Wisdom of the Ages: Toward Delivering the Children's Web with the Link-based AgeRank Algorithm

Karl Gyllstrom
karl.gyllstrom@cs.kuleuven.be

Marie-Francine Moens
sien.moens@cs.kuleuven.be

Department of Computer Science
Katholieke Universiteit Leuven
Leuven, Belgium

ABSTRACT

Though children frequently use web search engines to learn, interact, and be entertained, modern web search engines are poorly suited to children's needs, requiring relatively complex querying and filtering of results in order to find pages oriented to young audiences. To address this limitation, we designed AgeRank, a link-based algorithm that ranks web pages according their appropriateness for young audiences. We show its effectiveness through a multipart evaluation that demonstrates AgeRank to be accurate in page-labeling, widely-spanning in page coverage, and with high potential to improve children's search. As a fast, scalable, and effective algorithm, AgeRank can be adopted by search engines seeking to more effectively address the needs of young users, or easily fitted to complementary machine-learning based classification approaches.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

General Terms

Algorithms, Experimentation, Measurement

Keywords

Children, Link-analysis, Mechanical Turk

1. INTRODUCTION

Most web search engines are not highly usable by children because they present results for general audiences rather than results that are most suitable for young audiences. Currently, the only way to find pages for children through general web search engines is via targeted queries (e.g., appending "for kids" to a query) and filtering results. Unfortunately, children generally perform worse than adults at these tasks (9; 12), and may often interact with the web without parental assistance (6). Children are at a disadvantage in accessing the web, and there is a need for search engines that better serve young audiences by delivering results that are targeted to their age level.

To address this need, we created AgeRank, a link-based approach to label pages by the likelihood that those pages are appropriate for children. This labeling can be integrated into traditional search functionality, such as PageRank (10) and text-based search, in a number of ways, including as a complementary component to the ranking function, or as a way of filtering results altogether. Implemented in a search engine, children could use simple queries to access the general web much like adults do.

Philosophically, our focus is not to *limit* the information available to children by filtering explicit material, but rather to *expand* the available information by promoting results that are more suitable to them. Suitability is defined by features such as reading level, subject level, interactivity, and appeal. 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. AgeRank provides a way to rank results such that a child using a search engine would be more likely to receive the simpler page than the advanced one.

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.

We show through an multipart evaluation a number of important qualities of AgeRank, including its accuracy of labeling, its coverage of web pages and search results, and

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

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

its capacity to improve search for children through child-customized rankings of search results.

2. RELATED WORK

Bilal's research on children's web searching behaviors has identified a number of key areas that need to be addressed by the web search community, spanning interfaces and content (8; 9). Children have more difficulty issuing good queries and are unlikely to view more than five results (12); meaning they are not likely to sift through results to find child-oriented results. Not only are their queries more likely to produce weaker results, but their ability to deal with weaker results is poorer than the ability of adults.

There are some publicly available web search engines specialized for children, including *KidsClick!* (3) and *Yahoo! Kids* (7). These engines attempt to produce more child-friendly results, but suffer the limitation that the URLs and sites they index are manually selected by experts, constraining search to a limited corpus that cannot evolve as quickly as the web. The goal of our work is to minimize the role of expert effort in finding suitable pages, which would complement the efforts of these engines nicely.

PageRank is a popular ranking method, being both effective at ranking and having an intuitive appeal (10). Simply put, PageRank identifies a page to be credible if many credible pages point to it. While effective in web ranking, this approach is at odds with the goal of finding children's pages. Pages are more likely to be linked to if they have general appeal, meaning they are less likely to specifically target any particular audience, most notably children. Hence, PageRank will tend to promote pages with less specific demographic appeal. (We explore this in Section 5.1.)

TrustRank is a method to identify spam pages on the web, based on observations about how legitimate and spam pages link to one another (14). Our work shares TrustRank's general goal of distinguishing between two classes of pages, although the spam-web and children's web contain quite different linking patterns and necessitate distinct approaches. For example, TrustRank severely punishes pages whose path of some number of hops contains a spam page. We observe that children's pages quite often link to non-child pages, and try to account for that by considering the proportion of links.

Topical PageRank modifies PageRank to favor pages of a given topic, as represented by a manually selected set of pages (15). It operates by replacing the dampening function of PageRank to simulate a higher probability of PageRank's "random surfer" switching to the biased set. While this concentrates more of the PageRank scores in pages connected to the bias set, it has a limitation in our context, as, like PageRank, it will still promote more highly linked pages. For example, pages within English Wikipedia will likely maintain high PageRank across different English topics, even though it is inferior to the simple Wikipedia for young audiences. By considering each page in terms of proportion of children's to adults' links, rather than accumulated PageRank, we can prevent this, as highly linked pages that are also more heavily linked with adults' pages will be penalized.

Conceptually, our work incorporates themes from both topical PageRank and TrustRank. Like topical PageRank, we try to promote pages common to a positively labeled set, and like TrustRank, we try to demote pages from this set that are related to a negatively labeled set.

In a similar vein, much work is done on identifying communities, or clusters of pages with relatively higher intra-linking (e.g., (13)). Though our goals overlap, a community is too strict a definition: children's web pages comprise millions of web pages across many different topics, with relatively sparse intra-linking as a whole. Additionally, a page can be potentially valuable to children without strictly being a member of the community; we desire a relative likelihood rather than strict classification.

