Characterizing Web Content, User Interests, and Search Behavior by Reading Level and Topic

Jin Young Kim
University of Massachusetts, Amherst
140 Governors Drive
Amherst, MA U.S.A. 01003
jykim@cs.umass.edu

Kevyn Collins-Thompson
Paul N. Bennett, Susan T. Dumais
Microsoft Research
One Microsoft Way, Redmond, WA 98052
{kevynct, pauben, sdumais}@microsoft.com

ABSTRACT

A user's expertise or ability to understand a document on a given topic is an important aspect of that document's relevance. However, this aspect has not been well-explored in information retrieval systems, especially those at Web scale where the great diversity of content, users, and tasks presents an especially challenging search problem. To help improve our modeling and understanding of this diversity, we apply automatic text classifiers, based on reading difficulty and topic prediction, to estimate a novel type of profile for important entities in Web search – users, websites, and queries. These profiles capture topic and reading level distributions, which we then use in conjunction with search log data to characterize and compare different entities.

We find that reading level and topic distributions provide an important new representation of Web content and user interests, and that using both together is more effective than using either one separately. In particular we find that: 1) the reading level of Web content and the diversity of visitors to a website can vary greatly by topic; 2) the degree to which a user's profile matches with a site's profile is closely correlated with the user's preference of the website in search results, and 3) site or URL profiles can be used to predict 'expertness'— whether a given site or URL is oriented toward expert vs. non-expert users. Our findings provide strong evidence in favor of *jointly* incorporating reading level and topic distribution metadata into a variety of critical tasks in Web information systems.

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

General Terms: Experimentation, Human Factors, Measurement.

Keywords: Web search, log analysis, domain expertise, reading level prediction, topic prediction.

WSDM'12, February 8-12, 2012, Seattle, WA, USA.
Copyright 2012 ACM 978-1-4503-0747-5/12/02…$10.00.

1. INTRODUCTION

The World Wide Web is a diverse source of information for billions of Web users. This variety provides a significant challenge in enabling a user's access to information because large portions of the Web may fall outside of a particular user's overall interests, comprehension level, or comprehension within a particular domain. Specifically, an expert user in a given domain would likely be interested in different types of content than novice users. Furthermore, this expertise or comprehension level may vary for a particular user across domains. For example, a person might be an attorney by profession and have high proficiency in understanding content in the legal domain, but have much more limited knowledge in the medical domain, and thus prefer less technical material when searching for symptoms or remedies on health topics.

In this work, we aim to model this diversity for information retrieval systems by defining a novel form of probabilistic profile that can be used to describe users, queries, or websites – major entities of Web search. Our profile is a *probability distribution* of *reading level and topic* that we call an *RLT profile*. To compute a RLT profile for any entity, such as a website, user, or query, we first get a set of one or more URLs associated with that entity using sources such as click data or Web domain relationships. For example, a user profile might be associated with the URLs of previously clicked search results, or a website profile might be associated with the URLs making up the website content. We use automatic text classifiers to compute the RLT profiles (distributions over reading level and topic) for each URL in the set. Finally, we aggregate the distributions of the individual URL profiles to obtain the combined RLT profile of the entity. The resulting distribution is a compact yet general representation that enables novel characterizations of users, queries and websites and the interactions between them. We can then derive further useful properties of the profile's distribution, such as the expectation and entropy, to characterize interesting properties such as the diversity of topics available at a website.

Because our probabilistic profile is a distribution, this also enables a principled comparison between any two entities, by comparing their RLT distributions using information theoretic divergence metrics. Finally, using the relationship between users, queries and websites extracted from session logs, we can characterize each entity in terms of related entities. For instance, we can build a profile of a given website based on the profiles of its visitors to compare/contrast characteristics of the content of a site and its target audience.

Using search log data of 7,600 users over a 10-week period, and millions of URLs, we show how these profiles can be built and used to gain new insights into content and behavior on the Web. Specifically, we provide a new characterization of important search tasks, assess new features for personalization, and distinguish between websites oriented toward expert or non-expert users.

2. RELATED WORK

To our knowledge, this is the first study that characterizes Web content and user behavior in Web search in terms of both reading level and topic prediction at Web scale. Previous work can be divided into three main areas: using topic and reading difficulty prediction independently to improve Web search, modeling user and content familiarity, and characterizing domain expertise.

In Web retrieval, previous work tried to address the diversity among users by personalizing search results using either a user's topical interests or their desired reading level [6] [2] [8], but not both together. For example, topic predictions based on the Open Directory Project (ODP, dmoz.org) have previously been used as metadata to improve Web search effectiveness. Bennett *et al.* [2] derived an ODP class distribution for a query based on clicks and used this in combination with ODP class distributions computed for Web pages, to obtain new ranking features. Song *et al*. [15] used the entropy of the ODP category distribution as a site characteristic, analyzing the specialization of the search content at that site. Web sites with low entropies were considered to have higher search focus. The topic entropy of a site was a weak feature for predicting accurate site recommendations. White *et al.* [18] also used topic predictions to model users' short-term interests. They computed ODP labels for the top 10 search results, for clicks on the search engine result page (SERP), and for pages visited following SERP clicks. A user's short-term topic profile was the aggregate of these ODP page categories, weighted by dwell time and the length of time between a click and the current query. They found that topic-based short-term profiles helped for ambiguous queries, and that per-query optimal weights gave better improvement than globally optimal weights. Our work is different in that we used the reading level information together with the topic prediction results, and used long-term search session logs to estimate profiles and analyze user behavior.

Previously, Collins-Thompson *et al*. [6] used reading level metadata to perform Web search personalization. They computed reading level features for search result captions and the underlying full pages. Tan *et al.* [16] have also recently explored personalized content selection by modeling text comprehensibility. In contrast, our focus in this study is on general characteristics of pages, websites, users, and their relationships in Web search, as measured in terms of reading level and topic. We do not perform re-ranking or personalization as a task, although some of our results could be applied to obtain new features for personalization or contextual search tasks.

The use of reading level and topic metadata together for Web content retrieval was previously used in an intelligent tutoring application by Collins-Thompson and Callan [4]. About 20 million Web pages were tagged with both ODP topic category and reading level distributions in order to provide personalized material for improving vocabulary acquisition. However, this was a specialized retrieval application, not general Web search, and no analysis was provided on how topic and reading level interacted in pages, sites, or with users.

