Collecting Relevance Feedback on Titles and Photographs in Weblog Posts

Amy Campbell
Department of Linguistics
University of California, Berkeley
1203 Dwinelle Hall
Berkeley, CA 94720-2650
amycampbell@berkeley.edu

Christopher Wienberg and Andrew S. Gordon
Institute for Creative Technologies
University of Southern California
12015 Waterfront Drive
Los Angeles, CA 90094
cwienberg@ict.usc.edu, gordon@ict.usc.edu

ABSTRACT
We investigate new interfaces that allow users to specify topics of interest in streams of weblog stories by providing relevance feedback to a search algorithm. Noting that weblog stories often contain photographs taken by the blogger during the course of the narrated events, we investigate whether these photographs can serve as a proxy for the whole post when users are making judgments as to the post's relevance. We developed a new story annotation interface for collecting relevance feedback with three variations: users are presented either with the full post as it appears in a weblog, an embedded photograph, or only the title of the post. We describe a user evaluation that compares annotation time, quality, and subjective user experience across each of these three conditions. The results show that relevance judgments based on embedded photographs or titles are far less accurate than when reading the whole weblog post, but the time required to acquire an accurate model of the user's topic interest is greatly reduced.

Author Keywords
Weblogs; photographs; relevance feedback; user interfaces for machine learning; user study.

ACM Classification Keywords
H.5.2. User Interfaces (Interaction Styles, Evaluation).

INTRODUCTION
Machine learning algorithms are now routinely used in commercial web applications, and many enlist the users directly as providers of training data. For example, a movie rental service will incorporate a user's ratings of previous rentals when recommending new ones (e.g., Koren et al., 2009). Similarly, a personalized news service will select articles for a particular topic based on rules learned from a user's previous judgments of topic relevance (e.g., Stefik & Good, 2011). These applications and others have bolstered research interest in intelligent user interfaces that best facilitate the user's task of providing training data (Amershi et al., 2011). Successful approaches strike a balance between the user's need for a quality end-user experience and the system's need for copious amounts of quality training examples.

In this paper, we address the problem of training a system to identify weblog posts that are relevant to a user's interest. Specifically, we focus on training a system to recognize when a new weblog post is a personal story about an activity of interest to a particular user, given his or her previous relevance annotations. For example, a cardiologist may be interested in reading personal stories of people's experiences of having heart attacks. A used car salesman may be interested in reading personal stories from car buyers describing their experiences in negotiating with other salesmen. Parents of children with life-threatening diseases may be interested in reading personal stories from other parents who have gone through similar experiences of diagnosis, treatment, and recovery. More generally, people are interested in stories from other people that are relevant to their own personal lives in some direct way.

From a technology perspective, this problem is similar to that of Topic Detection and Tracking (TDT), a long-standing research challenge of developing algorithms for the automatic organization of news stories by the real-world events that they describe (Fiscus & Doddington, 2002). As with news articles, the problem of topic detection can be addressed by learning the lexical features from the documents that are predictive of topic relevance. However, the unique characteristics of personal stories in weblog posts do not lend themselves to other TDT concerns. We expect bloggers to author only one narrative about any particular event in their lives, so there is little concern for the TDT tasks of story segmentation, topic tracking, first story detection, and link detection. Conversely, the unique characteristics of weblog storytelling present different challenges and afford new opportunities.

In this paper, we seek to exploit one of the unique characteristics of weblog storytelling: stories frequently include photographs taken during the course of the narrated events. For example, a blogger telling the story of a fishing

IUI'12, February 14–17, 2011, Lisbon, Portugal.

trip might include a photo of the fish that they caught, or of the damage to their boat when they crashed into a rock. Taken out of the context of the weblog story and viewed in isolation, the activity contexts of these photos are often still recognizable; a photo of a fish on the end of a fishing pole is strongly suggestive of the topic of fishing trips. If this is true often enough, then we postulate that photographs in weblog stories may serve as an effective proxy for the content of the entire post. This would allow users to quickly train a system to recognize their topic of interest by judging the relevance of photographs extracted from unlabeled posts, rather than reading the actual text of the story.