AgeRank has complementary information to feature-based classification approaches such as Support Vector Machines. Our goal with AgeRank was to study links in isolation, as links are valuable sources of information about the relationship among pages, and are useful to understanding the topology of the children's web. Links can be especially useful in cases where page contents are difficult or impossible to classify; for example, if they contain relatively little text and a lot of multimedia data that is more difficult to classify. Furthermore, classification requires a concrete definition of what a child's page is. A page for a child may contain links to pages that are not obviously for children and which may not fit this classification, but may nonetheless contain resources that the page author determined to be contextually relevant and appropriate for children.

3. APPROACH

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.

3.1 Calculating AgeRank

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 (18). 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., (14)). 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 (18), 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:

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

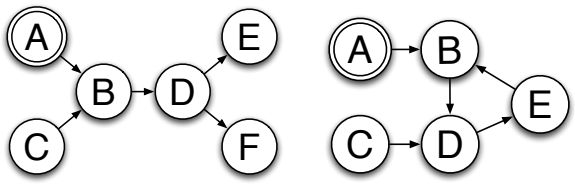
Consider the web graph example (a) depicted in Figure 1, where page A is from L_+. At the initial state, A has values $(1, 1, 0, 0)$. After iteration 1, B's score is $(0, \frac{1}{2}, 0, 0)$, since the similarity score is $\frac{1}{2}$ since B has two incoming links. At iteration 2, D's score is $(0, \frac{1}{2}, 0, 0)$; 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. 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.



(a) Simple web graph (b) Web graph with cycle

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

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 Discussion

In this section we provide some justification for the design decisions behind AgeRank.

PageRank is a useful algorithm for determining the credibility of pages, but there are reasons that we found it to not be ideally suited for our task. PageRank has the effect of accumulating weight from the graph. Conceptually this makes sense in terms of credibility: pages with many incoming links have relatively higher PageRanks, and the pages to which they link inherit even more. In our approach, we are looking for the probability that a page is suitable for young audiences, which is not relative. In other words, if we have two pages that are both deemed very appropriate for children, we have no need to further order them according to this appropriateness.

Another problem with PageRank is that it converges to

a state that is relatively independent of its original state. After a large number of iterations, the PageRank scores will concentrate on the pages at which the random surfer stops following links. As a consequence, weight shifts away from pages to which few other pages link. With our approach, the scores (positive and negative) for each individual page continue to accumulate with each iteration.

Our approach considers both the outward and inward links for a page. As algorithms like PageRank show, useful inferences about a page can be drawn by a page's incoming links (e.g., page A is most likely to link to page B if it considers B to be credible or relevant to A's topic). We believe that outgoing links are important in the context of age-appropriateness because the author of a child's web page is likely to take more care to ensure that the outgoing links are equally appropriate for children.

Our use of clamping stems from (18) but we justify the intuition by saying that the pages in L_+ and L_- are assumed to be valid labels and these labels should not be affected by iterations of the algorithm because, as a label-propagator, evidence from AgeRank is inherently less reliable than from the originally labeled pages. Further, clamping ensures that none of P_{out}, P_{in}, N_{out}, or N_{in} can decline, avoiding the aforementioned problem of convergence to an input-independent state.

An important advantage of the link-based approach in AgeRank is that it is highly parallelizable and amenable to MapReduce (11) in a manner akin to PageRank. All that is needed is the transition matrix of the web graph, (including its transpose, if scoring from outward links is desired). In this way, it is useful to search engines because no further indexing or preprocessing is required, as may be the case with classification approaches.

4. EVALUATION: SCORE ASSESSMENTS

In the following sections we describe our evaluation, which is divided into three areas. First, we directly evaluate the accuracy of AgeRank by comparing the scores it produces to actual user ratings of pages' child-appropriateness. Next, we evaluate the effect of AgeRank in a more applied, traditional web search context, separating discussion into search results (Section 5) and search queries (Section 6). We begin with a description of the experimental set up.

4.1 Experimental set up

The application of AgeRank requires an initial set of labeled pages. To this end, we constructed a ground truth set of positive pages (L_+) and negative pages (L_-). To construct these sets, we used two sources. The first source is the *Open Directory Project* (4) (DMOZ), which is a large, hierarchical collection of user-generated categorizations of web pages. Within DMOZ there exists the category *Kids and Teens*, under which a large number of subcategories and web page URLs have been placed by the community of users. These URLs pertain to pages from the general web, not those affiliated with DMOZ itself. The second source is the *simple-language* collection of Wikipedia[3], a variant of Wikipedia in which articles are written in simplified English with the explicit intent to be more approachable to children and non-native speakers. To construct L_- we collected all of the URLs within the following DMOZ topics: *Business*, *Computers*, *News*, and *Science*. From L_- we removed 1023 URLs that were also contained L_+.

We implemented AgeRank in Python using the Disco MapReduce framework[4]. We executed AgeRank on the first 108,160,000 pages in the *ClueWeb09* dataset (2), a massive crawl of web pages from 2009. We removed pages from L_+ and L_- that are not contained within the *ClueWeb09* dataset, as well as pages whose IDs were larger than 108,160,000. This produced 37,662 unique URLs for L_+ and 310,607 unique URLs for L_-. We ran AgeRank for 7 iterations, producing ratings for 11,594,120 pages[5]. From this set of rated pages, we removed those appearing in L_+ or L_- so that the set contained rankings that were assigned purely from the algorithm. We refer to this filtered collection as L_{AR}.

4.2 Page rating accuracy

We sought to measure the degree to which the individual AgeRank scores (P_{out}, P_{in}, N_{out}, N_{in}, and Tot) correlate with actual user ratings. To this end, we conducted an online evaluation with human raters using the Mechanical Turk (1). We first created a pool of URL/AgeRank pairs to evaluate. For our first experiment, we selected the 2 million URL/AgeRank pairs with the highest total evidence (positive or negative) to minimize noise. We then randomly sampled 1663 URLs from this pool for human ratings.