Previous work has also studied the relationship between a user's *topic familiarity* and search behavior and effectiveness. Kelly and Cool [10] found that with increased familiarity, document reading time decreases while search efficacy improves. Freund, Toms and Waterhouse [9] compared search behavior of work-related Web search to general search and found that work-related sessions used longer queries with a higher proportion of technical terms. The TREC Hard Track in 2003 [1] included a "familiarity" feature of a query, defined as user background knowledge on topic.

In contrast, Kumaran *et al.* [12] defined familiarity as a property of a document independent of user or query. They trained a classifier to label documents as either Introductory or Advanced, using features that included stopwords, reading level estimates, and various page-based features such as the amount of non-anchor text. Unlike their representation of a page, we use the entire vocabulary of the page, not just stopwords, and we explicitly model the interaction of reading level and topic predictions. Moreover, we consider the user's interests and background profile as critical to the nature of content difficulty – not as a static, user-independent property of documents.

Domain expertise on the Web is another closely connected area of research. White, Dumais and Teevan [17] characterized expert *vs.* non-experts according to their Web search behavior, across four domains: Computer Science, Legal, Medicine, and Finance. They also used interaction features from queries and sessions to predict whether a user was a domain expert. The expert and non-expert websites used in their study were labeled by human judges, whereas we learn to identify expert *vs.* non-expert websites automatically for a given domain.

This work extends existing research in the following ways. We introduce a novel probabilistic RLT profile and present a large-scale analysis on the interaction between reading level and topic profile on the web, particularly for content appearing in search results. We also demonstrate the value of building a profile using both reading level and topic, whereas existing work [6] [2] [8] considered them separately in building a profile. Finally, we introduce a technique for classifying expert vs. non-expert content, whereas existing work [17] focused on identifying domain experts.

3. READING LEVEL & TOPIC PROFILES

In this section, we present our methodology for building reading level and topic (RLT)-based profiles for important entities of Web search – users, websites and queries. We first describe how we obtain reading level and topic distributions for the content of individual URLs (Web pages), and how these are aggregated to form distributions that are the RLT profiles for entities such as website, queries, and users. We then define measures of the difference between profiles of individual entities, and the ambiguity or coherence of profiles for *groups* of entities. With these measures, we then explore interesting connections across entity types, or between entities and user search behavior.

3.1 Predicting Reading Level and Topic

We begin by describing how we characterize the content of a URL based on its reading level and topic distributions.

3.1.1 *Predicting Reading Level Distribution of Text*

We represent the reading difficulty of a document or text as a random variable R_d taking values in the range 1-12. These values correspond to American school grade levels, although they could easily be modified for finer or coarser distinctions in level, or for different tasks or populations. We computed reading level predictions for the full body text extracted from the underlying Web pages.

The reading difficulty prediction method that we use is a variant of an existing statistical language modeling approach that has been extensively evaluated on Web content and shown to be effective for both short, noisy texts, and full-page Web texts [4]. Unlike traditional measures that compute a single numeric score, this approach provides extra information about score reliability by computing the likely *distribution* over levels. We show later that having the entire distribution is important for higher-quality predictions. Moreover, language models are vocabulary-centric and can capture fine-grained patterns in individual word behavior across levels. Thus, they are ideal for the noisy, short, fragmented text that occurs on the Web in queries, titles, result snippets, image or table captions. Our variant incorporates feature weight estimation similar to that used in a second, recently introduced reading difficulty prediction model that computes a word's estimated age of word acquisition from a corpus of labeled examples. Evaluated on a test corpus, our classifier has a Root Mean Squared Error of approximately 1.7 grade levels (see [11] for more details).

3.1.2 *Predicting Topic Distributions from Text*

We chose to use the Open Directory Project (ODP) for topic classification because of its broad, general purpose topic coverage and availability of reasonably high-quality training data. For the experiments reported in this study we used 219 topical categories from the top two levels of the ODP.

To train topic classifiers, we used a crawl of ODP from early 2008. We first split the data into a train (70%) and validation (30%) set, then identified the topic categories (some categories like Regional are not topical and were discarded) that had at least 1K documents. This resulted in 219 categories at the top two levels of the hierarchy. To simplify the prediction-phase of the classification model, we simply flattened the two levels to an *m*-of-*N* (where N = 219) prediction task. A logistic regression classifier using an L_2 regularizer was trained over each of the ODP topics identified. When optimized for the F_1 score in each ODP category, the classifier has a micro-average F_1 of 0.60 (see [2] for more details).

3.2 Building Reading Level & Topic Profiles for Websites, Users, and Queries

We use the term *entity* (e) to refer to a website (s), a user (u) or a query (q). We use capital letters (S, U, Q) to denote a group of each entity type. We now define our novel reading-level and topic-based *profile* of an entity as the *probability distributions* of the given entity's reading level (R), and topic (T), or the combination of reading level and topic (RT). For example, a reading level and topic profile of a user or a query can be written as $P(RT|u)$ and $(RT|q)$, respectively.

Figure 1 summarizes typical relationships between entities in a Web search session, where a user issues a query, the query surfaces a website, and a user visits the website. The arrows here are bi-directional because a profile for any entity can be constructed from data describing any of the other entities. For instance, user information from site visits can be used to create a site profile, and site visits can be used to create user profiles.

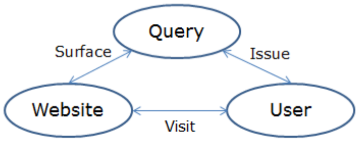

Figure 1: Relationship between entities in a search session.

One motivation for modeling the profile as a distribution, as opposed to a summary statistic, is to capture *focus effects* that allow us to distinguish between, for example, two websites or two users with identical average reading level but very different diversity in the topics or range of difficulty levels they cover.

Table 1 lists the entities used in this paper and potential sources of information for building them. We can use the information associated with the entity itself or the information aggregated from related entities to build a profile of each entity. We next describe how an entity-specific profile can be estimated.

3.2.1 *Profiles based on the Entity Itself*

Here we describe how we can build profiles of important entities based on the entity itself. Given a set of URLs associated with each entity, the joint distribution of reading level and topic is built by aggregating the distributions of the individual URLs computed by URL-level classifiers. To prevent the bias arising from the imbalance in the number of samples used, we used a fixed number of URLs (see below) to estimate the profile of each type.

Table 1: Summary of profile types used in this paper and sources used to build them.

