Clash of the Typings

Finding Controversies and Children's Topics Within Queries

Karl Gyllstrom and Marie-Francine Moens

Katholieke Universiteit Leuven, Belgium
{karl.gyllstrom,sien.moens}@cs.kuleuven.be

Abstract. The TadPolemic system identifies whether web search queries (1) are controversial in nature and/or (2) pertain to children's topics. We are incorporating it into a children's web search engine to assist children's search during difficult topics, as well as to provide filtering or mitigation of bias in results when children search for contentious topics. We show through an evaluation that the system is effective at detecting kids' topics and controversies for a broad range of topics. Though designed to assist children, we believe these methods are generalizable beyond young audiences and can be usefully applied in other contexts.

Keywords: controversy detection, children's search

1 Introduction

Consider the following claims about three topics in popular culture: (1) the recent *Twilight* book/film series is controversial as to whether or not it is sexist; (2) the subject of *capitalism*, though often discussed in mature news media, is likely to be interesting to many children; and (3) there is disagreement about whether *Nas*, an American rap artist, is a member of the *illuminati*. If you share the view of the authors, you will likely find these claims to be surprising, obscure, or even nonsensical. Yet, much data supports them: the query "twilight is sexist" has been issued to Google 190,000 times, "twilight is not sexist" has been issued 141,000 times, "capitalism for kids" has been issued 3,620,000 times, and "nas is/is not illuminati" 206,000 and 126,000 times, respectively[1]. Though they may not be obvious to us, many aspects of topics are evident in the way users seek information. In this work, we explore the use of queries to identify both controversial and children's topics. To this end, we have developed the TadPolemic system, which identified the above examples, among many others.

There are many potential applications of this system, but we mention here a few that we plan to incorporate into a children's search engine that we are building. First, TadPolemic can inform search engines that are designed to provide assistance or supervision to child users, who may require special treatment when searching the web. For example, if we detect that a query does not pertain

[1] As determined via the Google Suggest feature [2], described later.

to typical children's topics, we may assume it to be of advanced nature (e.g., calculus), and alert the system to provide greater assistance to the child (studies have called for such support, e.g., [6]). Controversial topics, for which children are still forming personal positions, may need extra care by the system to reduce the volume of biased information, as children may be less capable than adults at filtering information and making informed judgement of the material [10]. Next, TadPolemic's entity detection can create an extension to existing kids' topic listings, such as that of Dmoz [1], to enable users to browse by topic, as children often prefer to do [4].

The importance of providing such assistance – and protection – to young users cannot be overstated. A 2005 study revealed the prevalence of computer use in US households: 77% of children aged 5-6 had used a computer, 42% of whom had used websites, and 22% of whom were able to browse to websites unassisted [5]. A UK study reports that a surprisingly large number of children sometimes access the web with no parental supervision: as many as 68% for children aged 5-7, and 84% for children aged 8-15 [11]. Indeed, children are capable web users, and systems should embrace young audiences rather than assume vigilant supervision. However, we stress that although our work is designed to assist children, we believe it is generalizable beyond young audiences and can be usefully applied in other contexts.

To address the detection of both kids' entities and controversies, TadPolemic applies a simple, query-side approach in detecting the topical nature of queries. Specifically, frequently-issued queries are used as a measure of community sentiment toward various topics. As an example of the former, appending the terms "for kids" to a query (e.g., transforming "science puzzles" to "science puzzles for kids"), then determining that the new query is frequently issued, is an indication that the topic is interesting to children; on the other hand, the low frequency of a query such as "multivariate calculus for kids" indicates that the subject of "multivariate calculus" is unlikely to be appealing to children.

