu and Z. Ghahramani. Learning from labeled and unlabeled data with label propagation. Technical Report CMU-CALD-02-107, Carnegie Mellon University, 2002.

A Additional Material

Table 6: Complete list of features

term frequency ”child”	FORECAST score	avg unique entities per sent.	avg Wiktionary definition length	neighboring kid’s pages
term frequency ”kid”	Kincaid score	avg entities per sent.	avg Wiktionary parts-of-speech	neighboring general pages
text length (words)	LIX score	avg proper nouns per sent.	avg Wiktionary senses	URL term freq. ”child”
text length (sentences)	SMOG score	avg nouns per sent.	avg Wiktionary translations	URL term freq. ”kid”
avg sentence length (words)	OOV simple Wiktionary (12951 terms)	avg verbs per sent.	to-reference share	URL length
avg word length (characters)	OOV simple Wikipedia (48234 terms)	avg adjectives per sent.	about-reference share	domain length
avg syllables per word	OOV Basic English (850 terms)	avg prepositions per sent.	n HTML tags	domain term LM score
complex word share	OOV General Service (2284 terms)	kid’s character LM	n non-tag tokens	min image size
punctuation character share	OOV most frequent (1000 terms)	simple Wiktionary character LM	tag/token ratio	max image size
capital letter share	OOV Dale-Chall (3000 terms)	simple Wikipedia charcter LM	n images	avg image size
stop word share	OOV Ext. Basic English (2050 terms)	kid’s unigram LM	image/word ratio	image size σ
total words/unique words	OOV academic (570 terms)	simple Wiktionary unigram LM	n scripts	n jpgs
ARI score	person entity share	simple Wikipedia unigram LM	script/word ratio	n gifs
Coleman-Liau score	location entity share	kid’s 3-gram LM	n links	n pngs
Flesch score	organization entity share	simple Wiktionary 3-gram LM	link/word ratio	n bmps
Gunning-Fog score	total entities/unique entities	simple Wikipedia 3-gram LM	internal/external link ratio	commercial intent score