This relationship between text and images has been exploited in user interfaces of previous research. Early work on context-based multimedia retrieval (Dunlop & van Rijsbergen, 1991) sought to use the textual context to enable the retrieval of media that, at the time, could not be retrieved by content features (e.g. images). Even as content-based image retrieval has advanced, researchers have continued to use textual queries as an effective way of beginning a search for relevant images (e.g. Villa et al., 2010). Recently, Zha et al. (2010) explored a novel method for interactively improving image search by presenting a user with photographs that are exemplars of additional query terms that further disambiguate their query. In our research, we pursue an analogous approach: using images (photographs) to interactively improve the retrieval of relevant textual content. In doing so, we consider the relationship between text and images in the opposite direction than seen in context-based image retrieval. Our aim is to use images as a way to retrieve text similar to the text that surrounds them, i.e. image-based context retrieval.

In the sections that follow, we describe an experiment to determine whether the annotation of photographs appearing in weblog stories could serve as a suitable proxy for the full post when a user trains a system to recognize his or her topic of interest. We begin with a review of this genre of social media and a summary of the role of photography in existing corpora of weblog stories. We then describe a web-based user interface for annotating the relevance of weblog stories, with variants that show users only a photograph from an unlabeled post, the title of a post, or the full text of a post as it appears on the web. We describe a user evaluation to compare the effectiveness of these three variants across three different topics, and compare the time and accuracy of annotations of photos and titles to annotations of full web posts. We conclude with a discussion of the implications of these findings for future applications.

PERSONAL STORIES IN WEBLOGS

The phenomenal rise of weblogs over the last decade has created new opportunities for researchers to study personal communication on a massive scale. Within computer science, the stories that bloggers post in their weblogs have been seen as valuable sources of knowledge, both for people and computers. Gordon (2008) reviews the utility of weblog stories for organizations of people, particularly as sources of real-world experiences to serve as the basis for fictional scenarios in immersive training environments. Gordon et al. (2011) investigates the utility of weblog stories as a knowledgebase for computers, capitalizing on the causal structure of stories to guide automated reasoning systems to make causal inferences in commonsense situations. These efforts and others are predicated on the availability of very large corpora of personal stories. In this section, we describe an existing corpus of nearly one million personal stories from weblogs that we used in our experiments, along with a discussion of photographs within this dataset.

Stories in the ICWSM 2009 Spinn3r Dataset

Gordon and Swanson (2009) estimated that only 4.8% of all non-spam weblog posts are personal stories, which they define as non-fictional narrative discourse that describes a specific series of causally related events in the past, spanning a period of time of minutes, hours, or days, where the storyteller or a close associate is among the participants. Using supervised machine learning methods for text classification, these authors identified nearly one million personal stories among the 25 million English-language weblog posts from the ICWSM 2009 Spinn3r Dataset (Burton et al., 2009). We obtained this corpus of nearly one million personal stories for use in our experiments.

In our early user interface designs, we realized that it would be necessary to present users with content from these stories that is not encoded in the dataset format, e.g. a photograph that was embedded in the text of a post. For this reason, we needed to download the full post as it appears on the web. However, the posts in this corpus were nearly three years old at the time of our experiments, and many of them were no longer available. A study of 100 stories sampled at random from the corpus revealed that 26% of URLs no longer linked to the post content. For example, the popular blogging platform Vox.com had been the fifth largest source of weblog stories, accounting for 2.5% of the corpus, but had shut down entirely in September of 2010. We executed a series of scripts to identify broken links in the million-story corpus, and filtered out these items in our subsequent experiments. Additionally, we removed entries that had been marked as private, marked as containing adult content, or contained JavaScript that would prevent content from being embedded within our interface prototypes. After applying these filters, the story corpus was reduced to 627,782 stories from its original 960,098 stories (65%).

Photographs in Weblogs Stories

Ever since photography became easily accessible to the public at large, it held an important role in how we tell personal stories (Chafen, 1987). In much the same way that a physical photograph can be a catalyst for storytelling in face-to-face communication, digital photography

contributes an element of immersion to the textual stories that people write on their weblogs and in other social media. In our examination of the million-story corpus of Gordon and Swanson (2009), we were struck by the sheer frequency of photographs that appeared in weblogs, and how often these photographs had been taken in the context of the narrated events. To investigate the use of photographs in more detail, we conducted a series of analyses to quantify the relevance of photograph content to stories in which they were embedded.