We specify 5 advantages of our query-side approach. (1) It is simple and easily adoptable. (2) Queries have a potentially faster time-to-discovery over a possible content-based approach that must await web articles being written. For example, a sudden news event about a novel topic may trigger a spike in queries about that topic, which would be immediately available to a query-side system, while a content-side approach would have to diligently crawl new articles within which to identify controversies. As we show later, the use of queries to detect kids' and controversial topics may have a temporal advantage over content-based approaches. (3) Queries are potentially more reflective of the general public's sentiments. Whereas articles, such as Wikipedia, must generally fit a neutral standard, there are no such limitations on queries; hence, they may be more revealing of users' feelings. (4) Queries are the most compact representations of users' information interests, unlike web pages which contain dramatically more information that must often be filtered, cleaned, or otherwise reduced. Finally, (5) systems such as Google's and Yahoo!'s suggestion features demonstrate that the access of related queries (which we exploit) can be accomplished in real-time, allowing TadPolemic the benefit of operating at interactive speed. In fact, this quality enables TadPolemic to be implemented as a search engine layer that responds at query time.

2 Kids' Entity Detection

2.1 Overview

We refer to a *KidsEntity* as a subset of query terms that represent a topic of interest to children, and the objective of TadPolemic is to discover KidsEntities from queries. For example, for the query "when was Mickey Mouse created?", TadPolemic discovers "Mickey Mouse" as a KidsEntity. For this, we use a query-based approach that applies the knowledge of the community. Specifically, we use the Google Suggestion (GS) service [2], which provides query suggestions that are based on queries that other users issue. The interaction with GS works as follows: TadPolemic sends a query to the service, which responds with a set of 0-10 queries that are recommended completions of the query, as well as the frequency of each suggested query being issued to Google's web search. Conceptually, this serves as a coherence check for a query; if it is frequently issued, it is more likely that the query is meaningful [9]. The task of KidsEntity detection is as follows:

1. For query q_i comprising terms $t_i \in q_i$, find each subsequence $s_{\langle i,k \rangle}$ of k adjacent terms from 1 to the query length.
2. For each subsequence $s_{\langle i,k \rangle}$, form a *KidsQuery* by appending the terms "for kids" to the subsequence's terms.
3. This KidsQuery is issued to GS, and the suggestions (if any) are checked for the presence of the query itself (i.e., the query is a suggestion for itself).
4. The longest subsequence $s_{\langle i,k \rangle}$ for which its KidsQuery is in the GS is used as the KidsEntity for q_i.

For example, the query "who is Mickey Mouse" would generate the subsequences "who", "is", "Mickey", "Mouse", "who is", "is Mickey", "Mickey Mouse", "who is Mickey", "is Micky Mouse", as well as the full query itself. Of these subsequences, "Mickey Mouse" is the only one for which its KidsQuery "Mickey Mouse for kids" is in the GS suggestions.

Related work Entity detection within queries has not been studied extensively. Paşca describes a method where query templates (e.g., "how much does X cost") are discovered by processing the surrounding text from a large number of queries containing a particular term [12]. Our work adopts a similar template-based approach, though our focus is different: rather than attempt to identify templates through which similar entities can be discovered, we seek to identify binary properties – e.g., being for kids – and, in the case of controversy (described later), the nature of that property (i.e., the particular dimensions along which disagreements align).

2.2 Evaluation

The detection of KidsEntities should be both accurate (i.e., they actually correspond to kids' topics) and have a high coverage (i.e., they are detected in cases where queries pertain to kids' topics). This section is divided into the separate evaluation of these criteria, but we begin with our experimental setup. Let D_{kids} be the set of topics on DMOZ [1] within the *Kids and Teens* main category, which contains a nested hierarchy of children's topics as well as links to child-appropriate web pages within those topics. Let D_{adult} be the set of topics on DMOZ not within the Kids and Teens category. We filtered both D_{kids} and D_{adult} to the *Science*, *Arts*, and *Society* subcategories, then collected all of the subcategories within these two sets of topics by collecting the name of any topic appearing as a directory within the hierarchy. For example, DMOZ lists *Kids_and_Teens/Sports* and *Kids_and_Teens/Sports/Basketball*, from which we would draw *Sports* and *Basketball* as topics. Conceptually, D_{kids} represents a set of topics that are more likely to be for children, or to have child-friendly facets. Conversely, D_{adult} represents a set of topics that are less likely to meet this criterion (note that this distinction does not imply being adult-oriented).