Profile Type		**Built From**	**Source**	**Weighting**	**Section where we use**					
Page		Text classifiers	*Page text/content*	-	**4.2**				**5.1**	**5.2**
User		Aggregate	*Page profiles of visited URLs*	Num. of visits			**4.4**		**5.1**	
Session		Aggregate	*Page profiles of visited URLs in session*	Num. of visits			**4.4.2**			
Query		Aggregate	*Page profiles of top-ranked URLs*	Uniform				**4.5**		
Website	Content view	Aggregate	*Page profiles*	Uniform					**5.1**	**5.2**
	Usage view	Aggregate	*Page profiles, user-viewed*	Num. of visits		**4.3**				**5.2**
	Demographic /visitor view	Aggregate	*User profiles of visitors to site*	Uniform						**5.2**
	Surfacing queries	Aggregate	*Profiles of surfacing queries*	Uniform						**5.2**

We consider a website (or 'site') to be a set of URLs within a full URL domain. Site profiles can be built using one of two sources – site content, or user interaction with the site. In the *content view* of a site, we simply examine a sample of representative content without regard to user interaction with the site. In the *user view* of a site, we focus on the subset of URLs that users actually viewed on the site. This we call a *search-biased* user view – that is, we filter URLs that were clicked in search results.

To estimate the profile of a user, we use the URLs visited by a user during Web search sessions. In this work, we randomly choose 25 URLs to estimate the site-level or user-level profiles (we chose one fixed number to control for the amount of data used to estimate a profile).

For the case of queries, we use the top 10 URLs as of the profile for the query. When past click data is available for the query, one can use the list of clicked URLs with associated frequency, as was done in Bennett *et al.* [2].

3.2.2 Profiles based on Entity Relationships

Profiles can also be constructed from other profiles: a user's profile could be computed not only based on the webpages visited by the user, but alternatively using the profiles of websites visited by the user, or the profiles of queries issued by the user.

This is an important source of information because related entities provide a rich context by which we can make more accurate judgments on each entity. For instance, since a profile of a website is built using many webpages, it can be a more reliable source of information than using individual webpages.

In theory, relationship-based profile construction could be done in a way that introduced circular dependencies – for example, by having user profiles computed from site profiles of visits, which themselves were created from user profiles. We avoid such circularity issues here by using profiles based only on the entity itself, and not based on summaries of other entities, when building a profile based on entity relationships. In general, this type of aggregation is commonly used as an approximate way to infer properties of related entities in relational learning.

3.3 Characterizing and Comparing Profiles

Given the entity-specific profile built as above, now we define the measures used to characterize and compare entities and groups of entities.

3.3.1 Characteristics of an Individual Entity

We first describe the measures we use to characterize an individual entity. Since each profile is a probability distribution, we use the expectation and entropy to summarize the distribution. We denote expectation of reading level for a given entity e as $E[R|e]$, the expectation of topic distribution as $E[T|e]$, and the expectation of the joint distribution as $E[RT|e]$.

As a measure of variation, we use the entropy of reading level, topic and the joint distribution of reading level and topic. If we denote the probability that an entity e has a specific reading level $R_d \epsilon R$ as $P(r|e)$, the reading level entropy of the entity $H(R|e)$ can be derived as follows:

$$H(R|e) = \sum_{R_d \epsilon R} P(R_d|e) \times \log_2 P(R_d|e)$$

We can similarly define topical entropy $H(T|e)$ and joint entropy $H(RT|e)$.

3.3.2 Characteristics of a Group of Entities

In addition to summarizing the characteristics of each entity, we need to represent groups of entities. We build the profile of an entity group by aggregating the distributions of individual entities. Here, we aggregate using what may be considered a weighted centroid of the individual distributions, which for brevity we call a 'group centroid'. For instance, if we represent the probability that we observe a user u in a user group U as $P(u|U)$, we build the reading level profile for U as follows:

$$P(R|U) = \sum_{u \epsilon U} P(u|U) \times P(R|u)$$

In building a site profile based on its visitors, we estimate $P(u|U)$ based on the frequency of a user's visitation over sites. Once we have this aggregated representation, we can use the same metrics as in the case of individual entities.

In addition to using the group centroid to characterize the group profile, we can represent the diversity of the group in terms of its members. Here, we measure the diversity of a group using the average distance of members from the group centroid. In the case of reading level this is:

$$Div_R(U) = \sum_{u \epsilon U} P(u|U) \times D_R(U,u)$$

We use several metrics of comparison between an entity and a group to measure the distance, $D_R(U,u)$, which we explain in detail in the next section.

3.3.3 Comparisons between Entities or Groups

In many applications, we need to compute the profile similarity (or distance) between two entities, or between an entity and an entity group.

For the case of reading level, the simplest metric of comparison is the difference in the expectation of reading level between entities e_1 and e_2 as follows:

$$Diff_R(e_1||e_2) = E[R|e_1] - E[R|e_2]$$

However, it is not clear how we can define such difference metric for the case of topical category or the joint distribution of both. Also, the difference metrics captures only the mean of the distribution.

As an alternative, we use the Kullback-Leibler (KL) Divergence and Jensen-Shannon (JS) Divergence [7] to compare the similarity or distance between the full probability distributions of two entities. For reading level distribution between entities e_1 and e_2 these measures are defined as:

$$KL_R(e_1||e_2) = \sum_{R_d \epsilon R} P(R_d|e_1) \times \log_2 \frac{P(R_d|e_2)}{P(R_d|e_1)}$$

and

$$JS_R(e_1||e_2) = KL_R(e_1||\tfrac{e_1+e_2}{2}) + KL_R(e_1||\tfrac{e_1+e_2}{2}).$$

KL-divergence and JS-divergence for topic and joint distributions are defined similarly. To handle the zero frequency problem in calculating KL divergence, we used absolute discounting with ε = 0.001.

