Predicting User Interests from Contextual Information

Ryen W. White
Microsoft Research
Redmond, WA 98052
ryenw@microsoft.com

Peter Bailey
Microsoft
Redmond, WA 98052
pbailey@microsoft.com

Liwei Chen
Microsoft
Redmond, WA 98052
liweich@microsoft.com

ABSTRACT

Search and recommendation systems must include contextual information to effectively model users' interests. In this paper, we present a systematic study of the effectiveness of five variant sources of contextual information for user interest modeling. Post-query navigation and general browsing behaviors far outweigh direct search engine interaction as an information-gathering activity. Therefore we conducted this study with a focus on Website recommendations rather than search results. The five contextual information sources used are: social, historic, task, collection, and user interaction. We evaluate the utility of these sources, and overlaps between them, based on how effectively they predict users' future interests. Our findings demonstrate that the sources perform differently depending on the duration of the time window used for future prediction, and that context overlap outperforms any isolated source. Designers of Website suggestion systems can use our findings to provide improved support for post-query navigation and general browsing behaviors.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval – *search process, information filtering.*

General Terms

Algorithms, Experimentation, Human Factors.

Keywords

Context, user interest modeling, Website recommendation.

1. INTRODUCTION

Modeling user interests to meet individual user needs is an important challenge for personalization and information filtering applications, such as recommender systems [2]. Information behavior is embedded within an external *context* that motivates the problem situation and influences interaction behavior [12]. Meeting user requirements involves a thorough understanding of their interests expressed explicitly through search engine queries or implicitly through browsing behavior *and* search context.

The information retrieval (IR) community has theorized about context [12], developed models for context-sensitive search (*e.g.*, [27,30]), and performed user studies investigating the role of context in the information-seeking process (*e.g.*, [16]). Large-scale IR systems such as Web search engines assume queries are context-independent. This abstraction is necessary given the scale constraints under which these systems operate. User modeling systems have fewer constraints and typically process past user consumption data, search-related interactions, or explicit ratings to obtain a representation of user interests stored in a *user interest model* (*e.g.*, [10,37]). Such models are suitable for predicting future behavior, augmenting search engine queries, or suggesting relevant items during post-query navigation or general browsing.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
SIGIR'09, July 19–23, 2009, Boston, Massachusetts, USA.
Copyright 2009 ACM 978-1-60558-483-6/09/07...$5.00.

The historical information employed in user interest modeling is one source of contextual evidence about the current session. Others include time of day, user gender, age, ethnicity, locality, etc. The *polyrepresentation principle* [11] suggests that the overlap between numerous contexts associated with the current session can be used to locate pertinent items. The querying and result examination behavior of search system users supports the development of rudimentary user interest models that are based solely on the interaction context (*e.g.*, [36]). These interest models can be effective for identifying aspects of user information needs; however, users spend more time engaged in post-query navigation and general browsing than using search engines [34]. Although context information has been used to support post-query navigation and general browsing (*e.g.*, attentive systems can offer Website suggestions [4,21]), little is known about the value of different contextual sources for this purpose.

In this paper we describe a systematic, log-based study of numerous contextual sources for modeling user interests during Web interaction. The core task for any user modeling system is predicting future behavior, and we evaluate the informativeness of different sources of contextual evidence based on their informativeness for predicting users' future interests at different temporal durations. We assume that the user has browsed to a Web page and the task is to leverage context to predict their future interests. The use of the current page and five distinct sources of context are evaluated: (i) *interaction*: recent interaction behavior preceding the current page; (ii) *collection*: pages with hyperlinks to the current page; (iii) *task*: pages related to the current page by sharing the same search engine queries; (iv) *historic*: the long-term interests for the current user, and; (v) *social*: the combined interests of other users that also visit the current page. This is the first study to systematically assess contextual variants for user interest modeling. We also study the use of overlap between sources as a stronger source of contextual signal. As we will show, the performance of contextual variants depends on the time duration used to represent future interests, and overlap between contexts yields more effective interest models than any model itself. Understanding which sources and source combinations best predict future user interests is critical for the development of effective Website recommendation systems.

The remainder of this paper is structured as follows. Section 2 presents related work on at least contextual IR, user modeling, and recommendation systems. Section 3 describes the log data used to perform our study. The user interest models developed based on each contextual source are described in Section 4. We describe their evaluation in Section 5, and present the findings in Section 6. We discuss our findings in Section 7, and conclude in Section 8.

2. RELATED WORK

This work explores issues at the intersection of contextual IR, user studies based on Web browser or search engine interaction logs, data mining, implicit feedback, user modeling, collaborative filtering, and personalization. Each area has its own wealth of published work; this review focuses on relevant aspects.

Traditional IR models regard the retrieval problem as matching a query with a set of documents [28], and are inadequate for modeling personalized and contextual search. Previous work [27, 30] has used statistical modeling for context-sensitive search, but rely on a single source of contextual evidence. The *principle of polyrepresentation* [11,12] is based on a cognitive approach to IR and signifies that overlaps between a variety of *contexts* associated with the interactive IR process can be exploited to reduce the uncertainty and thereby improve IR performance. The small number of polyrepresentation studies to date have focused on improving retrieval within small, well-defined test collections by eliciting multiple information need representations from users [16] or mining inter-document references and intra-document structure [19,29]. In contrast, we apply polyrepresentation to tackle the challenge of user interest modeling during Web interaction. Although our study is aimed at providing better Web page recommendations for users engaged in browsing activity, the findings could also potentially improve the design of context-sensitive search applications.