Accuracy Let $Q_{all} = \mathrm{D}_{kids} \cup \mathrm{D}_{adult}$, comprising all the topics contained in either D_{kids} or D_{adult}, used as queries. Each query in Q_{all} was run through TadPolemic's entity detection process, generating a query set $Q_{TP} \subseteq Q_{all}$ comprising queries for which a KidsEntity exists. For each query $q \in Q_{TP}$, we checked the source (D_{kids}, D_{adult}) from which it was derived, allowing us to measure the extent to which the various sources included KidsEntities. As reported in Table 1, D_{kids} includes a substantially larger proportion of KidsEntities than D_{adult}; this serves as a validation, in that queries determined to be for children by TadPolemic are more likely to have explicit child labels by DMOZ.

We extended upon these findings to include an assessment of the child-friendliness of pages produced by web searches using these queries. This contributes a more applied assessment, as actual web searches and pages are used. As child-friendliness labels are not generally available for the web (as they are in DMOZ), we assessed child-friendliness by using demographic information made available by the Alexa database [2]. For a given site, Alexa may list the distribution of visitors who have children, relative to the general population, on a scale of -2 to 2, with -2 being much less likely to have children than the general population, 2 being much more likely, and 0 being similar to the general population. Our belief was that child friendly sites are more likely to be visited by households with children than households without children. We confirmed this by comparing the Alexa scores between a large sample of 5899 pages from D_{kids} and 1695 pages from D_{adult} which showed the average D_{kids} score to be statistically significantly higher ($\mu_1 = -0.17, \mu_2 = -0.43$, t-test p-value $\ll 0.0001$).

[2] `http://alexa.com`

Source	$\cap Q_{TP}$	Total	Ratio	Δ
D$_{kids}$	1132	1923	0.59	$\ll 0.001$
D$_{adult}$	2448	18529	0.13	

Table 1. Inclusion of Q_{TP} within various sources. Δ indicates difference between sources' proportions, reported as p-value of Fisher's test.

Source	$\in$		#	KidsQuery		Regular Query	
	Q_{TP}	D$_{kids}$		$\bar{x}$	Δ_{prev}	$\bar{x}$	Δ_{prev}
$Kids_1$	Yes	Yes	1132	0.37	N/A	-0.47	N/A
$Kids_2$	Yes	No	2448	0.28	$\ll 0.01$	-0.48	0.90
$Kids_3$	No	Yes	791	-0.02	$\ll 0.01$	-0.44	0.03
Non_-	No	No	16081	-0.32	$\ll 0.01$	-0.61	$\ll 0.01$

Table 2. Alexa ratings. Δ_{prev} indicates difference with previous row's mean Alexa rating, reported as p-value of Student's t-test.

For each query in Q_{all} we issued a web search[3], collecting 5 results, for which we looked up the Alexa ratings where available. In addition, we executed a search for each query using its KidsQuery variant (i.e., by appending "for kids" to it). We created 4 categories of queries for this examination: $Kids_1$ include queries from D$_{kids}$ that were identified to have a KidsEntity by TadPolemic (D$_{kids} \cap Q_{TP}$); $Kids_2$ have a KidsEntity, but were not in D$_{kids}$, while $Kids_3$ include queries from D$_{kids}$ without a KidsEntity; Non_- include queries that were neither found to have a KidsEntity, nor were in D$_{kids}$. This serves to compare the quality of Dmoz and TadPolemic labels. Table 2 depicts the average Alexa score of the top 5 search results for each query. The differences among sources in terms of Alexa ratings for KidsQueries are significant in each case. In the case of the regular (non-altered) query, the first 3 variants perform similarly. From the data we conclude that a query having a KidsEntity or appearing in D$_{kids}$ are both strong indicators of that query's pages' child-friendliness ratings, and, in the case of the KidsQuery variants, having a KidsEntity is a *stronger* indicator of Alexa rating than being in D$_{kids}$. TadPolemic's labels have similar quality to Dmoz, though TadPolemic can be applied to new topics, while Dmoz is limited to the categories created by the site authors. This adds evidence that the detection of KidsEntities can help orient web searches to children's web pages.