First, we sought to determine the ratio of photographs to other types of images embedded in corpus documents. We began by extracting every image tag (<img>) in the HTML of every story in our corpus. As prepared by Spinn3r.com, the weblog posts in this corpus are presented with all extraneous "chrome" HTML removed, including sidebars, advertising, and navigation structure. Consequently, the image tags in the corpus HTML are only those included by the author as a part of a given post. However, only a subset of these images consisted of photographs of real life events. We randomly sampled 100 image links where the image was still available on the web, and labeled each as either a photograph or a non-photographic image. Of these, 71% of the images were photographs, with others consisting largely of navigation icons, computer generated images, and images used to track hits on a page.

Second, using the same random subset of 100 images, we considered several simple heuristics for automatically distinguishing between photographs and non-photographs. We found that simply establishing a minimum width and height for images served as an effective photograph filter. Requiring that an image was at least 50 pixels wide by 50 pixels tall yielded 0.90 precision and 1.0 recall of photographs. An even better heuristic was to check the aspect ratio of the image to see if it matches common aspect ratios of photographs. We compared the ratio of the length of an image's shorter edge to its longer edge to the ratios 2:3, 3:4 and 9:16, corresponding to three of the most popular aspect ratios for still photography. By permitting a 5% deviance from these pre-set aspect ratios, this heuristic achieves a precision of 1.0 and recall of 0.94 in identifying photographs in the random sample. We then applied this heuristic to every image still available in our corpus of personal stories, identifying 534,514 photographs.

Third, we investigated the variations in how these photographs were distributed across the stories still available on the web. Only 13% of the stories still available on the web contained photographs that were also still available. In these stories, the mean number of photographs was five, and 40% had only one photograph. Assuming that stories with and without photographs disappear from the web at the same rate, we estimate that 19.5% of stories included a photograph when they were originally posted.

Fourth, we investigated the relevance of photographs to the narrated events of the story. We observed that not all of the photographs contained in weblog posts of personal stories had a direct relationship to the text of the post, e.g. a story about living in a retirement community might include a photo of the author's grandchildren. To determine what percent were relevant, we selected a random sample of 100 photographs from the set identified using our photograph heuristic. We annotated each as to whether it was relevant to the story in which it was embedded. Here we defined relevance to mean that the photograph depicted or was likely to have been taken in the context of the events that are narrated in the story. In cases where our heuristic failed or the post was not a personal story, we marked the photograph as non-relevant. In this sample, 82% of photographs were relevant to the stories in which they were embedded.

Fifth, we sought to develop a heuristic for selecting the most likely relevant photograph in stories that contained more than one. That is, we wondered whether the first, middlemost, or last photograph embedded in a personal story most often depicted the events that were narrated in the text. We randomly selected 100 stories that contained at least two photographs, then annotated each photograph as either relevant or irrelevant to the story surrounding it. In total, we annotated 672 photographs in this sample, and found 530 photographs (79%) in this sample to be relevant to the story in which they appear. We then compared the relevance of the first photograph, the middlemost photograph, or the last photograph. We found that 79% of stories on our sample have a relevant first photograph, 79% have a relevant photograph in the middlemost position of all photographs in the story, and 77% of stories have a relevant final photograph. Given no clear preference, we selected a heuristic that agreed with our intuition, and selected the middlemost photograph as the one most likely to be central to the story. We then applied this heuristic to each of the 105,456 available stories that contained photographs, selecting a single photograph for each that best represented its content.

In the course of these analyses, we saw several different ways that people use photographs in weblog storytelling. We made several qualitative observations that we feel are important to understanding the character of weblog story photography. The most common way authors present their photography was to interweave them with the text of the story. When posts are written in this style, the photographs often carry at least as much burden of telling the story as the narrative text. In some cases the photographs constitute the narration, with nearly all of the text consisting of photograph captions. These cases of photographs interwoven with narrative text are analogous to sitting around the family photo-album, or watching the slides of a friend's vacation. Here photographs provide a visual storytelling experience, where the author presents the photographs with enough description to bind them together into a coherent narrative, but allows the photographs to provide the details.