Supporting information-gathering behavior beyond search engine interaction has been actively studied. Recommender systems such as *Letizia* [21] and *Watson* [4] suggest items to users based on inferences made about user interests gleaned from their task environment ($e.g.$, recently-viewed Web pages or the contents of active desktop applications). *StumbleUpon* (stumbleupon.com) is a recommender system that uses collaborative filtering (CF) (an automated process combining human opinions with machine learning of personal preference) to create virtual communities of like-minded Web surfers. Rating Web sites updates a personal profile (a blog-style record of rated sites) and generates peer networks of Web surfers linked by common interest. These social networks coordinate the distribution of Web content, so that users "stumble upon" pages explicitly recommended by friends and peers. However, recommendations from CF systems typically require explicit action from a large community of users [9].

Interaction log analysis has provided researchers with insight into user behavior. One element of context that influences user behavior is the type of the information-seeking task. Various taxonomies of these types have been suggested (for both task nature and task complexity), including [3,5,15]. The nature of the information seeking task can lead to differences in user behavior. Kellar and colleagues' study [15] also examined the differences in dwell time for different tasks. Terai and colleagues investigated these differences in a user study exploring informational and transactional tasks on the Web [32]. They found significant differences both in the number of individual pages read and the time taken in reading them between these two task types. Kim and Allen studied both task differences and users' cognitive differences [18], Thatcher explored relationships between different tasks and the search strategies employed by people of differing degrees of Web experience [33], and White and colleagues [35] studied differences in the search behaviors of domain experts and non-experts. These investigations typically involved examining user behavior through query log analysis and user studies.

Without explicit user relevance judgments, user preference can only be inferred from their activities ($e.g.$, clicking on a hyperlink, viewing/saving/bookmarking a page). A range of applications for this concept have been explored, falling under the category of implicit feedback. Recent studies include those reported in [1,14,27] and have shown to effectively improve retrieval performance across a range of scenarios, especially Web search. Examining the applicability of implicit feedback for recommender systems has also been studied [7,22]. Applications of implicit feedback to Web page recommender systems are also available [10,17]. These systems typically establish historical click trails of a user or a community of users, and assess the accuracy of statistical machine learning models which predict future page visits.

A natural application of implicit feedback is in personalized search engines, which incorporate an individual's historical activities as part of a ranking system. There are many challenges with such personalization. For example, when exploiting short-term search history one must detect session boundaries first, so that only those searches with the same information need are used. Unfortunately, most existing studies on long-term search context fail to address this problem, although they still get positive results; studies often use all available context as a whole (or divide it into chunks by time), without distinguishing between relevant and irrelevant parts. Such work includes [23], which interpolates the current query with different chunks (time periods) of history (browsed Web pages) for personalized search, and [25,31], which construct user profiles from indexed desktop documents for search result re-ranking.

Modeling user interests is common practice for the construction of recommendation engines at e-commerce sites such as Amazon and Netflix. These can be derived both from explicit actions by users ($e.g.$, buying a product or requesting a movie) or interaction log behavior (clicking on certain categories of product or movie). In the Web search arena, user models constructed from interaction logs have been used to create automated Web search engine evaluation facilities by Dupret and colleagues [8]. The work most similar to ours is that by Piwowarski and Zaragoza [24] in which they explore three different predictive click models based on what we term historical and social context, but in a Web search setting trying to predict relationships between queries and clicked documents. In that work, they built a probabilistic user-centric model, a group model, and a global model, and a model that combined all three. The best of their models was able to achieve either accurate prediction (50% of the clicks) with high recall (75% of the time), or low recall (5% of the time) but very high accuracy of 98% prediction correctness.

We now describe the primary data source for our log-based study.

3. LOG DATA

The primary source of data for this study was the anonymized logs of URLs visited by users who opted in to provide data through a widely-distributed browser toolbar. These log entries include a unique identifier for the user, a timestamp for each page view, a unique browser window identifier (to resolve ambiguities in determining which browser a page was viewed), and the URL of the Web page visited. Intranet and secure (https) URL visits were excluded at the source. In order to remove variability caused by geographic and linguistic variation in search behavior, we only include entries generated in the English speaking United States locale. The results described in this paper are based on a sample of URL visits during a four-month period from August 2008

through November 2008, representing billions of URL visits from 250,000 unique users. The user sample was selected at random from a larger set of five million users after we had pre-filtered the data to remove extremely-active outlier users (top 1%), all of whom viewed many thousands of pages per day and were likely automated traffic. For each user we required an adequate number of Web page visits to create their *historic context* ($i.e.$, a model of long-term interests). Therefore, in addition to removing outliers, we also only chose users who visited at least 100 Web pages in the time period from August 2008 through September 2008.

From these logs we extracted hundreds of millions of *browse trails*, as defined by [34]. Browse trails consist of a temporally-ordered sequence of URLs comprising all pages visited by a user per Web browser instance or browser tab. Trails terminate with either: (i) a period of user inactivity of 30 or more minutes, or (ii) the termination of the browser instance or tab. The 30-minute threshold has already been used to demarcate sessions in other Web log analyses ($e.g.$, [34]). Access to browse trails let us study users' post-query navigation and general browsing behaviors.

We extracted millions of *context trails* from the set of browse trails that allowed us to study real-user interests. Context trails exist within browse trails and comprise a terminal URL, u_t, and the list of five Web pages preceding u_t in the browse trail, $u_{t-5}, \dots, u_{t-1}$. The five pages preceding u_t forms the immediate session-based *interaction context* introduced in Section 1. Five pages gave us sufficient information about user interests for the perceived situation prior to u_t and a low likelihood of being affected by significant shifts in those interests. Around five million terminal URLs were obtained by randomly sampling by frequency the URLs across the historic browse trails in August and September 2008 (referred to hereafter as T_h) ($i.e.$, each URL had a chance of being selected proportional to its frequency). The set of all terminal URLs, $\{u_t\}$, are the starting points from which we derive contextual information for u_t from the five contextual sources we study.