Category	Count	R_1	R_2	R_3	R_4	R_5	R_6	R_7	R_8	R_9	R_{10}	R_{11}	R_{12}	E[R|T]	SD[R|T]
Reference	20,959	0.00	0.00	0.00	0.02	0.17	0.10	0.15	0.04	0.02	0.03	0.20	0.27	8.80	2.86
Health	42,145	0.00	0.00	0.00	0.03	0.18	0.08	0.13	0.04	0.04	0.10	0.27	0.11	8.53	2.65
Science	19,816	0.00	0.00	0.00	0.06	0.23	0.09	0.07	0.02	0.01	0.08	0.27	0.17	8.44	2.97
Computers	93,204	0.00	0.00	0.00	0.06	0.24	0.19	0.03	0.01	0.01	0.02	0.32	0.12	8.11	3.00
Business	113,122	0.00	0.00	0.00	0.05	0.22	0.16	0.09	0.03	0.02	0.04	0.26	0.12	8.08	2.86
Society	232,791	0.00	0.00	0.00	0.02	0.23	0.07	0.35	0.03	0.01	0.01	0.22	0.06	7.62	2.42
Adult	31,044	0.00	0.00	0.00	0.05	0.28	0.26	0.14	0.05	0.02	0.01	0.13	0.06	6.98	2.41
Kids_and_Teens	10,253	0.00	0.00	0.02	0.23	0.26	0.13	0.09	0.02	0.01	0.02	0.15	0.08	6.60	2.81
Games	27,528	0.00	0.00	0.00	0.19	0.36	0.10	0.11	0.02	0.02	0.03	0.12	0.03	6.39	2.44
Recreation	48,619	0.00	0.00	0.00	0.11	0.44	0.19	0.08	0.02	0.02	0.02	0.09	0.02	6.18	2.15
Arts	162,762	0.00	0.00	0.00	0.08	0.40	0.27	0.10	0.05	0.01	0.01	0.06	0.02	6.18	1.94
Home	20,577	0.00	0.00	0.02	0.19	0.41	0.14	0.04	0.03	0.01	0.03	0.09	0.04	6.08	2.40
News	19,370	0.00	0.00	0.00	0.04	0.41	0.33	0.14	0.02	0.02	0.01	0.03	0.01	5.99	1.45
Shopping	109,875	0.00	0.00	0.01	0.22	0.29	0.24	0.09	0.03	0.01	0.02	0.07	0.02	5.98	2.09
Sports	31,942	0.00	0.00	0.00	0.09	0.56	0.11	0.10	0.03	0.03	0.02	0.06	0.02	5.94	1.93

Table 2: Reading level distribution and the average reading level for top ODP categories. Rows are ordered by the expectation of reading level. Cells are shaded according to their probability mass.

4. CHARACTERIZING THE WEB USING READING LEVEL & TOPIC PROFILES

We now show how assigning a RLT profile to websites, users, and queries reveals interesting new relationships and task characterizations that have implications for improving search.

We first provide a description of the datasets used in this study, and an analysis of Web content and websites with respect to topic and reading level. We then examine how we can characterize Web users, websites and queries using profiles.

4.1 Data Set

The primary source of data for this study is a proprietary data set containing the anonymized logs of URLs visited by users who consented to provide interaction data through a widely-distributed browser plug-in. The data set contained browser-based logs with both searching and browsing episodes from which we extract search-related data. Log entries include a browser identifier, a timestamp for each page view, and the URL of the Web page visited. To remove variability caused by geographic and linguistic variation in search behavior, we only include log entries generated in the English-speaking United States locale.

The results described in this paper are based on URL visits during 10 weeks from August through early October 2010, representing millions of Web page visits from thousands of unique users who visited at least 25 pages during the period. From these data we extracted search sessions from Bing, using a session extraction methodology similar to White et al. [18]. Search sessions begin with a query, occur within the same browser and tab instance (to lessen the effect of any multi-tasking that users may perform), and terminate following 30 minutes of user inactivity.

From these search sessions we extracted search queries and for each query, we obtained the top ten search results retrieved by Bing and the titles and the snippets for each result. In total, we built long-term profiles for 7,613 users based on more than 2 million clicks. For websites, we built both content-view and user-view profiles for 4,715 websites which had more than 25 clicked URLs during the period. Finally, we created profiles for 141,325 unique queries in our data set.

A second dataset, which we call the 'web content' dataset, comprised reading level and ODP topic predictions for a snapshot of 8 billion Web documents from April 18, 2011. Each page was tagged with a reading level distribution over American grade levels 1-12, and top-3 most likely ODP categories as described previously.

4.2 Characterizing Web Content

We start by examining the properties of Web content in terms of reading level and topic, using the 2 million URLs visited by Web users as described in the previous section. While this dataset is a search-biased subset of the Web, it represents a significant sample of content that is both broad in topic coverage and of explicit interest to users.

Table 2 summarizes the reading level distribution and the mean and standard deviation of reading levels for the 15 top-level ODP categories. Topics are sorted by descending mean reading difficulty $E[R|T]$. The resulting ordering fits with our general expectation: the categories with highest average reading difficulty have aspects that were more technically oriented, namely Reference, Health, Science, Computers, and Business. The lowest-difficulty categories were more broadly-oriented topics: Sports, Shopping, News, Home, and Arts. In addition, especially for the more technical topics, there is noticeable bi-modality in reading level distribution, with one mode in the grade 4-7 range and a second in the grade 10-12 range, perhaps reflecting the co-existence of 'layperson' and 'expert' content respectively for these topics. The standard deviation of per-topic reading level $SD[R|T]$ also varies considerably across topics. Computers, Science, Reference, and Business have the highest standard deviations in reading level. Kids & Teens also has high variance, perhaps due to its goal of providing content to a broad age range of children. Lowest variance in reading level

Table 3: The correlation between site profile entropy and site visitors' group-level profile diversity, for different entropy and group-level diversity measures.

Website Content Entropy	*Visitor Group-Level Diversity*		
	Div$_R$(U|s)	**Div$_T$(U|s)**	**Div$_{RT}$(U|s)**
E[R|s]	0.052	0.081	0.095
H[R|s]	0.025	0.127	0.143
H[T|s]	0.094	0.336	0.324
H[RT|s]	0.057	0.260	0.264

were News, Sports, Arts, Shopping, and Recreation. The low variation among News pages in particular suggests that news content tends to be more consistent and homogeneous in style than generic Web content.

We next summarize reading level and topic distributions for the three entities of interest – websites, user and queries, and then show how these characterizations can be used to model the selection of search results.

4.3 Characterizing Websites

Here we move from individual page properties to website profiles built by aggregating reading level and topic distribution predictions over many individual pages from the website. Website-level properties help capture the overall scope of content at the site, and they provide a useful background model for dealing with previously unseen pages from the site.