The analogy to physical photo albums breaks down when considering the topics of the stories containing photographs. Although stories with photographs are often compelling, the events that are narrated in weblogs tend to be more mundane than those that might be documented in a family photo album. Owing perhaps to the medium of weblogs and to the proliferation of digital photography, photographs in weblog stories often depict events common in everyday life. While a family album might be reserved for photographs from vacations, holidays, and other extraordinary moments in life, a weblog story is at least as likely to have photographs from the author's daily walk around their neighborhood, the leaky pipe under their kitchen sink, or their halfway finished art project.

Nearly all of the photography we encountered was taken in the context of leisure time, with almost none depicting the professional lives of the authors. This is to be expected; it is often inappropriate to take photographs in a work environment. However, there are many stories without photographs that describe situations related to work, such as the author complaining about his or her boss or celebrating an accomplishment. This observation reveals a more general point about weblog story photography: stories featuring photography are not representative of all weblog stories as a whole. Rather, weblog stories with photography over-represent stories that are particularly photogenic, where photography is not only appropriate, but illuminative of the activity. For instance, stories without photographs may be more likely to describe daily commutes and writing novels, i.e. things that are not particularly visual. Some stories about these topics will have photographs, but these stories generally have an additional, visual component, e.g. a daily commute where a stone had cracked the windshield of the author's car.

Photographs that are not relevant to the story are generally relevant to the author's life in some other direct way. In these cases, the author generally includes a statement cross-referencing another story to which the photograph relates. Generally, the author implies that his or her readers will have enough knowledge about the author's life—knowledge gained from either previous posts or personal acquaintance—to understand the photographs and the context in which those photographs were taken.

The author of the post is almost always the one who originally took the photograph, as indicated by his or her own descriptions.

ANNOTATION INTERFACES FOR STORY RELEVANCE

The central aim of our research effort was to develop and evaluate new interfaces that would allow end-users to train a machine learning algorithm to recognize weblog stories that are relevant to their specific interests. As a user interface design project, our primary concerns were to minimize the time required to gather copious amounts of high quality training data from the user, while maximizing the quality of the subjective user experience. We designed and implemented three story annotation interfaces, and paired them with a back-end story retrieval system so as to evaluate the effectiveness of these interfaces in situations that approximate those found in web-scale search applications.

In each of the three interfaces, a user would be presented with content from weblog stories selected by the back-end system as relevant to the user's interest, and asked to rate them as either relevant or not relevant to this interest. Relevance judgments would then be used to adjust the system's relevance model, and additional materials would be selected. The three interfaces varied primarily on the type of content that was presented to the user. In the "full text" condition, the full weblog post would be presented to the user as it appears on the web. In the "photo-only" condition, a photograph from the post would be shown, selected using the heuristic described in the previous section. In the "title-only" condition, the title of the post would be the only content presented to the user. In each condition, the full text of the annotated weblog story was used as relevance feedback. For example, if a photograph was annotated as relevant to the user, the full text of the story in which it was embedded would be used to adjust the system's relevance model.

Our hypothesis was that users would be able to quickly and accurately judge the relevance of embedded photographs to their topic of interest. Given that not all photographs are relevant to the events narrated in the story, we expected that a user's relevance feedback would contain inaccuracies. However, we anticipated that noise in this feedback would be overcome by the quantity of training data that could be annotated by a user in a given period of time, as compared to the full text interface. Our third condition, the "title-only" interface, provided an alternative to photographs that could also be quickly judged by users. However, we expected that titles alone would not provide sufficient information to make correct relevance judgments often enough to overcome the noise that was introduced.

Our designs were implemented as a Django (Python-based) web application with two back-end services for database access and textual search. We used the Terrier IR Platform (Ounis et al., 2007) as our back-end text search engine, using the default Divergence from Randomness retrieval model on an index of each of the 627,782 stories in the collection. Beginning with an initial topic description, the system employed a simple relevance-feedback method to update the query after each user annotation (Rocchio, 1971). The updated query is then run against the corpus, yielding an updated list of search results from which new content is chosen for display and annotation. This type of interactive, example-based machine learning is known to be a promising approach when dealing with large, unstructured data sets (Amershi et al., 2011), and is well suited to our goal of finding clusters of related stories in a large corpus of weblog posts.