We uploaded the URLs to Mechanical Turk for human ratings[6]. On the rating page, we provided a link to the page to be rated, as well as a selection widget which allowed users to specify on a 7-point Likert scale (from -3 to 3) the degree to which the page is oriented toward or appropriate for children, with 0 being a neutral score. Users could also specify that the URL did not load properly. We asked users to perceive child-appropriateness in terms of reading-level, subject matter, and interest to children.

We measured the relationship between the human ratings and the AgeRank scores through linear regression of the human rating – which we will refer to as H – on AgeRank. Our ratings consisted of a list of 6-tuples ($url \times H \times P_{out} \times P_{in} \times N_{out} \times N_{in} \times Tot$). In each rating, we normalized each of the values P_{out}, P_{in}, N_{out}, and N_{in} by dividing them by the sum of the four values. Conceptually, this characterizes the amount of total evidence concentrated within each individual score. This point is important because the amount of total evidence is highly variable: e.g., a page can have a high Tot score while having low scores for P_{out} and P_{in} if they are larger than N_{out} and N_{in}.

One of our concerns was that AgeRank score would be more heavily concentrated among pages that are more proximal to L_+ or L_-. To isolate the usefulness of AgeRank in cases of greater distances, we also measured the regressions on a sample of pages from L_{AR} whose distances from a page in L_+ were at least 4 hops, calling this score $Tot_{hops \geq 4}$. We randomly sampled 1277 separate ratings for this set.

[3] `http://simple.wikipedia.org`

[4] `http://discoproject.org/`

[5] We did not experimentally determine the ideal number of iterations for which to execute AgeRank. We arrived at 7 based on observations that label propagation would likely attenuate scores substantially after 7 iterations such that further iterations would change the scores by very small amounts.

[6] The URL contents may have changed since the crawl, but *ClueWeb09* contains insufficient information to reconstruct pages (e.g., no images or style sheets).

For comparison, we evaluated the original DMOZ scores. We created a pool of equal numbers of positive and negatively rated pages by sampling from L_+ and L_-. We randomly sampled 500 pages from this pool, then placed the pages on Mechanical Turk for a human rating as described above. We calculated the linear regression of human score on DMOZ category; to assist in this calculation we assigned a DMOZ-score of 1 to a page that is in L_+, and a DMOZ-score of 0 to a page that is in L_-. This matches the range of the scoring system in AgeRank such that the slopes from the linear regressions are comparable. All of these regression scores are reported in Table 1.

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
$Tot_{hops \geq 4}$	0.2129	0.0460
DMOZ	0.5206	0.0001

Table 1: Linear regression of various AgeRank scores on human scores and Dmoz on human scores.

These results indicate that the values P_{out}, P_{in}, N_{out}, and N_{in} reflect the hypothesis that paths to L_+ are predictive of a page being more child-appropriate, and that paths to L_- are predictive of the opposite. Inward and outward links are both predictive along both positive and negative scores, although N_{in} was not statistically significant. Tot, as a combination of values, also exhibits a significant positive relationship with human scores. $Tot_{hops \geq 4}$ reflect Tot to a lesser extent, indicating that AgeRank is still useful after relatively more hops.

Interestingly, the DMOZ scores are not dramatically higher than the AgeRank scores. This has both positive and negative implications. The positive result is that AgeRank is able to find relatively good pages compared to L_+ and L_-. This further indicates that locality among children's pages is reasonably strong since pages within a few hops of L_+ are likely to be children's pages as well. The negative finding is that DMOZ is not an ideal source of labels, as human scores often do not agree with it, although, as our results in subsequent sections show, we can still achieve good retrieval using it for label sources L_+ and L_-.

5. EVALUATION: SEARCH RESULTS

Where the previous section explicitly assessed the accuracy of AgeRank scores, this section examines the effects of AgeRank in an applied information retrieval context. We begin with a description of our experimental setup. We generated a simulated set of children's queries with which to evaluate our approach, as, to our knowledge, there is no existing collection of children's web search queries. We constructed this set by collecting the titles of leaf subdirectories under the top-level topic "Kids and Teens" from DMOZ, which included titles such as "dinosaurs" and "Egypt", producing 4237 queries. We refer this query set as Q_{kids} (See Section 6.2 for investigation of these queries).

For each query $q \in Q_{kids}$, we executed a text-based search upon the *ClueWeb09* corpus using the Indri search system[7], using Dirichlet smoothing with $\mu = 1600$ and the default values for all other parameters. For each query q, the system produced a result list r_q, which included 1000 results listed in order of the system-assigned relevance score. We say that each result in r_q has a relevance-rank, or a numeric position in the ordering corresponding to the degree to which Indri assesses its relevance to q.

5.1 PageRank comparison

PageRank is effective in determining credibility among web pages, but it is not inherently age-sensitive. And, as we argued previously, the fact that PageRank promotes pages of general appeal can cause rankings that are not ideal for children. Here, we measure the correlation between AgeRank and PageRank scores for web pages, which gives us an indication of the usefulness of AgeRank as a complementary ranking. For example, if higher AgeRank pages also have higher PageRank scores, this would reduce the relative potential improvement of incorporating AgeRank into page ranking.