In the next section we describe the user interest models created based on the context trails and their surrounding contexts.

4. USER INTEREST MODELS

We developed user interest models based on u_t and the five sources of contextual information used in our study. The sources were chosen based on elements of a nested model of context stratification proposed by Ingwersen and Järvelin [12]. The dimensions of that model represent the main contextual influences affecting users engaged in information behavior: (i) *object structures*: signs ($i.e.$, discrete units of meaning), page features (u_t), and cognitive structures (user); (ii) *inter-object contexts or structures*: between-object relations such as hyperlinks or citations; (iii) *interaction context*: evidence of interaction behavior during the search session; (iv) *social, systemic, domain-work task context*: peer group (*social context*), retrieval system (systemic), real work or daily-life tasks (*task context*); (v) *economic techno-, physical-, and societal context*: prevailing infrastructures that influence all elements in the nested model of context, and; (vi) *historic context*: the experiences of the cognitive actor (user) that affect how they perceive and interpret situations. The context stratification is illustrated in Figure 1, with the user at a given Web page, u_t, at the core of the model, and with the dimensions used in our study underlined and shown in boldface. The dimensions not chosen ($e.g.$, intra-object structures, signs, and emotions) could not accurately be modeled in a log-based study since we lacked access to Web page content (only their URLs), the user's cognitive and affective state at session time, or infra-structure details.

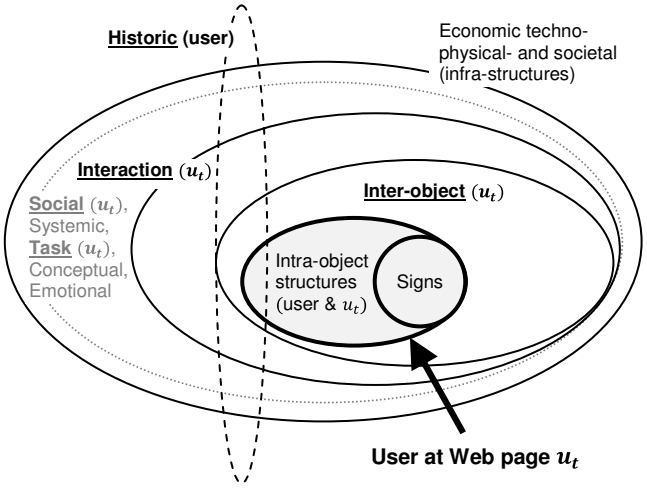

Figure 1. The nested model of context stratification for information seeking and retrieval (based on [12]).

For each context trail extracted from the logs, we created a user interest model for u_t, the *interaction context* $u_{t-5}, \dots, u_{t-1}$, and the other contextual variants (collection, historic, task, and social). To define user interests in a manageable way for all models, we classified the Web pages sourced from each context into the topical hierarchy from a popular Web directory, the Open Directory Project (ODP) (dmoz.org). Given the large number of pages involved, we used automatic classification. Our classifier assigned labels to pages based on the ODP in a similar way to Shen and colleagues [26], by starting with URLs present in the ODP and incrementally pruning non-present URLs until a match was found or miss declared. In a similar way to [26], we excluded Web pages labeled with the "$World$" and "$Regional$" top-level ODP categories, since these categories are location-based and are typically uninformative for constructing models of user interests.

User interests were represented as a list of ODP category labels assigned to URLs from each source. The ODP labels in the list were ranked in descending order based on each label's frequency in the context. For example, the top portion of a user interest model for a British golf enthusiast might resemble:

ODP Category Labels	**Freq.**
$Sports/Golf/Courses/Europe/United\ Kingdom$	102
$Sports/Golf/Driving\ Ranges$	86
$Sports/Golf/Instruction/Golf\ Schools$	63
$Games/Video\ Games/Sports/Golf$	55

The following interest models are created for each contextual variant using this approach:

No context (u_t only): One ODP label is assigned to the terminal URL based on the output of the ODP classifier. This label serves as the interest model for the terminal page in the context trail.

Interaction context ($u_{t-5}, \dots, u_{t-1}$): One ODP label is assigned to each of the five pages immediately preceding u_t in the context trail. The labels are aggregated and label frequencies (based on the number of pages in the interaction context with each label) are used to create a ranked list of labels. The ranked list is the interest model for the *interaction context* of u_t.

Task context: The interest model for the *task context* is created using ODP labels assigned to Web pages visited by other users attempting the same or similar tasks. To realize this goal we used the query and search result page click-through logs from a large commercial Web search engine. One month of logs from October 2008 was used to create a graph from each u_t to each query, q_{u_t}, with a result-page click on u_t. For each u_t, we traversed the graph to set of related URLs, $\{u_r\}$, via the set of queries, $\{q_{u_t,u_r}\}$, that led to a click on both u_t and u_r. Figure 2 illustrates this process.

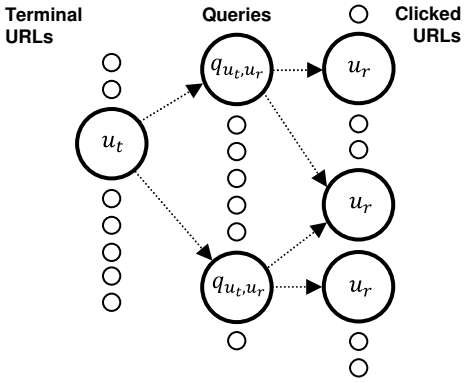

Figure 2. Creating *task context* using queries and result clicks.