Design Principles

Our interface design choices were guided by a small set of general principles. First, we limited ourselves to using *readily available, weblog author-supplied content*. This principle informed our choice of content items to be annotated in each interface. An interface displaying the full text of each story was necessary to establish baseline performance. In determining which sub-story content items could serve as potential proxies for the full text we chose story titles because they are tagged in the ICWSM 2009 Spinn3r Dataset, and photographs because they are easily stripped from the full story HTML. We ruled out an interface that would ask users to annotate query-based summarizations, for instance, because it would require additional processing to generate that content.

A second principle that informed our design was *user choice*. We display multiple content items on a single screen and do not require users to annotate them in any particular order; in fact, we instructed experimental users to work at their own pace in any order they choose. The title- and photo-based interfaces display six content items in a grid arrangement. While there isn't sufficient screen space for a grid in the full text interface, we retain the ability for users to choose among content items by including a set of six tabs along the side of the page that control which story is displayed in the main frame (see Figure 1 below). In all three interfaces we allow users to skip content items, rather than forcing them to make a decision about every item.

The final principle we employed in designing our system was *consistency of appearance and functionality* across interfaces. Shneiderman (1998) urges to "strive for consistency" as the first golden rule of interface design. Specifically, in similar situations the terminology used, controls displayed, and actions required should be kept as consistent as possible. This principle was important to us because each subject in our user study would be using three different interfaces. A consistent interaction mechanism across each interface, with a similar look and feel, allowed the user to focus on the primary annotation task, rather than on the interface itself.

Learning algorithm

To support the comparative evaluations of our interface designs, we implemented a standard query-refinement algorithm based on relevance feedback (Rocchio, 1971). Beginning with an initial query, each user annotation updates the query by combining query terms with the terms in the annotated document, following a weighted mixture scheme. Following common practice, Rocchio parameters were set to $\alpha = 1$, $\beta = 1$, and $\gamma = 0$, which weights the initial query and relevant documents equally, and effectively ignores negative feedback. In order to maintain query times that supported real-time interaction, it was necessary to restrict refined queries to the 50 most information-laden terms, as indicated by their tf-idf scores computed over the entire corpus. Queries were updated after each annotation