From L_{AR} (described in Section 4.1), we sampled 3 million pages, for each page collecting its page ID, AgeRank score, and PageRank value[8]. We calculated the linear regression to determine the dependency of PageRank on AgeRank. This produced a tiny slope, with an R^2-value of 0.0003 and P-value of $2.095e-57$. From these measurements we conclude that the correlation among AgeRank and PageRank is significantly negligible[9]. PageRank can not be used as a surrogate for AgeRank, so AgeRank can provide complementary information about a page.

Since pages with high PageRank reflect those that are most important on the web, it is important that our method produces labels for a large amount of these pages. Therefore, we examined the portion of pages with high PageRank for which AgeRank produced a score. We selected the 7 million pages with the highest PageRank scores from *ClueWeb09*, and, for each page in this list, checked whether the page was also in L_{AR}. Figure 2 depicts a plot of the percentage of pages for which a label exists across the top PageRanked pages. Of the top 100,000 pages, 91.4% of pages have AgeRank scores; of the top 7 million pages, 19.9% do, indicating a good coverage of the top pages by AgeRank, using L_+ and L_-.

5.2 Search Results Coverage

We further explore the coverage of L_{AR} in terms of search results. Specifically, we measure the degree to which AgeRank can affect search results for queries. Ideally, across children's topics, and even topics of marginal interest to children, a large number of pages have AgeRank values, meaning our approach is finding good pages, and giving us greater power to reorder results in a more suitable order for children.

For each result set r_q, we measured the number of results that are contained in L_+, L_-, and L_{AR}. Table 2 depicts the number of queries for which at least one page from the source appears in the search results for the query. Results from L_{AR}

[7] `http://www.lemurproject.org/indri/`

[8] PageRank values for the pages were distributed with the *ClueWeb09* dataset.

[9] Linear regression on log-transformed PageRanks was equally conclusive

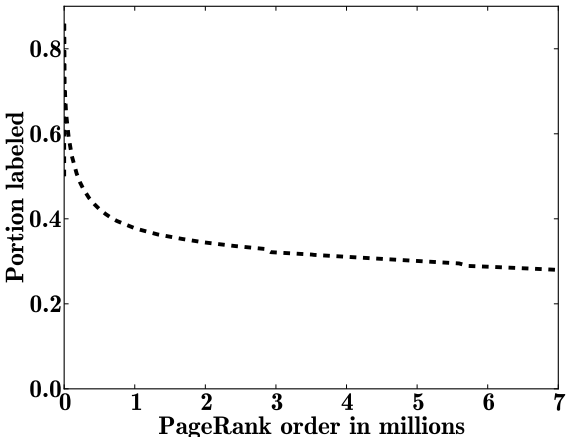


Figure 2: Percentage of AgeRank coverage over top 7 million PageRank pages.

appear in nearly all queries, meaning that in most cases, we could achieve a reordering in terms of child-appropriateness.

Source	# Queries	Portion
L_+	2238	0.53
L_-	1494	0.35
L_{AR}	4195	0.99

Table 2: Number of queries for which at least one result appears from each source.

The above results do not express the positions of results within query results. To that end, for each label source in L_+, L_-, and L_{AR}, we measured the *precision-at-N* (10, 100, 1000) across all r_q, where a result is considered "relevant" *if it is contained in the source*. Additionally, we calculated the Discounted Cumulative Gain (DCG) (16) for each source across all r_q, where, for relevance score, we used a 1 if the page was in the source, and 0 otherwise. DCG is effective here in reflecting both the number of results as well as their ranks. To be clear, inclusion in a source is not proof of topical relevance; we use these measures not to assess relevance, but rather to characterize where pages from these sources appear within result lists as a way of comparing how well the sources cover search results. To emphasize this distinction, we refer to these measures as *L-precision* and *L-DCG*. These results are depicted in Table 3.

Source	L-P@10	L-P@100	L-P@1000	L-DCG
L_+	0.0079	0.0034	0.0014	0.2182
L_-	0.0018	0.0012	0.0010	0.1257
L_{AR}	0.1184	0.1367	0.1432	17.6019

Table 3: Mean L-P@N and L-DCG scores across Q_{kids}.

These results show that the use of AgeRank dramatically expands upon the information available in L_+ and L_- (i.e., DMOZ). One reason that we show this comparison is that L_+ represents a baseline of what is currently available to children who use search. In other words, if a child's web search were restricted to pages within L_+ (i.e., pages for which an explicit child label exists, e.g., KidsClick!), the availability of pages is much lower, as reflected by the poor performance of L_+ in terms of precision and DCG.

5.3 Existing rankings

The previous section demonstrates that the use of AgeRank can expand the portion of web search results for which we have data – positive or negative – regarding their child-appropriateness. However, it does not show whether this data would be sufficient for a better ranking of results, i.e., a ranking more suitable for children. We explore this in the following two sections.

First, we measured the degree to which the relevance-rank of results agrees with the AgeRanks of results. As simulated children's queries, we expect that queries' results' relevance scores should correlate with their child-appropriateness because a page about a child's topic should be both more relevant to the query and more child-appropriate. Our approach was to compare two rankings of results – ascendingly and descendingly by AgeRank – by their relevance scores, with the hypothesis that for children's queries, the descendingly AgeRanked lists should be the more relevant of the two.

First, r_q was condensed (17) by removing any results for which no AgeRank was assigned. For example, if the results at positions 1 and 5 had AgeRank values, but not the results at positions 2, 3, and 4, the position of the result at 5 becomes 2 after condensing. Condensing is a method for assessing the performance of retrieval systems on result lists containing pages without assigned relevance values. Since we are comparing the relative performance of different orderings of AgeRanked pages, condensing allows us to focus on these pages and reduce the noise introduced by results for which no AgeRank exists.