An ODP label is assigned to each of the related URLs discovered by traversing this graph. The assigned labels are aggregated and their frequencies used to create a ranked list of labels. The ranked list represents the interest model for the *task context* of u_t.

Collection context: The interest model for the *collection context* was created using Web pages containing hyperlinks that refer to u_t. We obtained the set of in-links for each u_t from the index of a large commercial Web search engine. An ODP label was assigned to each in-link, and in a similar way to other contexts, we created a ranked list of the labels based on their frequency. This list formed the interest model for the *collection context* of u_t.

Historic context: The interest model for the *historic context* was created for each user based on their long-term interaction history. To create each user's *historic context*, we classified all Web pages they visited in T_h, and created a ranked list of ODP labels based on label frequency. This list represents the interest model for the *historic context* for all u_t visited by that user.

Social context: The interest model for *social context* was created by combining the *historic contexts* of users that also visit u_t. Note that this differs from the task context in that we focus on other users' long-term interests rather than only leveraging common querying behavior to find related URLs. From the browse trails in T_h we found users who have also visited u_t, and combined their interest models (*historic contexts*) to create a ranked list of ODP labels based on label frequency. This list formed the interest model for the *social context* of u_t.

Implementing these context variants allowed us to systematically evaluate the effectiveness of different sources of context for user interest modeling. We evaluated each source and combinations of sources based on their predictive value.

5. EVALUATION

In this section we describe the evaluation of our interest models. We describe the preparation of the data for our study, the study methodology, and the measures used to compare the models.

5.1 Data Preparation

We divided the set of browse trails described in Section 3 into two subsets: *historic* (T_h) and *current* (T_c). T_h was defined earlier as the source of the *historic context* for each user. From the October 2008 browse trails, T_c, a set of "unseen" context trails, T_x, was extracted. For each context trail, we constructed interest models for u_t and obtained ground truth data about future user interests. We used browse trails for the current user taken from October 2008 and, if needed, November 2008, *that began after the visit to* u_t as a source of future user behavior. ODP labels were assigned to pages in the future, aggregated by label, and a ranked list of ODP labels was created based on label frequency in the same way as with the interest models described in Section 4. The futures were specific to each user and each u_t, and were used to gauge the predictive value of each of contextual source for the context trail.

Interest model effectiveness may vary depending on temporal distance from u_t to some future time point. We made predictions using three temporal durations: (i) *short*: within one hour of u_t; (ii) *medium*: within one day of u_t, or; (iii) *long*: within one week of u_t. The futures are overlapping: *medium* contains *short* and *long* contains both *short* and *medium*. We could have extended future beyond one week, but felt it would be unreasonable to expect a model to accurately predict longer-term future interests based on a single u_t.

To help ensure experimental integrity, we did not use all context trails; we filtered the trails based on the following criteria:

- The coverage of our ODP classifier with URL back-off was around 65%. A missing ODP label for u_t meant that we could not create the *no context* baseline. Any number of missing ODP labels for $\{u_{t-5}, \dots, u_{t-1}\}$ significantly skews the user interest model for the *interaction context*. Therefore, it was a requirement that for each context trail ODP labels could be assigned to u_t and all Web pages in $\{u_{t-5}, \dots, u_{t-1}\}$.
- The ODP category label for search engines, portals, or social networking sites ($e.g.$, google.com, yahoo.com, bebo.com) is uninformative for defining user interests. If used in interest models, it will affect their future predictions either by making the prediction task too difficult ($e.g.$, the future is likely unrelated to search engines or Web portals), or too easy ($e.g.$, we do not reward models for predicting continued frequent use of social networking Websites). Therefore, it was a requirement that u_t and $\{u_{t-5}, \dots, u_{t-1}\}$ did not contain search engines, portals, or social networking sites. These categories of Websites were also removed from the ground truth data.
- Since the ground truth is based on interaction behavior and not explicit ratings, we require many data points for it to be reliable. Therefore, we required that for each context trail, the short/medium/long futures be based on 10/20/50 Web page visits respectively. In addition, at least 50% of the pages in each future must have a label assigned by our ODP classifier.
- Since highly-active users may bias our sample, we selected at most 10 context trails from each user.

Although the size of T_x dropped to around 15% of the original, filtering the set of all context trails based on these criteria was necessary to create a high-quality data set for our study.

5.2 Methodology

As stated previously, the evaluation task was to predict future user interests following a visit to u_t based solely on u_t or on the available contextual information. We divided T_x into ten equally-

sized sets (and discarded the small remainder) to facilitate more reliable statistical testing. Context trails were randomly assigned with the constraint that each set contained only one trail per user. Each set contained 20,550 context trails.

The experimental procedure involved performing the following on the 20,550 context trails in each of the 10 experimental sets:

1. Find the short-, medium-, and long-term futures and build ground-truth interest models for each of them;
2. Build user interest models for different context sources, and;
3. Determine the accuracy of the context-based models in predicting the ground truth.

In the next section we describe the measures used to evaluate the predictive performance of our interest models.

5.3 Measures

The practical use of successful contextual modeling would most likely be in providing a surrogate for user interests and in the selection of sites to recommend to users as they browse the Web. The ODP labels in the six models (u_t plus five context variants) were stored as ranked lists in descending order of estimated informativeness. The ground truth labels were stored similarly. Therefore, we used standard IR measures to evaluate the predictive accuracy of the context-based models. We gave higher scores to the models for placing actual future interests high in the predicted list. For this reason, we focused on measures that scored the interest models well for achieving high early precision.