4.3.1 Topic-specific Analysis

We first look at the distribution of websites in terms of the expectation of reading level and the probability that they belong to a specific topic, where we used content-based profiles. Figure 2 shows results for two topics, Kids & Teens (left) and Computers (right). Each point corresponds to a website. The *x*-axis shows the probability $P(T|s)$ of the example topic given the site while the *y*-axis shows the expected reading level over all pages for that site. For the Kids & Teens topic, there is a clear negative correlation between the predicted 'Kids' nature of a site and the average reading level of its pages: the more strongly predicted the Kids & Teens topic is for a site, the lower the site's overall reading level. The Computers topic, on the other hand, exhibits the opposite trend: the more computer-focused a website, the higher its overall reading level is likely to be.

4.3.2 Site Content Profile by User Profiles

In this section we examine the relationship between a site's profile and its visitors' profiles. Table 3 shows the correlation between entropy for site's profile (based on content) and the divergence among site visitors' profiles (based on a sample of site URLs visited). The divergences are based on topic, reading level, and the joint distribution of topic and reading level. The results show that the entropy in a given site's topic content profile is positively correlated with the topic diversity in profiles among its visitors. This means that sites with topically diverse contents attract diverse visitors. However, the correlations are lower in the case of sites' reading level profile.

The overall correlation between reading level and site diversity is low ($corr(E[R|s], Div_{RT}(U|s)) = 0.095$) in the top right cell of Table 3. However, when we analyzed this further by breaking down the results across topics, as shown in Figure 3, we discovered significant variation, with much higher correlation for some topics. This is an excellent example of how the joint analysis of reading level and topic can reveal new insights.

For example, Figure 3 shows that for the Computers and Reference categories, higher reading level of the site led to more coherent visitors. For the Kids and Teens category, however, higher reading level meant more diverse visitors. One possible interpretation of this result is that the range of users is more restricted for technical categories like computers, whereas a site for children may have content that can bring interests from more diverse user groups (e.g., parents and teachers) if it has a high reading level.

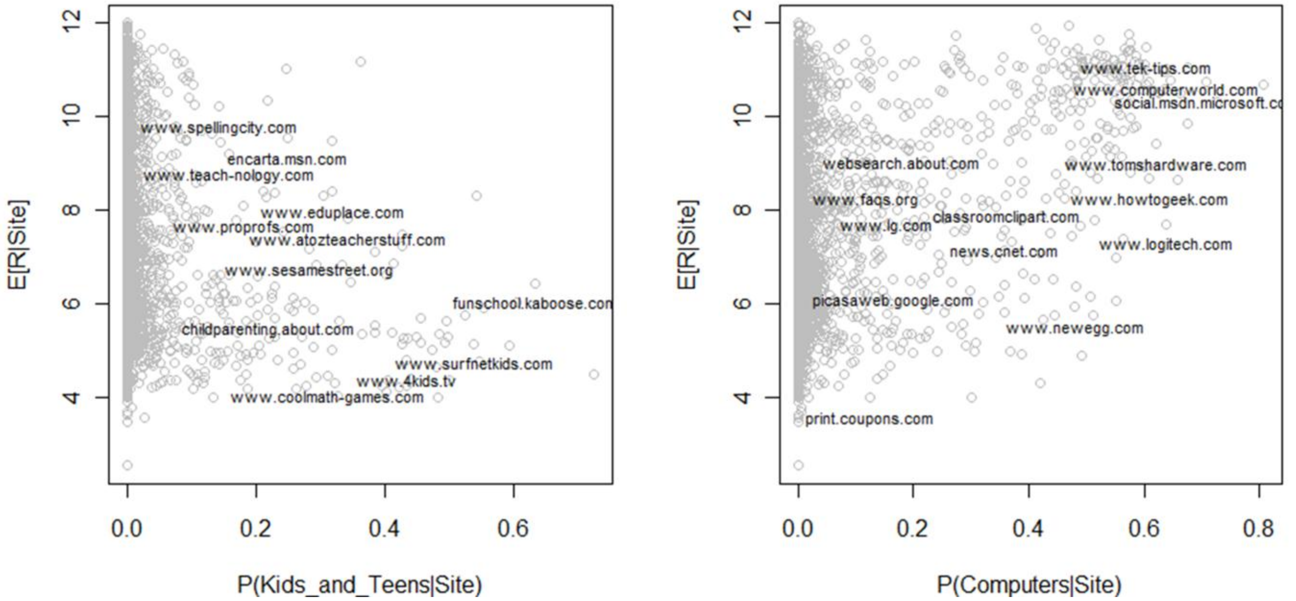

Figure 2: Scatterplot of website content-based profiles for average reading level against the probability of belonging to a topic (left: Kids and Teens, right: Computers). Labels are displayed for a representative subset of websites in this space.

4.4 Characterizing Web Users

We next look at the profile characteristics of Web users, similarly to Section 4.3, in terms of the expectation of reading level and the probability that they belong to a specific topic.

4.4.1 *Topic-specific Analysis*

We investigated the correlation between a user's reading level and the probability of the user belonging to a category, as estimated from their profile over a 10-week period. Figure 4 shows the results for two categories – Kids and Teens (left) and Computers (right). Each point represents a single user profile plotted as a function of dominant topic probability (*x*-axis) and expected reading level (*y*-axis). The results show that for some classes, such as the Computers topic (right), there is a clear separation into readers in terms of user preference for higher- vs. lower-level content, as the probability of a user belonging to that topic increases. Other topics, such as Kids and Teens (left) do not exhibit this behavior.

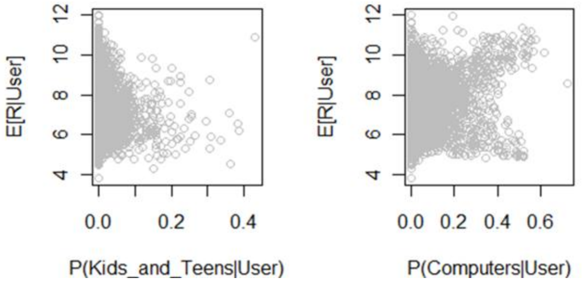

Figure 4: Scatterplot of user profiles plotted by average reading level against the probability of belonging to a topic (left: Kids and Teens, right: Computers).

4.4.2 *Users' Deviation from Their Own Profiles*

Most previous work that has applied reading level or topic categories to search tasks has assumed that users want material matching their profile as closely as possible. In this section we provide a characterization of key tasks for which users *deviate* significantly from their typical profile during a search session. Our hypothesis was that users demonstrate *higher motivation to obtain information* when they exhibit *stretch reading* behavior – spending significant time reading material that involves a higher cognitive effort.