Consider result list r_q as a list of 2-tuples ($r \times a$), where r is the reciprocal of the relevance rank (i.e., 1 over the position of the result in the list from 1 to 1000), and a is the AgeRank of the result. We created two new rankings of this list, r_q^+ and r_q^-, by reordering results according to their AgeRanks, descendingly and ascendingly, respectively (i.e., in r_q^+, results with higher AgeRank values appear at more relevant positions in the list). We used the reciprocal of the relevance rank as an approximation of a relevance score: results appearing at less relevant positions will have lower reciprocal relevance rank scores.

Conceptually, r_q^+ represents a possible (though simplistic) way to integrate AgeRank into result lists, by reordering results according to their AgeRank values. We compared these two rankings to determine which one was more relevant, where more relevant means having a greater concentration of relevance scores at better ranks. We measured relevance in terms of DCG, which captures these two qualities, creating scores DCG_q^+ and DCG_q^-. The difference between these values provides a measurement of the degree to which the relevance is concentrated into child-oriented results.

Indeed, 96.4% of queries in Q_{kids} produced results where DCG_q^+ was higher than DCG_q^-. This means that scores with higher AgeRanks tended to appear at more relevant position in the results. (See Section 6.1 for a more detailed examination of this.)

Next, we isolated the performance of AgeRank at the extremes. For each r_q we did the following. First, from r_q^+, we

drew the top and bottom 20 results and placed them into $best_q$ and $worse_q$, respectively. These represent the most and least child-appropriate pages. We measured the relative placements of the results in these lists within r_q for comparison. We call $before_n$ the set of the entries in $best_q$ with higher relevance ranks than the most highly ranked entry in $worse_q$, and $after_n$ the set of entries in $worse_q$ with higher relevance ranks than the lowest relevance-ranked entry in $best_q$. Conceptually, this portrays the relevance-overlap of the two sets. Table 4 depicts measurements characterizing the overlaps of $before_n$ and $after_n$ in terms of $best_q$ and $worse_q$ (left), and the L-P@10 and L-DCG scores for $best_q$ and $worse_q$ (right).

	N	%
$before_n$	13.32	81.36
$after_n$	6.27	39.63

Side	L-P@10	L-DCG
$best_q$	0.01936	0.01785
$worse_q$	0.00091	0.01493

Table 4: Average values of $before_n$ and $after_n$ across Q_{kids}(left) and L-P@10 and L-DCG for highest and lowest 20% AgeRank across all queries in Q_{kids}(right).

At this point we are left with potentially contradictory findings. The ordering of results for children's queries show that AgeRank is positively correlating with pages that are more relevant to the query, and hence more likely to be oriented toward young audiences, which provides evidence that AgeRank is performing correctly. However, the ordering also provides evidence that search engines already determine a reasonable ordering of results for children's queries. We argue that the latter is indeed not the case through the following counterpoints: First, consider the results in Table 4 (right). This table depicts the L-P@10 and L-DCG scores of $best_q$ and $worse_q$. Though $best_q$ is performing better than $worse_q$, it is not still not performing well in terms of producing good results for queries. The pages with the highest AgeRank values are not consistently appearing at the best positions. These positions, as shown in Section 5.2, are mostly populated by results for which no AgeRank value exists; though we cannot say for sure whether they are child-appropriate or not, we believe a system for children should limit children's exposure to pages for which we cannot make assessments. In this sense, without a reordering of results by AgeRank, relevance-based search is less effective at producing age-appropriate results for young audiences.

An alternative, such as limiting search to domains with a pre-existing classification, such as DMOZ, is likely to be inferior. Consider the L-P@10 scores in Tables 3 and 4 (right), which depict the number of appearances of a page with an AgeRank score in the top 10 results: the relative positions of scores from L_+ are worse than from L_{AR}. Highly relevant results are more likely to have higher AgeRank values.

5.4 Evaluated rankings

In the previous sections we demonstrated the accuracy of AgeRank by showing its correlation with human assessments of child-appropriateness. We demonstrated the potential effectiveness of AgeRank by showing that a reordering by descending AgeRank would produce more relevant results than a reordering by ascending AgeRank. One limitation is that we did not evaluate the child-appropriateness of search results by human assessors, and the findings of Section 4.2 are not necessarily importable to search since the evaluated pages were randomly sampled from L_{AR} rather than from real search results.

We expanded upon the findings in the previous sections to answer the question: given a reordering of search results based on AgeRank, how would the human perceptions (rather than system perceptions) of age-appropriateness be affected? The previous section showed that there is a tendency of results with higher AgeRank values to also have better relevance scores for children's queries. The question is whether manipulating result lists to promote the rank of results by their AgeRank would produce a more child-appropriate ordering in terms of human perceptions.

One problem with human assessments is that they are expensive – even at $0.01US per rating – and we had many results to be evaluated. Rather than evaluate each result, we instead isolated a region of interesting results. For a sample of results r_q, we created a pool of results by merging the $before_n$ and $after_n$ sets that we used in Section 5.3. Conceptually, this pool represents a region of ambiguity, where the results had either relatively high or low AgeRank, but not moderate. This is an important region for two reasons: First, given the fact that it contains high AgeRank values, it allows us to assess the pages that would appear first in an ordering of results by AgeRank; and second, that this region contains results whose positions would be relatively dramatically affected by a reordering of results by AgeRank values.