Our evaluation used precision, mean reciprocal rank, normalized cumulative discounted gain, and $F1$. We computed these measures separately for short-, medium-, and long-term futures. We now describe how we interpreted them for our study.

P@1: This measure required that the top predicted category label pl_1 for a context trail matched its top actual label l_1 in the specified future duration. If so, the user interest model would be given a score 1, and 0 otherwise. The scores over all context trails were then averaged to provide a final $P@1$ score for each set.

P@3: This measure compared top predicted category label pl_l for a context trail with any of its top 3 actual labels l_1, l_2, l_3 in the specified future duration. If there was a match, the user interest model would be given a score 1, and 0 otherwise. Scores were averaged to compute final $P@3$ scores as before. $P@3$ assumes that at most one label prediction would be used in a real system, but correctly predicting any of three dominant interest is useful.

Mean reciprocal rank: A standard alternative measure used often in Web search evaluation tasks is mean reciprocal rank (MRR); $e.g.$, Chowdhury and Soboroff's investigation reported its use in [6]. To compute this measure the top actual category label l_1 from a context trail was compared progressively down the ordered list of predicted category label predictions $pl_1, \dots, pl_p$ for the specified future duration. If l_1 matched pl_i, the score assigned was the reciprocal of the prediction rank position, $1/i$, and 0 otherwise. The scores over all context trails were then averaged to compute a final MRR for each set.

Normalized discounted cumulative gain: Another measure used was a variant of normalized discounted cumulative gain ($nDCG$) [13]. $nDCG$ biases towards the early retrieval of highly-relevant documents, although it also includes a recall component to the calculation. In our case the documents are ODP labels, such that the list of actual labels for a context trail is generated based on the specified future duration and is considered an ideal vector, with each actual label given a relevance score of 1. (An alternative approach would be to assign the label its corresponding frequency count as its relevance score.) The list of predicted labels $\{pl_1, \dots, pl_p\}$ generated by the user interest model is then compared to the ideal vector, and a discounted cumulative gain score is computed using a standard $log2$ discount factor. Our modification of the standard computation of $nDCG$ was to restrict the depth of the comparisons between the two label vectors to the minimum length of the two. The score was then normalized by dividing it by the maximum possible value that could be obtained to this depth. The scores over all context trails were averaged to provide a final $nDCG$ score for each set.

$F1$: Evaluation measures in similar settings such as the KDD Cup 2005 [20] often use the $F1$ measure (also known as test accuracy) which computes the harmonic mean of precision and recall. We include the $F1$ score to allow comparability to past work. For any context trail, the recall depth is computed based on the number of predicted labels for that trail. The scores over all trails were averaged to get the final $F1$ score for each set.

6. FINDINGS

6.1 Context source comparison

We performed a comparison of the predictive accuracy of user interest models generated based on u_t only and the five sources of contextual evidence. Table 1 shows the results of this comparison for each of the interest models, at each future time duration ($i.e.$, short, medium, long). Evaluation measures were computed over each experimental set and the results averaged. The maximum of the standard errors between the means is also reported.

The results show that the *interaction context* predicts user interests most accurately in the time immediately following the visit to u_t. This is likely because u_t does not represent the beginning or the end of the current task, and the interaction that occurs before and after u_t is task-related. The findings show that the interests of the user within one day of u_t are most accurately predicted by the *task context*, suggesting that the active work task may be lengthy. The findings also show that the long-term interests of the user ($i.e.$, those within one week of u_t) are most accurately predicted by the *historic context* of the user, but also the *social context* comprising the interests of other users who also visit u_t. These other users may share interests with the current user, making their long-term interests similar (and hence similarly predictive). Given the large sample sizes, the observed differences between the models for each measure are statistically significant using paired t-tests (all $t(20549) \geq 1.96$, all $p \leq .05$).

The observed variation in model performance for each of the three time durations suggests that different sources of contextual information may be suited for different tasks. For example, if a Website recommendation system must predict user interests immediately ($e.g.$, to recommend Websites that support task completion) it should leverage u_t, *interaction context*, and *task context*. However, if the system needs to predict longer term interests ($e.g.$, to recommend Websites of general interest) *historic context* and *social context* should be used. The *collection context* performed particularly poorly across all time durations, perhaps because it was related to u_t rather than the user, their task, or their observed interaction behavior.

Since $F1$ correlated strongly with the other measures we used it for the additional analysis in the remainder of this section.

Table 1. Predictive performance of contextual sources for different temporal durations (bold = best performing).

Context	Short					Medium					Long				
	$P@1$	$P@3$	MRR	$nDCG$	$F1$	$P@1$	$P@3$	MRR	$nDCG$	$F1$	$P@1$	$P@3$	MRR	$nDCG$	$F1$
None	0.52	0.52	0.52	0.54	0.52	0.35	0.35	0.35	0.36	0.35	0.21	0.21	0.22	0.22	0.21
Interaction	**0.62**	**0.62**	**0.62**	**0.64**	**0.62**	0.37	0.40	0.38	0.41	0.39	0.25	0.26	0.26	0.26	0.27
Task	0.58	0.60	0.60	0.62	0.61	**0.39**	**0.41**	**0.42**	**0.45**	**0.44**	0.23	0.26	0.25	0.30	0.30
Collection	0.14	0.16	0.15	0.14	0.11	0.09	0.08	0.07	0.12	0.08	0.02	0.04	0.02	0.08	0.03
Social	0.10	0.18	0.16	0.12	0.13	0.15	0.19	0.16	0.20	0.19	0.29	0.30	0.30	0.32	0.31
Historic	0.16	0.24	0.21	0.16	0.18	0.20	0.33	0.29	0.25	0.31	**0.37**	**0.43**	**0.38**	**0.38**	**0.40**
Std. errors	≤ .02	≤ .03	≤ .02	≤ .02	≤ .02	≤ .02	≤ .02	≤ .02	≤ .03	≤ .02	≤ .02	≤ .03	≤ .02	≤ .02	≤ .02

Table 2. $F1$ scores for varying levels of ODP category label back-off (bold = best performing).