With the introduction of reading level metadata, we now have a way to measure approximate cognitive load of the search results a user sees. If we could identify such situations automatically, the search engine might customize its retrieval or interface to support such high-motivation needs. To our knowledge this is the first such estimate of cognitive load that does not require specialized client hardware such as eye-tracking [3] and thus can be widely used on a very large scale.

To perform our analysis we built a unigram language model for the titles of all 'stretch pages' that were at least four grade levels above the average reading level for a given user's profile, and which had clicks indicating that they were satisfied with the result. Based on models developed by [14], we use dwell times of 30 sec. or more to denote satisfaction. We compared this 'stretch reading' title language model to a background language model based on the titles of all clicked content. To ensure that we identified terms of interest to a broader set of users we restricted the language models to only consider words that were associated with clicks from at least ten unique users. For each word, we then computed a score using the log ratio of its probability in the 'stretch' model, to its probability in the background model. Words with log-ratio score much greater than zero are much more likely in the 'stretch reading' model than the background model.

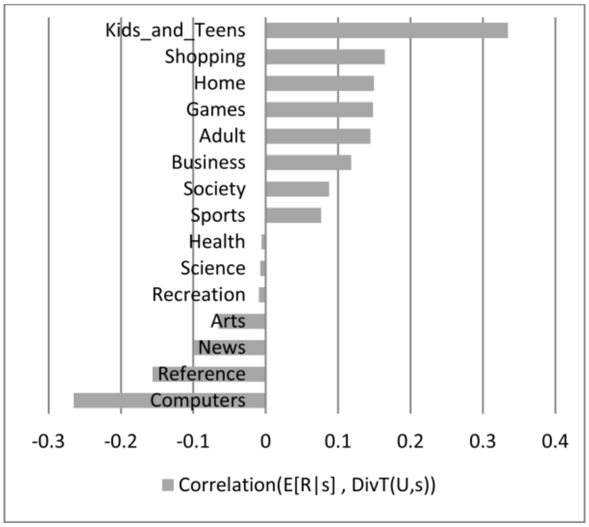

Figure 3: The correlation between a website's expected reading level and the profile diversity of its visitors, varies greatly depending on the topic of the website.

The highest- and lowest-scoring stretch words are shown in Table 4. The top terms are 'test' and 'tests' – and looking at more detailed search context, we found the main topic areas of the corresponding queries to be a combination of education and medical tests. Other education-related terms also appear in the top set, including 'education' and 'learning'.

In general, we found the great majority of stretch reading could be classified into five main areas that correspond to important but sometimes challenging activities required in everyday life: education-related reading such as test preparation and fact-finding about schools; government-related material such as

Table 4: The strongest and weakest words associated with stretch reading, as estimated from titles of pages with a satisfied SERP click, and computed log ratio against a background model of all SAT click titles.

Highest association with stretch reading		Lowest association with stretch reading	
Title word	Log ratio	Title word	Log ratio
tests	2.22	best	-0.42
test	1.99	football	-0.45
sample	1.94	store	-0.46
digital	1.88	great (deals)	-0.47
(tuition) options	1.87	items	-0.52
(financial) aid	1.87	new	-0.53
(medication) effects	1.84	sale	-0.61
education	1.77	games	-0.65
forms	1.76	sports	-0.78
plan	1.74	food	-0.81
pay	1.71	news	-0.82
medical	1.69	music	-1.02
learning	1.62	all	-1.35

Table 5: The effectiveness of different user profile-based metrics for predicting users' preference between clicked and skipped items in search results, as measured by proportion of clicked items with a correct preference prediction. Results are subdivided by degree of user profile focus based on profile distribution entropy. Prediction generally improved for users with more focused profiles, and when reading level and topic were used together.

User data		*Page-based measures*			*Site-based measures*		
User Group	*# Clicks*	*$Diff_R(u,d)$*	*$KL_R(u,d)$*	*$KL_T(u,d)$*	*$KL_R(u,s)$*	*$KL_T(u,s)$*	*$KL_{RT}(u,s)$*
↑Focused Profile	5,960	55.03%	60.15%	69.16%	59.23%	60.79%	65.27%
	147,195	51.08%	50.43%	55.78%	52.25%	54.20%	54.41%
	197,733	49.74%	50.84%	54.20%	52.75%	53.36%	53.63%
↓Diverse Profile	15,610	49.05%	50.90%	54.54%	53.81%	53.40%	52.90%
Total	366,498	50.33%	50.83%	55.10%	52.70%	53.82%	54.10%

instructions for filling out forms, or legal/judicial information; medical content on the effects of prescriptions, diseases, and tests and procedures; financial aid and planning materials; and job-seeking. All of these scenarios are ones where it is more likely the user is highly motivated to obtain information and thus more willing to spend time outside of their typical topic and reading-level zone.

The least-likely stretch terms gives indications of very different topical content that match with our other findings, that topics such as sports, games, shopping have lower average reading level, while also being popular. The predominance of these topics could suggest that people either don't read above their reading level in these areas, or that there don't typically exist pages with high reading levels that cover those topics. This is the subject of future work.

Being able to automatically identify tasks where the user is highly motivated to learn or require help has important implications for search engines. It means that, on detecting such scenarios, a search engine could automatically suggest easier website or page alternatives if available, or provide support in other ways, such as vocabulary assistance, specialized vertical search, or enhanced suggestions through a social network of friends or people with related goals.

4.5 Characterizing Web Queries

In Section 3 we introduced the definition of a 'query profile' based on the combined reading level and topic distribution of the top-10 results returned for the query. Similar to the cases of websites and users, we investigated the correlation between a query's reading level and the probability of the query belonging to a topic category. Figure 5 shows results for two categories: Kids and Teens (left) and Computers (right). Each point represents a single user profile plotted as a function of dominant topic probability (*x*-axis) and expected reading level (*y*-axis).

The results show similar trends to those for websites: queries with higher probability for the Kids and Teens category had lower reading levels, and those with higher chance of belonging to the Computers category had higher reading levels. Again, this result shows that there exist significant correlations between reading level and topic probability within an entity profile.

We also compared the topical coherence of query profiles, as measured by the profile's topic entropy $H[T]$ over a query's results, between navigational and non-navigational queries. We found that on average, navigational queries had profiles with higher topic entropy (1.460) than non-navigational queries (1.397). This may be due to navigational queries having many minority intents after the first or second result, or results covering varied aspects of the navigational entity after the direct links. This is an area for further analysis.