We could infer from the results in Section 4.2 that the higher AgeRanked pages would be more appropriate for children, but we desired to reaffirm this in the context of a more applied retrieval task. For each result in r_q, we created a request for a human assessment through Mechanical Turk in the manner described in Section 4.2.

Depending on the query, some of the $best_q$ and $worse_q$ values had very similar scores. We removed the lowest $best_q$ and highest $worse_q$ in equal amounts whenever the difference between them was ≥ 0.5. This removed 157 results from the 753 for which we collected ratings. This was to reduce the noise from results with quite close scores. Not only were these ratings more ambiguous in terms of human ratings – unsurprisingly, given the closeness of their scores – but in practice these are uninteresting to consider since their relative ordering would be the least changed.

Our results are as follows. We compared the human scores between $best_q$ and $worse_q$ across the 753 ratings. We found a significant difference through both Student's t-test (P-value of 0.013) and Mann-Whitney-U test (P-value of 0.029). In these regions of relevance-overlap, and in the context of search, we reaffirmed the finding that higher AgeRank correlate with human assessments of child-appropriateness.

6. EVALUATION: SEARCH QUERIES

In this section we continue our search-based evaluation but focus on the query side.

6.1 Relative query ordering and prediction

One of our assumptions in this evaluation has been that the queries in Q_{kids} reflect children's queries, and this justified our finding that reordering descendingly by AgeRank improves relevance over an ascending order was a sign of AgeRank performing correctly. We further investigated the relationship between the child-appropriateness of children's queries and the AgeRank values of those queries' re-

sults. We hypothesized that, since AgeRank reflects age-appropriateness, queries whose results had higher AgeRank values would be more age-appropriate.

In terms of user assessments, evaluating the child-appropriateness of queries is more involved than for pages. A page can be assessed individually, and somewhat independently of knowledge of its topic. On the other hand, a query for which the rater knows little about is much more difficult to assess. The rater would probably assess the query by searching for it online and examining pages, before making a determination. After inspecting the queries in Q_{kids} we determined that we knew quite little about many of them, and decided that collecting human assessments across all Q_{kids} would be problematic.

As an alternative, we took the following approach. We observed that many queries in Q_{kids} pertain to video games. The US has a standardized rating system for the age-appropriateness of video games called the ESRB[10]. This rating system provides several categories, including, among others: *Everyone* (all ages), *Everyone 10+* (all ages over 10), *Teen* (teenagers), and *Mature* (older than 17). Given these ratings, we considered queries pertaining to video games and measured whether queries that we consider to be more child-appropriate (by virtue of the distance between rankings as described in Section 5.3) were more likely to pertain to video games with more child-appropriate ratings.

To assess this, we did the following. For each result in r_q, we executed a search for the ESRB rating by using the *GamesRadar* API[11], which allows queries by keywords. We filtered the queries in Q_{kids} to those with video game ratings, then ordered these queries descendingly by the difference between their DCG_q^+ and DCG_q^- values (explained in Section 5.3). We call this ordering $games^+$. Conceptually, this ordering reflects the degree to which their relevance is concentrated in results with higher AgeRank values. For each query, we identified its ESRB games rating, creating a 3-tuple $(q \times k \times t)$, where q is the query, k is its position in $games^+$, and t is its ESRB rating. Here, higher k is associated with a lower likelihood of being for young audiences. We separated these 3-tuples into four lists: one for each rating in *Everyone*, *Everyone 10+*, *Teen*, and *Mature*.

Given these lists, we compared the values of k among them. This provides a measurement for how age-appropriateness varies among each ESRB label. We first ranked the four ratings by the median value of k for their tuple list. Then, for each rating, we calculated the difference in $games^+$ ranks between the tuple lists of that label and *Everyone*, using the Mann-Whitney-U test. Table 5 depicts the results. The relative ordering of each label's median ranks reflects the order of ESRB age-appropriateness, with each being significantly higher than *Everyone* except for *Everyone 10+*. Next, we merged *Everyone* and *Everyone 10+* into *Everyone'*, and compared it against a combination of *Teen*, *Mature*, and a merging of *Teen* and *Mature*. Table 6 depicts these comparisons.

These results provide two complementary forms of evidence. First, they provide further evidence that a query's child-appropriateness can validate the performance of AgeRank. In Section 5.3 we assumed that Q_{kids} reflects children's topics to show that AgeRank was performing well, by showing that the children's topics were generally met with results whose relevance was more highly concentrated in higher AgeRank results (by virtue of $DCG_q^+ - DCG_q^-$). These findings show that not only is there a positive relationship between the child-appropriateness of a query and the AgeRank of its results, but that this relationship is granular among age groups from children to younger teenagers to older teenagers.

Second, the results provide evidence that we can predict the child-appropriateness of queries by results. This is an interesting extension upon the findings of AgeRank. Where previous experiments supported the hypothesis that higher AgeRank predicted more child-appropriateness for a page, this finding supports the notion that the AgeRanks of pages, considered in aggregate as results for a query, have cumulative predictive power.

ESRB Rating	Rank	# Queries	$\Delta_{Everyone}$
Everyone	1	83	N/A
Everyone 10+	2	57	0.46
Teen	3	138	0.05
Mature	4	53	0.01

Table 5: Difference in ESRB game labels across queries. $\Delta_{Everyone}$ reflects the difference in ranks from an ESRB category to *Everyone* in terms of the P-value from a Mann-Whitney-U test.

ESRB Rating	$\Delta_{Everyone'}$
Teen	0.030
Mature	0.004
Teen + Mature	0.006

Table 6: Difference of ratings to *Everyone'*, reported as a P-value from a Mann-Whitney-U test.