Coverage In this section, we study the use of TadPolemic on actual user queries to demonstrate how it might perform in natural contexts. For this purpose, we used the children's query log proposed by Duarte et al. [7], to which we refer as Q_{AK}. The query log is a subset of the AOL query log[4] that is likely to pertain to children's queries, based on the landing site of the query and various other properties of the search session. Though not proven that the queries are from or on behalf of children, for our purposes it suffices as a useful approximation.

For each query $q_i \in Q_{AK}$ we used TadPolemic to identify a KidsEntity if available. Of 2332 queries, KidsEntities were detected in 2119 (90%), providing

[3] We report results from Google web search, though results from Yahoo! and Bing did not significantly affect our results.

[4] We understand that the use of this query log is controversial due to privacy issues. The subset identified by Duarte et al. was constructed with care to avoid exposing potentially sensitive personal data [7].

further evidence of a relationship between the detection of KidsEntities and the child-appropriateness of queries. We then studied the queries from the side of web search results to validate the detected KidsEntities. We had two goals: first, to determine whether the results for queries containing a KidsEntity are more child-oriented, and second, to determine whether the KidsEntities extracted from TadPolemic are semantically meaningful representations of the query. First, we examined the child-friendliness of results. For each $q_i \in Q_{AK}$, we ran a Google web search on q_i and drew the top 5 results, which we refer to as SR_i. For each result $r_j \in SR_i$, we looked up the Alexa children's rating, and averaged the value of the 5 ratings across the SR_i. Of the 2108 queries with KidsEntities and ratings, the mean Alexa rating was 0.824, while the mean rating among the 212 queries without KidsEntities was 0.383. This difference is significant (p-value $\ll 0.0001$ by Fisher's test), echoes our previous findings, and reinforces the fact that KidsEntities are more likely to yield pages suitable for children.

Regarding the second goal, we examined the topical-similarity between a query and its KidsEntity to ensure that, in addition to being more suitable for children, the KidsEntity retained the semantic value of the original query (i.e., extracting the KidsEntity from a query did not substantially change that query's meaning). For each result $r_j \in SR_i$, we looked up the popular *del.icio.us* tags for that page, where available, using the API call *urlinfo*[5], adding the tag to the query's tag set T_{SR_i}. If the KidsEntity k_{q_i} for q_i was in T_{SR_i}, the search result was considered relevant, and k_{q_i} was therefore considered to be a topically-relevant representation of q_i. Of 1768 queries for which a tag existed within their web results, 1273 had a matching tag (72%), indicating a strong topical overlap. Note that our use of *del.icio.us* is an approximation of relevance, as there are not human relevance assessments of the queries and web pages used in this experiment.

3 Controversy Detection

3.1 Overview

Controversy detection uses a related approach to that of the KidsEntity detection. The approach applies the frequency of what we refer to as *claim queries*, or queries of the form "X [is/are/was/were] Y", which can provide insight into the community's sentiments – and disagreements – on popular web search topics. Given a topic T_i, we create a set of *claim* queries by appending, individually, the verbs "is", "are", "was", and "were", to create four query variants. For example, given the topic *war* we would create the queries "war is", "war are", etc. These queries are then dispatched to the GS service, and the list of suggestions are examined. For each suggestion s_k in this list, we draw the terms t_{s_k} appearing after the query's verb (e.g., from the query "global warming *is* real", we pull the term "real"), and add these terms to a set of *claim terms* C_{T_i} for T_i. Next, we create a set of *negation terms* C'_{T_i}: for each term $t_{s_k} \in C_{T_i}$, we create its antonyms