5. APPLICATIONS

In this section we describe two initial applications of RLT profiles to search-related problems: using profile-based metrics to predict user preferences on Web search results, and classifying expert vs. novice content on the Web.

5.1 Analyzing the Impact of Profile Match on Search Results Clicks

We analyzed how users' preferences on search results can be predicted using their profiles. Specifically, we compared the profile divergence between a user and each clicked document in the top 10 search results, and between the user and each skipped document right above the clicked document.

We used the same 10-week session data for the experiments described in this section, with two important differences. First, since we are interested only in search behavior, we used the users' clicks only on the top 10 search results. Also, to estimate user profiles in a data set different from the search log data we are analyzing, we split the data set by time period. We used the user profile estimated from the first five-week period for analyzing the search behavior of the second five-week period.

Table 5 summarizes the results. Each cell represents the percentage of cases where clicked documents have a higher profile matching score than the corresponding skipped document. Each column represents different metrics of divergence between a user and a document (or a corresponding

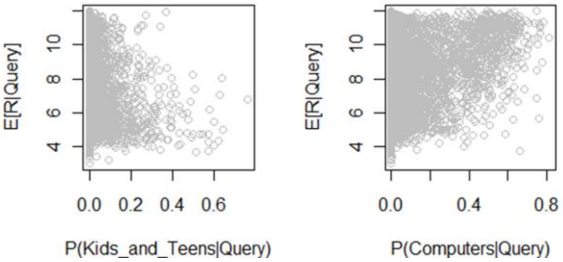

Figure 5: Scatterplot of query profiles plotted by average reading level against the probability of belonging to a topic (left: Kids &Teens, right: Computers).

Table 6: Summary of accuracy, precision and recall for expert site-finding classification task compared to the baseline accuracy of random selection.

Baseline (predict most likely class)	0.659
Classifier accuracy	0.822
Precision for expert sites	0.769
Recall for expert sites	0.683
Precision for non-expert sites	0.847
Recall for non-expert sites	0.893

website). We broke down our results by the degree of 'focus' in a user's profile, as measured by the entropy of the joint reading level and topic distribution $H[RT|u]$. For instance, the data in the top-most ('Focused') row of Table 5 are from users with relatively low profile entropy and thus with very focused search behavior in terms of both reading level and topic. Users in the 'Diverse' group, on the other hand, had broader reading level and topic profiles with relatively high entropy. First, we note that in general, all KL-divergence-based measures show accuracy higher than 50% overall, indicating that the profile does give some signal about users' click preferences.

Sorted by prediction effectiveness on the total user group, each metric's gain was significantly higher than the other metrics below it, as well as the random predictions, using a paired t-test with p-value < 0.01.

Our main observations are that preference prediction was stronger for users with more focused profiles across all metrics, and that for site-level profiles, the joint distribution $KL_{RT}(u,s)$ gives better results than using either reading level or topic distribution alone. Furthermore, site-based measures, which use the aggregated topic and reading level distribution of a page's website, performed almost as well as the measures that used the individual page, while requiring orders of magnitude less storage. The page-level KL divergence in topic between a user and a document $KL_T(u,d)$ did indeed perform better than the divergence in joint distribution between a user and a site $KL_{RT}(u,s)$. This is understandable, in that URL-level information is more fine-grained than site-level information.

Finally, within page-based measures, the feature based on the entire distribution $KL_R(u,d)$ showed better performance than using only the expectation of the same distribution $Diff_R$ (u, d), illustrating the benefit of having more information by modeling the profile as a distribution.

5.2 Predicting Domain Expertise Using Reading Level & Topic Profiles

In this section we apply RLT profiles to data from a previous study on domain expertise. We then describe and evaluate a classification model for distinguishing expert from non-expert websites using features of reading level and topic profiles.

5.2.1 Comparing Expert vs. Non-expert URLs

White *et al.* [17] used manually annotated expert vs. novice URLs to characterize the behavior of domain experts and non-experts, comprising the 150 most popular URLs within each of four domains: Medical, Computer Science, Legal, and Finance. We computed reading level and topic distributions for each URL in their data set. With these distributions, for each URL we computed expected reading level and the joint entropy of reading level and topic, $H(RT)$. Figure 6 illustrates this, showing that pages labeled as 'expert' have higher reading level and higher topic focus (lower entropy) than the 'non-expert' sites. The differences in both variables varied across domains, with medical domain pages showing highest difference in reading level between expert and non-expert Web pages.

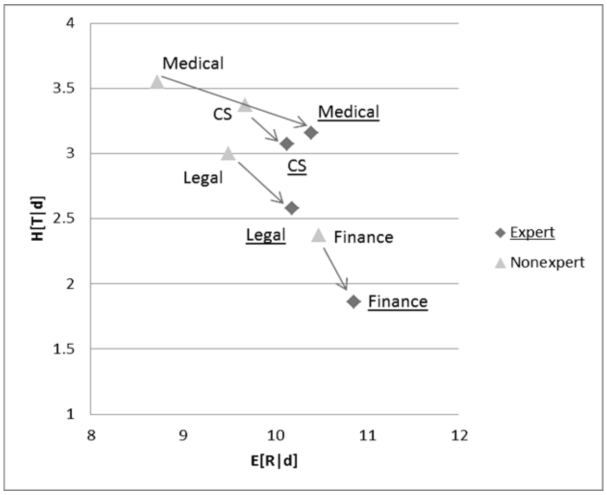

Figure 6: Expert-oriented pages had higher average reading difficulty and higher average topic focus (lower topic entropy) than non-expert pages in the same domain.

5.2.2 Predicting Expert Websites

Based on the observations in the previous section we developed a method to predict expert websites using reading level and topic profiles. We predict expertise at the site level rather than the

Feature	*Correl.*	*Description*
KL_{RT}(s,U)	-0.56	Divergence of average user's RLT profile from site's profile.
KL_{T}(s,U)	-0.55	Divergence of average user's topic profile from site's profile.
Div_{RT}(U|s)	-0.45	Average divergence of user's RLT profile from its centroid.
Div_{T}(U|s)	-0.41	Average divergence of user's topic profile from its centroid.
$Diff_{T}$(s,U)	-0.40	% of users whose top ODP category differs from that of site.
Div_{R}(U|s)	-0.29	Divergence of average user's reading level profile from site's profile.
$Diff_{T}$(s,Q)	-0.28	% of queries whose top ODP category differs from that of site.
H_{RT}(U|s)	-0.21	Average RLT entropy of site visitors.
H(T|s)	-0.13	Site topical entropy (user-viewed URLs).
E(R|s)	+0.23	Site reading level (user-viewed URLs).
E(R|Q)	+0.34	Expectation of surfacing queries' reading level.
E(R|U)	+0.44	Expectation of visitor's reading level.