6.2 Toward a children's query set

A major challenge in the evaluation of children's search is the lack of datasets. In particular, there is no existing collection of queries. We believe DMOZ topic titles are a reasonable approximation in the context of web search for children's pages, and can effectively be used by others seeking to experiment with children's search. In this section, we justify our use of Q_{kids} and attempt to address potential concerns.

One problem with our evaluation is that we used DMOZ as the source of both seed pages (L_+) and queries (Q_{kids}). The reader might question whether queries from DMOZ are inherently more likely to produce biased results, since Q_{kids} was derived from the categories from which L_+ was drawn. We address concerns about our use of DMOZ as follows.

First, we created a set of queries $Q_{small} \subset Q_{large}$ intended to be less biased toward DMOZ. We created Q_{small} as follows: First, we created a query set $Q_{virtual}$ by appending every 2-letter permutation of lowercase alphabet characters to the term "kids" (e.g., "kids aa", "kids ab"). We sent each of these queries to Google Suggest's (5) suggestion lookup table, which, for many of the queries in $Q_{virtual}$, produced a small number of suggested queries (based on them being frequently issued by Google users), and called this query set Q_{Google}. For each query $q \in Q_{Google}$, we removed the term

[10] `http://www.esrb.org/ratings/ratings_guide.jsp`
[11] `http://www.gamesradar.com/developer`

"kids" from the query, creating query q', and, if $q' \in Q_{kids}$, placed the query in Q_{small}. In total, this produced 89 queries.

Our rationale for the construction of Q_{small} is that, since Q_{kids} is based on DMOZ categories, there is a risk of bias toward pages already appearing in DMOZ and hence being more likely to have a higher AgeRank. By using Q_{small}, we filtered queries to those also appearing in Google's database, selecting queries with more evidence of having general appeal (as evidenced by being popular Google queries), and hence reducing the likelihood of the query being specific to DMOZ. Table 7 depicts a random sample of queries from Q_{small}.

Queries		
online stories	electronics	party games
japan	energy	events
homework help	science	jewelry
cuisine	music	opera
uganda	rhode island	costumes
sudoku	wrestling	projects
ozone layer	egypt	american idol
synonyms	radio	crafts
igre	issues	fashion
geometry	italy	nutrition

Table 7: Random queries from $Q_{small} \subset Q_{kids}$.

We separated the results of the experiments in Section 5.2 according to whether they were from queries in Q_{small} or $Q_{kids} \setminus Q_{small}$. The differences are reported in Table 8. The differences are small (though statistically significant) and in the direction away from bias (i.e., queries in Q_{small} perform better than queries from Q_{kids}).

Query source	L-P@10	L-P@100	L-P@1000
Q_{small}	0.126	0.175	0.177
$Q_{kids} \setminus Q_{small}$	0.118	0.137	0.143

Table 8: Comparison of precision between Q_{small} and $Q_{kids} \setminus Q_{small}$.

Finally, we conducted a Mechanical Turk evaluation of queries within Q_{small} to determine the likelihood that they were issued by children. As in Section 4.2, we presented queries to users and asked for a Likert assessment of the likelihood of the query being issued for or by children. In this evaluation, we collected 3 assessments for each query.

We removed from these rated queries any query that did not have at least 2 nonzero ratings that were of the same sign, considering the remaining queries to be in agreement. This produced 50 agreement queries, of which 39 were positive (i.e., rated as likely for or by children). These ratings suggest that Q_{small} are useful approximations of children's queries. We believe these ratings are generalizable to Q_{kids} given the overlap and similarity in scores.

7. CONCLUSIONS

In this section we reflect upon our findings and discuss future work. First, from our results, we offer a number of observations about the topology of the children's web. We then gather the findings of our evaluation to present our main claims. We finish with a discussion of limitations and future work.

7.1 Topology of the children's web

The data confirm our hypothesis that locality exists among pages that are appropriate for children. This locality is apparently both bidirectional and bipolar (i.e., applies to both being and not being a child's page), as evidenced by the correlations identified in Table 1. Furthermore, this effect was reflected in cases where the distance from L_+ was 4 hops or greater, providing evidence that the locality of the children's web may be deep. In other words, not only are children's pages more likely to link to children's pages, but they are more likely to link to pages that are themselves more likely to link to children's pages. Interestingly, the agreement among AgeRank and human scores were not dramatically different from the use of the original DMOZ labels and human scores. We believe this to be further evidence of the locality of the children's web, in that the propagation of labels across multiple hops did not strongly deteriorate the likelihood of the label being accurate. Finally, the children's web has a strong overlap with pages of high PageRank, as evidenced by the large overlap between high PageRank pages and L_{AR}, though PageRank does not correlate with being for children or for adults.

The "Kids and Teens" category of DMOZ is a reasonable representative of children's pages, and we assess it to be useful in the context of finding children's pages. However, it is not an ideal representation. As we showed in Section 4.2, the agreement between human age assessments and pages' inclusions in child or adult DMOZ categories was weaker than we expected. In particular, our selection of a set of pages for L_- was a weaker predictor of human assessment than the L_+. In fact, as mentioned in Section 4.1, a non-trivial portion of pages from L_- already existed within L_+ set and had to be removed. This emphasizes an important problem: the distinction between children's and adults' pages is often vague and subjective.

There is a trade-off between quantity and quality of assessments. A hand-picked seed set may produce AgeRank scores that are more consistent with human assessments, but have much lower coverage. Coverage is important both for number of pages assessed, but to help ensure that pages can accumulate both positive and negative evidence. The integration of a feature-based classifier may help clean and expand the seed sets.