[5] http://www.delicious.com/help/json

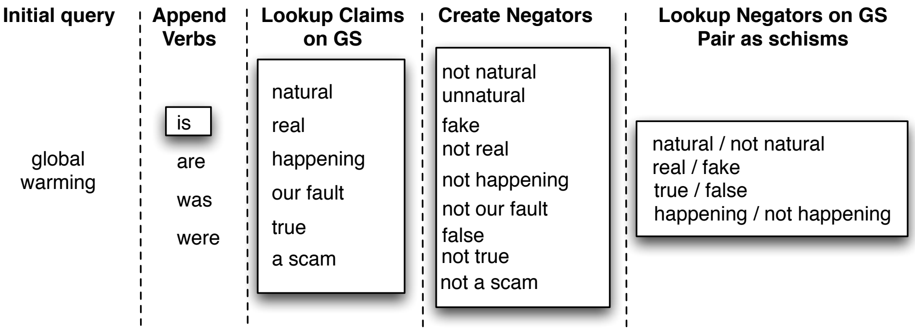

Fig. 1. Controversy lookup.

by (1) Wordnet [8] lookups, and (2) creating negating terms by prepending the word "not" to the term (e.g., "real" becomes "not real"). These negating terms are used to construct *anti-queries* (e.g., "global warming is *not real*"), and we remove those not appearing within GS. For each query/anti-query pair, we create a *schism* Z_j that is composed of a source term $t_{s_k} \in C_{T_i}$ and its negation term $t'_{s_k} \in C'_{T_i}$. See Figure 1 for a depiction of this process.

We treat the existence of a schism for a topic as evidence that the topic is (at least partially) controversial. Of course, controversy is a complex quality, and the nature and number of schisms should both be considered when assessing the depth of the disagreement. In fact, we later describe ways to identify highly contentious topics based on the presence of certain terms within their schisms. An important aspect of our work is not only that controversy is detected, but that we detect the particular dimensions along which disagreements align. Finally, an approach similar to the KidsEntity detection could be used for controversy detection on more complex queries (e.g., from the query "what is global warming" we could extract "global warming" as a popular entity, for which controversy detection is run).

Related work The problem of controversy ranking on Wikipedia was explored by Vuong et al. as a means to identify pages of more significant dispute [13]. Their approach studies the reversal of modifications to articles – particularly when the reversals and modifications are generated by established contributors for whom such events are rare. Our approach is complementary in that we study claims made within queries, although we share interest in Wikipedia as a source of controversy and platform for testing the detection thereof.

3.2 Evaluation

We explore the validity of our controversy detection through the use of Wikipedia as both an explicit reference of controversial topics via labels, as well as an implicit reference via the content of articles and their discussion pages.

Topic detection We created a list of topics to use as input for TadPolemic's controversy detection using popular Wikipedia topics; we selected the article titles of the 3000 most frequently viewed Wikipedia articles during an hour of August 25th, 2010 [6], which we refer to as T_{wiki}. Of these, TadPolemic labeled a subset V_{TD} as controversial. For comparison we used the list of controversial topics available on Wikipedia[7], which we refer to as V_{wiki}[8]. This page contains a listing of Wikipedia articles that, for various reasons, experience a large degree of edit conflicts, and represent topics for which disagreement and controversy likely exist. We identified 384 topics in V_{TD}, 277 topics in V_{wiki}, and 100 topics in both, meaning 36% of the topics in V_{wiki} were identified by TadPolemic, while 74% of the topics discovered by TadPolemic were not in V_{wiki}.

Comparison of topic qualities Though TadPolemic effectively identified many controversies contained within V_{wiki}, we further studied the topics in V_{wiki} not identified by TadPolemic and observed some limitations with these topics. First, many pertain to issues that may be obscure or otherwise unpopular. Very few of these topics had a *claim query* (only 28, or 15%), meaning that TadPolemic could not draw any useful information. We measured the relative unpopularity of these topics (as compared to those found by TadPolemic) by comparing the query frequencies for topics within V_{TD} and V_{wiki}, which are reported in Table 3 (frequency columns). The differences are quite pronounced, and each source is significantly different from the others. From this we conclude that the topics in V_{wiki} that were not identified by TadPolemic tend to be more obscure; this would make them less likely to appear as common queries, a property upon which TadPolemic relies, and therefore less likely to be issued by users.