Context	One-level			Two-level			Three-level			No backoff (table 1)		
	Short	Medium	Long	Short	Medium	Long	Short	Medium	Long	Short	Medium	Long
None	0.71	0.51	0.42	0.64	0.48	0.35	0.54	0.35	0.26	0.52	0.35	0.21
Interaction	**0.75**	0.51	0.43	**0.70**	0.48	0.31	**0.66**	0.42	0.28	**0.62**	0.39	0.27
Task	0.73	**0.55**	0.41	0.68	**0.50**	0.39	0.63	**0.46**	0.33	0.61	**0.44**	0.30
Collection	0.28	0.22	0.17	0.22	0.16	0.10	0.15	0.10	0.06	0.11	0.08	0.03
Social	0.21	0.26	0.42	0.18	0.23	0.37	0.17	0.20	0.35	0.13	0.19	0.31
Historic	0.23	0.37	**0.47**	0.19	0.34	**0.44**	0.19	0.32	**0.43**	0.18	0.31	**0.40**
Std. errors	≤ .03	≤ .02	≤ .02	≤ .03	≤ .02	≤ .02	≤ .03	≤ .02	≤ .02	≤ .02	≤ .02	≤ .02

Table 3. $F1$ scores for levels of label filtering based on page visit frequency (bold = best performing).

Context	Prediction ≥ 5			Ground truth ≥ 5			Pred. & Ground truth ≥ 5			No filtering (table 1)		
	Short	Medium	Long	Short	Medium	Long	Short	Medium	Long	Short	Medium	Long
None	n/a	n/a	n/a	0.58	0.38	0.24	n/a	n/a	n/a	0.52	0.35	0.21
Interaction	**0.65**	0.44	0.31	**0.64**	0.42	0.28	**0.68**	0.43	0.32	**0.62**	0.39	0.27
Task	0.63	**0.46**	0.33	**0.64**	**0.47**	0.33	0.64	**0.45**	0.36	0.61	**0.44**	0.30
Collection	0.14	0.10	0.03	0.15	0.12	0.05	0.16	0.15	0.07	0.11	0.08	0.03
Social	0.16	0.21	0.31	0.14	0.22	0.33	0.18	0.24	0.35	0.13	0.19	0.31
Historic	0.17	0.32	**0.42**	0.19	0.34	**0.43**	0.20	0.33	**0.45**	0.18	0.31	**0.40**
Std. errors	≤ .02	≤ .01	≤ .01	≤ .01	≤ .01	≤ .01	≤ .01	≤ .01	≤ .01	≤ .02	≤ .02	≤ .02

Table 4. Top-ranked source combinations based on $F1$ score.
Combinations with significant difference from best-performing context in Table 1 are marked (** = $p < .01$, * = $p < .05$).

Rank	Short		Medium		Long	
	Sources	$F1$	Sources	$F1$	Sources	$F1$
1	n, i, t, h, s, c	0.72^{**}	n, i, t, h, s, c	0.53^{**}	n, i, t, h, s, c	0.45^{**}
2	n, i, s, h, c	0.71^{**}	n, i, t, h, c	0.52^{**}	n, i, s, h, c	0.43^{**}
3	n, i, h, t, c	0.71^{**}	n, i, t	0.49^{**}	n, i, h, t, c	0.43^{*}
4	n, i, h, c	0.71^{**}	n, i, s, h, c	0.48^{*}	s, h	0.43^{*}
5	n, i, s, t, c	0.69^{**}	n, i, h, t	0.48^{*}	n, i, s, h, t	0.42^{*}
6	n, i, s, c	0.69^{**}	n, i, h, c	0.46^{*}	n, i, s, h	0.42
7	n, i, t, c	0.69^{**}	n, i, s, h, t	0.45	n, i, h, t	0.42
8	n, i, c	0.68^{*}	n, i, s, h	0.45	n, i, h	0.41
9	n, i, s, h, t	0.68^{*}	n, i, s, t, c	0.44	i, s, h, c	0.41
10	n, i, s, h	0.67^{*}	n, i, t, c	0.44	i, s, h	0.40

6.2 Handling near misses

The findings presented in Table 1 were derived based on matching the full ODP labels in the interest models with the full label in the ground truth. In that analysis we penalize the interest models for *any* mismatch between the predicted and actual label. However, small differences in estimates of user interests may be unimportant to a recommendation system. For example, interests represented by the ODP label "$Sports/Golf/Instruction/Golf\ Schools$" could also be represented by "$Sports/Golf/Instruction$" with only a slight loss in precision. In the analysis performed in Section 6.1 this would be regarded as a total miss, whereas it is actually a near miss. We investigate the effect of using label back-off involving the aggregation of ODP category labels under a parent node to mitigate the effect of near misses. We performed the experiment described in Section 6.1 an additional time by backing-off on all labels in the ground truth and in the predictions to a specified level. One-level back-off means convert all ODP to their top level ($e.g.$, "$Sports$"). Two- and three-level back-off means convert all labels to their top two and three levels respectively ($e.g.$, "$Sports/Golf$" and "$Sports/Golf/Instruction$"). In Table 2 we present the average $F1$ score for the additional analysis performed with back-off to different levels in the ODP hierarchy. As expected, the findings show an increase in the predictive accuracy of all models and for all time durations. The trends in the relative ordering of the interest models observed in Section 6.1 remain unchanged for label back-off (as does the statistical significance of the observed differences). The relative ordering of the interest models is insensitive to the granularity of the interest representation; any difference in model performance is not due to near misses.