Table 7: Features used for expert site prediction task, sorted by their correlation with expert/non-expert label.

URL level because sites tend to have a coherent profile in many cases, and much more data can be aggregated at the site level. We focused on predicting expertise of sites within the Computers topic. We selected websites occurring 25 or more times in our session log.

To create a site profile we used both 100 randomly chosen URLs on the site and 25 user-viewed URLs on the site. In other words, we extracted features from both content-based and user-viewed profiles of given website. We also used the aggregate profiles of site visitors and queries used to visit the website. To capture the diversity of visitors' profiles from the site profile, we used the divergence of the average user's profile from the site profile, i.e. how different site visitors were from the site profile itself, as well as the average divergence of a user's profile from the user profile group centroid, i.e. how different site visitors are among themselves.

Our prediction experiment is based on 10-fold cross validation, using a gradient boosted decision tree classifier [19]. The results in Table 6 show 82% overall accuracy. This accuracy is significantly better than the baseline accuracy of 65.9% across all 10 folds. Table 7 lists the classifier features and their correlation with the binary Expert/Non-expert label, showing that features based on profiles of visitors and queries provide strong signals for the 'expertness' of websites. The divergence of visitor's profiles shows highest negative correlation, indicating that users for expert sites have coherent profiles with the site and with each other. Also, the expectation of reading level for queries and visitors shows stronger positive correlation (+0.34, +0.44) with the 'expertness' of a site than the expected reading level of the site itself (+0.23) based on visited URLs.

6. SUMMARY AND CONCLUSIONS

We introduced a novel form of probabilistic profile, the RLT profile, which can be used to describe major entities of Web search such as users, queries, or websites – based on reading level and topic metadata produced by automatic text classifiers. Based on these profiles, we performed a large-scale analysis of Web content and search interactions. Our findings show that RLT profiles are effective for a variety of analysis and prediction purposes: they provide novel characterizations for websites, users and queries by combining distributional statistics of both topic and reading level distributions as well as their joint distribution. These representations can be used for a variety of search-related tasks such as understanding search result preferences, and predicting whether the content of a URL or site is targeted at domain experts or non-experts.

Our main finding is that reading level and topic metadata used together were more effective than either one used alone. We analyzed how reading level distribution of content on the Web varies across topics. Then, using features derived from RLT profiles, we found these features provided effective personalization signals, predicting a user's preference for Web pages and sites in search results. With these RLT profiles we characterized a user's behavior when they deviated from their profile to perform 'stretch' tasks. Finally, we applied RLT profiles of Web sites to analyze and predict the 'expertness' of these sites.

Future directions include applying these findings to various end-user tasks. The divergence metrics developed in this paper could be evaluated for their effectiveness as features for personalized re-ranking. We also plan to investigate users' stretch behavior further, so that we can detect and assist these patterns of behavior more effectively. Finally, the techniques developed for expert *vs*. novice site classification can be applied both for recommendation and ranking purposes.

Acknowledgments

We thank Ryen White for providing and assisting with the expertise URL dataset.

REFERENCES

[1] J. Allan. HARD Track Overview in TREC 2003: High accuracy retrieval from documents. In *Proceedings of TREC 2003*. NIST Special Publication.

[2] P. Bennett, K. Svore, and S.T. Dumais. Classification-enhanced ranking. In *Proceedings of WWW 2010*.

[3] M. J. Cole, J. Gwizdka, C. Liu, R. Bierig, N. Belkin, X. Zhang. Task and user effects on reading patterns in information search. *Interacting with Computers* (2011).

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

[5] K. Collins-Thompson and J. Callan. Information retrieval for language tutoring: an overview of the REAP project. In *Proceedings of SIGIR 2004*. ACM, New York, USA.

[6] K. Collins-Thompson, P.N. Bennett, R.W. White, S. de la Chica, D. Sontag. Personalizing web search results by reading level. In *Proceedings of CIKM 2011*. ACM, New York, USA.

[7] I. Dagan, L. Lee, and F. Pereira. 1997. Similarity-based methods for word sense disambiguation. In *Proceedings of EACL 1997*. Association for Computational Linguistics, Stroudsburg, USA.

[8] Z. Dou, R.Song, J. R. Wen. A large-scale evaluation and analysis of personalized search strategies. In *Proceedings of WWW 2007*.

[9] Freund, L., Toms, E.G., & Waterhouse, J. Modeling the information behaviour of software engineers using a work-task framework. In *Proceedings of ASIST 2005*, Charlotte, NC, USA.

[10] D. Kelly and C. Cool. The effects of topic familiarity on information search and use behaviors. In *Proceedings of JCDL 2002*.

[11] P. Kidwell, G. Lebanon, K. Collins-Thompson. Statistical estimation of word acquisition with application to readability prediction. In *Proceedings of EMNLP 2009*, Singapore.

[12] G. Kumaran, R. Jones. Biasing web search results for topic familiarity. In *Proceedings of CIKM 2005*. ACM New York, USA.

[13] F. Peng, N. Ahmed, X. Li, Y. Lu. Context sensitive stemming for web search. In *Proceedings of SIGIR 2007*. ACM New York, USA.

[14] K. Wang, T. Walker, Z. Zheng. Estimating relevance ranking quality from web search clickthrough data. In *Proceedings of SIGKDD 2009*, 1355—1364.

[15] Y. Song, N. Nguyen, L. He, S. Imig, and R. Rounthwaite. 2011. Searchable web sites recommendation. In *Proceedings of WSDM 2011*.

[16] C. Tan, E. Gabrilovich, and B. Pang. To each his own: personalized content selection based on text comprehensibility. In *Proceedings of WSDM 2012*. ACM, New York, USA.

[17] R. W. White, S. Dumais, J. Teevan. Characterizing the influence of domain expertise on Web search behavior. In *Proceedings of WSDM 2009*. ACM, New York, USA.

[18] R. W. White, P. N. Bennett, S. Dumais. Predicting short-term interests using activity-based search context. In *Proceedings of CIKM 2010*. ACM, New York, USA.

[19] Q. Wu, C.J.C. Burges, K. Svore, and J. Gao. Adapting boosting for information retrieval measures. *Information Retrieval*, 2009.