A related problem is the timeliness of articles. For example, since TadPolemic is based on queries, it is theoretically more adaptive to novel controversies than Wikipedia, as Wikipedia must await discussion by editors. We measured this effect, with the hypothesis that topics detected by TadPolemic, yet not appearing in V_{wiki}, would be more recent. This is depicted in Table 3 (recency columns): indeed, controversies discovered by TadPolemic are more recent than topics listed by Wikipedia as controversial. We would recommend the combined use of both TadPolemic and Wikipedia to maximize coverage of controversial topics.

Contentious issues We isolate certain schisms as *contentious*, or involving disagreement about topics that are polarizing or highly sensitive. We accomplish this by simply checking the schisms for the presence of a preselected set of sensitive terms, including *right, wrong, true, false, guilty, innocent, safe, dangerous, legal, illegal,* and *evil.* We call the set of topics containing these terms V_{hot}. We feel that this set is more important to capture, since the potential sensitivity among the audience is higher.

[6] Dumps of traffic are available at `http://dammit.lt/wikistats/`.

[7] `http://en.wikipedia.org/w/index.php?title=Wikipedia:List_of_controversial_issues&oldid=386446018`

[8] We removed from V_{wiki} any topic that was not contained in T_{wiki}.

Source	#	Frequency		Recency	
		$\bar{x}$ frequency	Δ_{prev}	$\bar{x}$ age (days)	Δ_{prev}
$V_{TD} - V_{wiki}$	284	233452813.38	N/A	2887.691	N/A
$V_{TD} \cap V_{wiki}$	100	192684800.00	0.42	3017.863	0.093
$V_{wiki} - V_{TD}$	177	60927305.08	0.0035	3249.621	$\ll 0.0001$

Table 3. Frequency and recency of topics from various sources. Δ_{prev} reports the difference of means between a row and its preceding row, reported as p-value of a Student's t-test.

Source	$S \cap V_{wiki}$	#	Ratio	Δ
V_{hot}	38	59	0.644	$\ll 0.0001$
$V_{TD} - V_{hot}$	65	269	0.242	

Table 4. Difference of inclusion between V_{hot} and $V_{TD} - V_{hot}$, reported as p-value of Fisher's test.

Source	# contr	# total	Ratio	Δ_{prev}
V_{hot}	64	93	0.688	N/A
V_{TD}	217	386	0.562	0.0344
T_{wiki}	1418	2949	0.481	0.0029

Table 5. Proportion of topic pages containing controversial terms. Difference reported as p-value of Fisher's test.

First, we compared V_{hot} to V_{wiki}. Our hypothesis was that topics of more severe disagreement would be more likely to appear in Wikipedia's controversial list. As depicted in Table 4, this is indeed the case: V_{hot} are substantially more likely to be included in V_{wiki}. Next, we continued with a simple, content-oriented approach. For each page in T_{wiki}, we checked the contents of the page for occurrences of the terms "controversy" or "controversial". We compared the prevalence of pages with these terms between V_{hot} and V_{TD}, with the result depicted in Table 5. The prevalence of these controversial terms is significantly higher in topics which TadPolemic has detected to be controversial, which are significantly higher than Wikipedia pages in general (as indicated by the prevalence for T_{wiki}).

Schism detection The approach we took in this experiment was to evaluate whether the schisms identified by TadPolemic were also visible in the editor discussions on Wikipedia about the topic. We believed that a particular schism (e.g., whether global warming is real or not) would manifest in the commentary about changes to the article, as users may more aggressively revise or delete text about these issues and document the modifications. For each topic $t_i \in V_{TD}$, we drew the text comments of the 3000 most recent edits[9] and placed them into a single document for the topic, which we then indexed into a Lemur text index [3]. For each topic $t_k \in V_{TD}$, we selected the terms from its schisms, creating schism-set S_{t_k}. Using this schism-set, we created a query set Q_{merged}, containing a query q_k for each S_{t_k} by merging the unique terms of S_{t_k} together. Next, we

[9] For example, the edit-history of the *global warming* topic is `http://en.wikipedia.org/w/index.php?title=Global_warming&action=history`.