6.3 Improving "judgment" confidence

The ODP labels used to represent user interests were assigned automatically based on page visit information extracted from our log data. Upon examining the predictions, we observed that due to the sparsity of page visits in the subset of the logs which constituted our ground truth, the labels periodically were of poor quality. Manual inspection of the context trails and the predicted labels implied the predictions were reasonable. We hypothesized that because a label assigned to the ground truth may represent only one visit to a Web page, and have a single user and one or two clicks, these sparse page visits may be distorting the evaluation measures. To investigate this, we performed the experiment of Section 6.1 an additional time, holding out labels assigned from the ground truth sets based on less than five visits. We repeated this process by holding out low-frequency labels from the prediction, and from the prediction and ground truth.

In Table 3 we present the findings of this additional analysis averaged across all experimental sets, for filtering the predictions, the ground truths, and both. Although the relative ordering of the models remains unchanged (and differences between them are still statistically significant), the $F1$ scores increase and the standard errors drop, giving us more confidence regarding conclusions drawn about the relative ordering of the context sources.

6.4 Combining contexts

A key aspect of the principle of polyrepresentation is the use of *cognitive overlap* between multiple contextual elements to strengthen the relevance signal of certain items [11]. We applied this principle directly in our study and in addition to considering contextual sources independently, we also considered source combinations. We performed the experiment of Section 6.1 an additional time, but systematically varied the combinations of contexts used. In total, 57 context combinations were tested.

For each combination, we obtained the specified external sources plus u_t if required. We selected the ODP category labels and their respective frequencies *for labels that appeared in all relevant interest models*; giving us the overlap between context sources. Some sources were more voluminous and may have higher frequency counts even though they had the same label ranking. Combining the frequency counts of all used sources would have biased the ranking. To rank items in the overlap, we adopted a simple strategy using the average rank position of each label across all used contexts, and the sorting based on that average.

In Table 4 we present the average $F1$ scores obtained for the top-10 best-performing combined models. To preserve space, the first letter of each source is used to denote its use in the model ($e.g.$, n=none ($i.e.$, u_t only), i=interaction, etc.). The findings show that using a combination of sources leads to more accurate future predictions in the short-, medium-, and long-term. Those combinations with an $F1$ score that is significantly different from the best performing model in Table 1 (using a paired t-test) are marked. For each time duration, there exists at least one context combination that significantly outperforms all contexts in isolation; this supports the principle of polyrepresentation. Data in Table 4 demonstrates that certain contexts are required to obtain high prediction accuracy ($i.e.$, u_t and *interaction context* in short-term predictions, *task context* in medium-term predictions, and *social context* and *historic context* in long-term predictions).

7. DISCUSSION

We studied the effectiveness of different sources of contextual evidence, and their overlap, for user interest modeling. The findings of our study suggest that the best-performing contextual sources are dependent on the duration between u_t and the end of the prediction window. This has implications for the systems that use contextual information to support post-query navigation and general browsing behaviors. For example, these systems must not treat all context sources equally. Weights should be assigned to each source depending on whether the system is recommending Web pages that are relevant to the immediate situation, the current work task, or the user's general interests. The contexts as defined could be implemented using server-side lookups (*task*, *collection*, and *social*) or client-side code (*interaction* and *historic*).

Our finding that interests within an hour of u_t could be predicted by local context information such as u_t itself and $\{u_{t-5}, \ldots, u_{t-1}\}$ suggests that topical interests, as represented by ODP category labels, are not highly changeable within a short period of time. Search queries and information needs may evolve during this time, but topical interest may be less dynamic. The high effectiveness of *task context* in predicting activities within one day of u_t may be due to its consideration of the current situation as well as similar situations encountered by other users. Since by definition *task context* is broader than *interaction context*, it is more likely able to include task variants that could appear within the next full day. The effectiveness of the *historic context* and *social context* in predicting longer-term user interests is likely related to their ability to predict the general interests of each user. They are effective at doing so since they have access to large amounts of long-term information for a user and similar users.

We demonstrated that polyrepresentation is viable for user interest modeling. As shown in Table 4, models based on overlap between sources (especially between all sources) performed better than any

individual source. More work is necessary to determine how best to combine sources beyond linear averaging, including using machine learning to automatically determine source weights.

The observed differences in this study may be related to the nature of the sources that were selected. For example, it may have been better to use anchor text rather than in-links as the *collection context*. However, given that this study was log-based, and that we had to transform all contexts to URLs for ODP labeling, the definitions of context we adopted seem reasonable. User studies conducted in tandem with human labeling of user interests are important next steps to validate our claims.

8. CONCLUSIONS

In this paper we have presented a systematic, log-based study of the effectiveness of five variant sources of contextual information for user interest modeling. Given the prevalence of post-query navigation and general browsing, we conducted this study within a framework of Website recommendations rather than search results. We extracted browsing contexts from toolbar logs and built a variety of user interest models based on the current page, contextual variants, and overlaps between contexts. The interest models were required to predict short-, medium-, and long-term user interests. Our findings show that the predictive value of each contextual sources varies according to the time duration of the prediction. We showed that the relative ordering of the contexts for each time duration was unaffected by coarser representations of user interests and higher-quality predictions or ground truths, and that context overlap was more effective than any individual context. Website recommendation systems should use context, because doing so outperforms not using it. However, the systems may need to vary the source depending on the modeling task. Our findings should enhance Website recommendation systems and facilitate improved information-gathering support for their users.