7.2 Summary and implications of results

Our evaluation covered many aspects of AgeRank, so we gather the findings here into a more congruent argument. (1) A user study across over 3000 rankings showed that AgeRank scores correlate with human assessments of age-appropriateness of a page (Section 4.2). (2) PageRank is a very weak predictor of AgeRank (Section 5.1). (3) For much of the 7 million web pages with highest PageRank scores, AgeRank produced scores (Section 5.1). (4) Across Q_{kids}, we found a mean L-P@1000 of 0.143 for AgeRank values, with 99% of queries having at least one result with an AgeRank score (Section 5.2). (5) For the results in Q_{kids}, reordering results by descending AgeRank produced a more query-relevant ordering than reordering by ascending AgeRank (Section 5.3). (6) In regions of overlap in relevance-rank

between results of high AgeRank and low AgeRank, user assessments showed that the higher AgeRank pages were more age-appropriate (Section 5.4). (7) The distance in query-relevance between descending and ascending ordering of result lists by AgeRank correlates with a granular ground-truth labeling of child-appropriateness classes (ESRB) (Section 6.1). (8) Queries in $Q_{small} \subset Q_{kids}$ were generally rated to be child-appropriate by human assessors. This is likely to be generalizable to Q_{kids} (Section 6.2).

We combine these findings to make the following claims, which comprise the major contributions of this work. First, AgeRank effectively captures the age-appropriateness of a page (1), and is necessary because PageRank alone cannot effectively do so (2); the usefulness of both directions and polarities of linking, in this context, present a potential advantage over single-direction, single-topic approaches such as topical PageRank. The coverage of the automatically labeled pages is high, including many important pages (3), and results for many queries (4). AgeRank can potentially improve children's search, as for child-oriented queries (8), currently the number of pages assessed to be child-appropriate that appear in the best positions is low, and reordering by AgeRank concentrates more relevance to the top pages than reordering by reverse AgeRank (5), and very likely concentrates more age-appropriateness as well (6). Beyond individual pages, AgeRank can also be used to predict the child-appropriates of queries (7). Finally, evidence suggests that our approach to generating children's queries is reasonable (7, 8), which both helps justify findings pertaining to search results (4-6), and can be used by others wishing to execute similar studies on children's IR.

7.3 Limitations and future work

Our results show that relative AgeRank ordering has a positive effect on child-appropriateness, but we did not identify a specific score threshold to determine with confidence whether a page is for children or not. We pursued this work with the goal of better understanding the locality and linking of the children's web to inform a more elaborate ranking method. We plan to continue this work by building a complete search engine, including (1) the integration a feature-based classifier, for which the link-based information could be a complementary approach, and (2) a principled combination of relevance, authority (e.g., PageRank), and AgeRank.

We used adult assessors via Mechanical Turk to determine the child-appropriateness of pages. We believe this was a reasonable form of evaluation that is more practical than the difficult task of having children make assessments. Once we have a more complete search engine, we intend to evaluate our work more holistically with children.

One problem with considering the pages to which a given page links is the risk of gaming the algorithm by malicious parties, such as spam advertisers, since a page has no control over which pages link to it (resilient to tampering), while full control over where it links (vulnerable to tampering). As AgeRank is a modular score that can be easily combined with others, it could be complemented by spam detection algorithms. Furthermore, though the outgoing link scores were effective predictors, they are not necessary for AgeRank to function.

Acknowledgements The research leading to these results has received funding from the European Community's Seventh Framework Programme FP7/2007-2013 under grant agreement n^o 231507.

References

[1] Amazon Mechanical Turk. `http://www.mturk.com/`.

[2] The Clueweb09 dataset. `http://boston.lti.cs.cmu.edu/Data/clueweb09/`.

[3] KidsClick!: Project background. `http://www.kidsclick.org/clickback.html`.

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

[5] Query Suggest FAQ. `http://labs.google.com/intl/en/suggestfaq.html`.

[6] UK children's media literacy. `http://www.ofcom.org.uk/advice/media_literacy/medlitpub/medlitpubrss/ukchildrensml/ukchildrensml1.pdf`.

[7] What is Yahoo! Kids? `http://sp.askkids.com/docs/askkids/index.shtml`.

[8] 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.

[9] D. Bilal and J. Kirby. Differences and similarities in information seeking: children and adults as Web users. *Information Processing & Management*, 38(5):649 – 670, 2002.

[10] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. *Comput. Netw. ISDN Syst.*, 30(1-7):107–117, 1998.

[11] J. Dean and S. Ghemawat. MapReduce: simplified data processing on large clusters. *Commun. ACM*, 51(1):107–113, 2008.

[12] A. Druin, E. Foss, L. Hatley, E. Golub, M. L. Guha, J. Fails, and H. Hutchinson. How children search the internet with keyword interfaces. In *IDC '09*, pages 89–96, New York, NY, USA, 2009. ACM.

[13] G. W. Flake, S. Lawrence, and C. L. Giles. Efficient identification of Web communities. In *KDD '00*, pages 150–160, New York, NY, USA, 2000. ACM.

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

[15] T. H. Haveliwala. Topic-sensitive PageRank. In *WWW '02*, pages 517–526, New York, NY, USA, 2002. ACM.

[16] K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of IR techniques. *ACM Trans. Inf. Syst.*, 20(4):422–446, 2002.

[17] T. Sakai. Alternatives to Bpref. In *SIGIR '07*, pages 71–78, New York, NY, USA, 2007. ACM.

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