Method	succeeded	failed	Success Ratio	$\bar{x}$ position (matches)	$\bar{x}$ results	Position ratio
Q_{merged}	269	108	0.714	27.710	112.724	0.246
Q_{indiv}	1172	1211	0.492	34.514	84.572	0.408

Table 6. Success rate and search results of queries in the two scenarios. Position ratio indicates the average region in which the correct result is found; for example, Q_{merged} queries tend to appear in the first quartile of results.

created query set Q_{indiv}, containing a query q_m for each schism $s_m \in S_{t_k}$ composed only of the terms within the particular schism. These queries were then issued to the Lemur index[10], retrieving a list of documents corresponding to the Wikipedia edit-history of the topic. Conceptually, this evaluation identifies links between schisms detected by TadPolemic and discussions regarding potential disagreement on Wikipedia. The number of queries for which the document was returned is depicted in the left columns of Table 6, revealing that 71% of merged schisms retrieved the topic from which they were derived. We also examined the positions of the correct results within these query results. Ideally, a schism should be a strong match for the discussion text of a Wikipedia article, as indicated by it appearing at a better (lower) position. Statistics about the positions of the correct result within query's results are depicted in the 4 right columns of Table 6. We observe that the TadPolemic schisms identified for a topic are generally matched to the Wikipedia discussion topics to which they correspond. Note that this experiment is quite coarse, as it assumes that schism terms are mentioned as comments in Wikipedia discussion pages (as they often are not). However, these results show that some degree of connection exists.

4 Concluding Remarks

In our evaluations, we presented a large number of findings that we summarize here:

1. The percentage of queries with KidsEntities appearing in D$_{kids}$ is high, and significantly higher than the queries appearing in D$_{adult}$ (Table 1).
2. For a query, there is a strong connection between the existence of a KidsEntity within it and a higher child-appropriateness rating for its search results (Table 2). This is nearly as strong as the topic being manually labelled as child-oriented (D$_{kids}$).
3. The connection in (2) is stronger than the also-positive connection between a query being in D$_{kids}$ and its child-appropriateness rating (Table 2), when using the KidsQuery variant.
4. For actual queries, the connection between a query having a KidsEntity and child-appropriate ratings was strong (Section 2.2).
5. For actual queries with children's landing pages, the prevalence of KidsEntities was high and query results tended to include pages pertaining to those KidsEntities (Section 2.2).

[10] Lemur was configured to use KL-divergence and default values.

6. 36% of the topics in Wikipedia's controversy list were detected, while only 26% of the topics TadPolemic identified as controversial were in the list (Section 3.2), and contentious topics detected by TadPolemic were even more likely to be listed on Wikipedia's list (Table 4).
7. The topics detected by TadPolemic were more popular and newer than those on the list that were not detected (Table 3); those that were not had low (15%) incidence of claim queries.
8. The Wikipedia pages for topics determined by TadPolemic to be controversial were more likely to contain terms like "controversy", and this was more pronounced with topics identified to be contentious by TadPolemic (Table 5).
9. The schisms identified by TadPolemic were correlated with the discussion pages of their Wikipedia topics (Table 6).

We connect these findings into the following conclusions: the KidsEntities detected by TadPolemic are accurate (1-3) and have a broad coverage in actual web queries (4-5). The controversies detected by TadPolemic are accurate (6, 8), and potentially more broad, popular, and timely than those specified by Wikipedia's controversial list (7). Evidence suggests that the schisms detected are accurate (9).

Though TadPolemic is still in a formative stage, we believe much of the technology could be simply implemented in a usable system. GS is designed to execute at interactive rates, specifically to offer suggestions as the user types a query. In terms of requests per user query to the service, our system generates a volume that is comparable to the volume that a typical Google search user would generate. For example, the KidsEntity extraction on the query "who is Mickey Mouse" would generate 10 requests to GS (each subsequence of 1 to 4 terms), while a user typing the query on the Google website would generate 19 requests (once per character entered). The controversy detection requires a similarly manageable number of requests. In this respect, an implementation of TadPolemic in an online search engine could be as simple as a thin layer between the user and a web search engine. The simplicity of our approach – a single URL call to Google – is an asset in this regard.