9. REFERENCES

[1] Agichtein, E., Brill, E. & Dumais, S.T. (2006). Improving web search ranking by incorporating user behavior information. *Proc. SIGIR*, 19-26.

[2] Bilenko, M. *et al.* (2008). Talking the talk vs. walking the walk: salience of information needs in querying vs**.** browsing. *Proc. ACM SIGIR*, 705-706.

[3] Broder, A. (2002). A taxonomy of Web search. *ACM SIGIR Forum*, 36(2), 3-10.

[4] Budzik, J. & Hammond, K. (1999). Watson: anticipating and contextualizing information needs. *Proc. ASIS*, 727-740.

[5] Byström, K. & Järvelin, K. (1995). Task complexity affects information seeking and use. *IP&M*, 31(2), 191-213.

[6] Chowdhury, A. & Soboroff, I. (2002). Automatic evaluation of world wide web search services. *Proc. SIGIR*, 421-422.

[7] Claypool, M. *et al.* (2001). Inferring user interest. *IEEE Internet Computing*, 5(6), 32-39.

[8] Dupret, G., Murdock, V. & Piwowarski, B. (2007). Web search evaluation using clickthrough data and a user model. *Proc. WWW Workshop on Query Log Analysis.*

[9] Goldberg, D. *et al.* (1992). Using collaborative filtering to weave an information tapestry. *CACM*, 35(12), 61-70.

[10] Gunduz, S.U. & Özsu, M.T., (2003). Recommendation models for user accesses to web pages. *Proc. Conf. on Artificial Neural Networks*, 1003-1010.

[11] Ingwersen, P. (1994). Polyrepresentation of information needs and semantic entities: elements of a cognitive theory for information retrieval interaction. *Proc. SIGIR*, 101-110.

[12] Ingwersen, P. & Järvelin, K. (2005). *The Turn: Integration of Information Seeking and Retrieval in Context*. Springer.

[13] Järvelin, K. & Kekäläinen, J. (2000). IR evaluation methods for retrieving highly relevant documents. *Proc SIGIR*, 41-48.

[14] Joachims, T. & Radlinski, F. (2007). Search engines that learn from implicit feedback. *IEEE Computer*, 40(8), 34-40.

[15] Kellar, M., Watters, C. & Shepherd, M. (2007). A field study characterizing web-based information seeking tasks. *JASIST*, 58(7), 999-1018.

[16] Kelly, D., Dollu, V.D. & Fu, X. (2005). The loquacious user: A document-independent source of terms for query expansion. *Proc. SIGIR*, 457-464.

[17] Khalil, F., Li, J. & Wang, H. (2008). Integrating recommendation models for improved web page prediction accuracy. *Proc. Australasian Conf. on Comp. Sci.*, 91-100.

[18] Kim, K. & Allen, B. (2002). Cognitive and task influences on web searching behavior. *JASIST*, 53(2), 109-119.

[19] Larsen, B. & Ingwersen, P. (2002). The boomerang effect: retrieving scientific documents via the network of references and citations. *Proc. SIGIR*, 397-398.

[20] Li, Y., Zheng, Z., & Dai, H.K. (2005). KDD CUP-2005 report: facing a great challenge. *SIGKDD Expl.*, 7(2), 91-99.

[21] Lieberman, H. (1995). Letizia: an agent that assists web browsing. *Proc. IJCAI*, 924-929.

[22] Oard, D. & Kim, J. (1998). Implicit feedback for recommender systems. *Proc. AAAI Wkp. on Rec. Sys.*, 81-83.

[23] Pitkow, J. *et al.* (2002). Personalized search. *CACM*, 45(9), 50-55.

[24] Piwowarski, B. & Zaragoza, H. (2007). Predictive user click models based on click-through history. *Proc CIKM*, 175-182.

[25] Qiu, F. & Cho, J. (2006). Automatic identification of user interest for personalized search. *Proc. WWW*, 727-736.

[26] Shen, X., Dumais, S. & Horvitz, E. (2005). Analysis of topic dynamics in web search. *Proc. WWW*, 1102-1103.

[27] Shen, X., Tan, B. & Zhai, C. (2005). Context-sensitive information retrieval using implicit feedback. *Proc. SIGIR*, 43-50.

[28] Singhal, A. (2001). Modern information retrieval: A brief overview. *IEEE Data Engineering Bulletin*, 24(4), 35-43.

[29] Skov, M., Larsen, B. & Ingwersen, P. (2006). Inter and intra-document contexts applied in polyrepresentation. *Proc. IIiX*, 97-101.

[30] Tan, B., Shen, X. & Zhai, C. (2006). Mining long-term search history to improve search accuracy. *Proc. SIGKDD*, 718-723.

[31] Teevan, J., Dumais, S.T. & Horvitz, E. (2005). Personalizing search via automated analysis of interests and activities. *Proc. SIGIR*, 449-456.

[32] Terai, H. *et al.* (2008). Differences between informational and transactional tasks in information seeking on the web. *Proc. IIiX*, 152-159.

[33] Thatcher, A. (2008). The influence of web experience and task type. *IP&M*, 44(3), 1308-1329.

[34] White, R.W. & Drucker, S.M. (2007). Investigating behavioral variability in web search. *Proc. WWW*, 21-30.

[35] White, R.W. *et al*. (2009). Characterizing the influence of domain expertise on web search behavior. *Proc. WSDM*.

[36] White, R.W. *et al*. (2005). A study of factors affecting the utility of implicit relevance feedback. *Proc. SIGIR*, 35-42.

[37] Zhang, Y. & Koren, J. (2007). Efficient bayesian hierarchical user modeling for recommendation systems. *Proc. SIGIR*, 47-54.