Limitations Due to the inherent difficulty of evaluating kids' entities and controversies, we used some comparison data sets that are imperfect. We emphasize that the use of the Wikipedia controversy list is a very coarse approximation of a gold standard. The inclusion of pages within this list is subject to the presence of "edit-wars", which are unlikely to occur for the vast majority of topics for which some disagreement exists. Furthermore, inclusion on this page is a ephemeral matter, and topics may enter and exit as disagreements are mediated via Wikipedia's community. Similarly, the use of Wikipedia articles containing the term "controversy" is also coarse; the presence or absence of the term "controversy" is an extremely simple test, and subject to the peculiars of Wikipedia's structure (e.g., many large topics are distributed among many linked articles, only one of which may contain discussion of controversies pertaining to the topic). Our approach also assumes that the particular schisms will be mentioned directly within the comments, though in practice this is not nearly comprehensive. Still, the connections we identified are a sign that there is reasonable overlap. Finally, the use of *del.icio.us* tags as relevance assessment suffers the limitation of sparsity in number and variety of tags; though it has the advantage of providing easy, fast, and cheap relevance information. Despite the flaws of these comparisons, we believe that our results characterize a system that is effectively performing the function that we intended it to perform.

On the other hand, we perceived the use of human assessment to also be sensitive to errors. The reason is that (1) controversy is a largely subjective matter, and (2) our system identified controversy in an extremely large number of topics, many of which we were not initially familiar with (though brief research confirmed their existence), and included schisms ranging from obscure (whether *oxygen* is flammable, and whether *Nas* (American rapper) is a member of the illuminati) to the very recent (whether *twitter* is useful, whether the films *Toy Story 3* and *The Last Airbender* were good or bad). Nonetheless, we consider human evaluation to be an essential future direction of our work.

References

1. ODP – Open Directory Project. `http://www.dmoz.org/`.
2. Query Suggest FAQ. `http://labs.google.com/intl/en/suggestfaq.html`.
3. The Lemur Project, 2001-2008. `http://www.lemurproject.org`.
4. D. Bilal. Draw and tell: Children as designers of web interfaces. *ASIST*, 40(1):135–141, 2003.
5. S. L. Calvert, V. J. Rideout, J. L. Woolard, R. F. Barr, and G. A. Strouse. Age, Ethnicity, and Socioeconomic Patterns in Early Computer Use. *American Behavioral Scientist*, 48(5):590–607, 2005.
6. A. Druin, E. Foss, H. Hutchinson, E. Golub, and L. Hatley. Children's roles using keyword search interfaces at home. In *CHI '10*, pages 413–422, New York, NY, USA, 2010. ACM.
7. 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.
8. C. Fellbaum. *WordNet: An Electronic Lexical Database*. Bradford Books, 1998.
9. K. Gyllstrom and M.-F. Moens. A picture is worth a thousand search results: finding child-oriented multimedia results with collAge. In *SIGIR '10*, pages 731–732, New York, NY, USA, 2010. ACM.
10. S. G. Hirsh. Children's relevance criteria and information seeking on electronic resources. *J. Am. Soc. Inf. Sci.*, 50(14):1265–1283, 1999.
11. Ofcom. UK children's media literacy, March 2010. `http://stakeholders.ofcom.org.uk/market-data-research/media-literacy/medlitpub/medlitpubrss/uk_childrens_ml/`.
12. M. Paşca. Weakly-supervised discovery of named entities using web search queries. In *CIKM '07*, pages 683–690, New York, NY, USA, 2007. ACM.
13. B.-Q. Vuong, E.-P. Lim, A. Sun, M.-T. Le, and H. W. Lauw. On ranking controversies in wikipedia: models and evaluation. In *WSDM '08*, pages 171–182, New York, NY, USA, 2008. ACM.