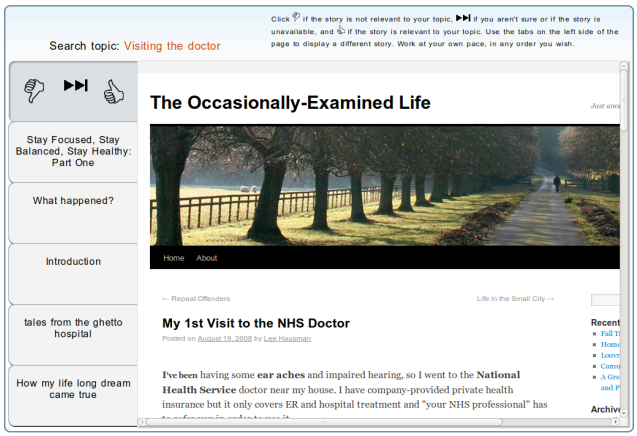

Figure 1. Full-text annotation interface

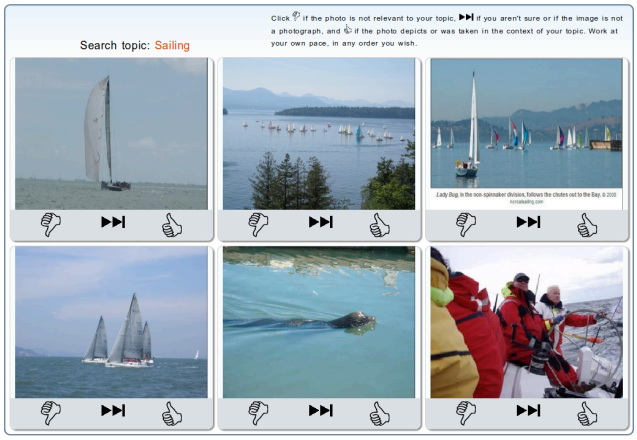

Figure 2. Photo-only annotation interface

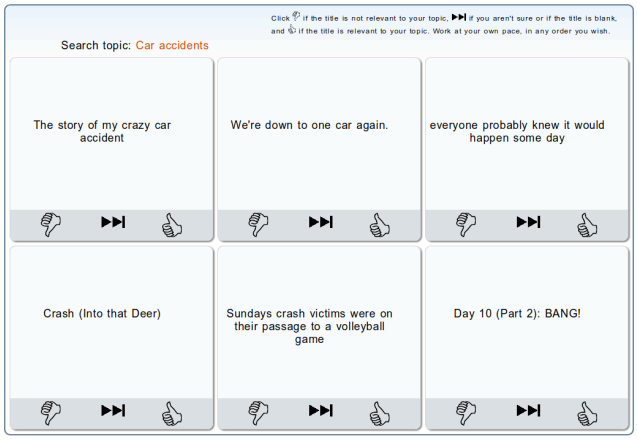

Figure 3. Title-only annotation interface

using this approach, and the highest-ranking story that had not yet been annotated by the user was selected to replace the item just annotated in the user interface.

Three annotation interfaces

In the full-text interface (Figure 1) users see the full text versions of weblog stories that match the search query, as they would appear on the authors' weblogs. These stories usually contain titles and may contain embedded photos. A set of tabs at the left side of the page allows users to change the displayed story. The tab corresponding to the current story is visually distinguished by its color and by the appearance of the annotation icons. Users are instructed to click thumbs-down if the story is not relevant to the search topic, skip if they aren't sure, and thumbs-up if the story is relevant. Once a story is annotated its text is used to revise the query and search the collection, and then the item is replaced by the highest-ranking story without an annotation.

The photo-only interface (Figure 2) displays photographs embedded in the matching stories. Although the overall look of the page is consistent with the full text system, the photo panels are arranged in a grid to maximize user choice. The annotation icons used in this interface are the same as those used in the full text interface. Users are instructed to consider a photograph relevant if it was taken in the context of the activity denoted by the search topic.

The title-only interface (Figure 3) displays only the titles of the weblog stories matching the search query. The page layout and the annotation controls are identical to those of the photo-based interface. Users are instructed to consider a title as relevant if it seems to be the title of a relevant story.

EVALUATION

In order to evaluate our system we designed an experiment to measure the effects of using titles or photographs as proxies for full stories in the annotation task. Objectively, we wondered how system accuracy would change with respect to annotation time and number of annotations in each of the three annotation interfaces. Subjectively, we wanted to determine how users perceived each interface's ease of use and enjoyment, as well as their confidence levels in the accuracy of each interface.

Participants

We recruited 18 people to participate in a user study; all but one were employees of the University of Southern California's Institute for Creative Technologies. There were 8 male and 10 female participants, and their ages ranged between 20 and 61. At a minimum, all participants had attended some college; 5 held a Bachelor's degree and 11 held a Master's or Doctoral degree. Job titles were diverse, with research and technical occupations being the most common, though administrative jobs were represented as well. 15 of the participants reported being “experienced” or “very experienced” computer users, 11 described themselves as regular readers of weblogs, and 4 stated that they maintain their own weblog. Each participant was paid $20 for approximately 45 minutes of participation.

Experimental Design and Procedures

All 18 participants used all three interfaces during the experiment. Subjects used the photo-only and title-only interfaces for exactly five minutes each. Subjects used the full-text interface for exactly 15 minutes, which allotted them enough time to read and annotate a suitable number of weblog posts. Subjects were asked to use each interface for the full amount of allotted time without stopping, but otherwise annotated at their own pace.

Since all users interacted with all interfaces, we needed to ensure that a user would not encounter the same content in different interfaces. For instance, if a user read a story in the full text interface, it would bias the results if the same user later annotated the title of the same story in the titles-

Topic	Initial "boring story" search query
Topic 1: Sailing	We went sailing. We motored our sailboat out of our slip in the harbor. We raised the main sail and the jib with the halyards, and trimmed the sails with the lines. The wind was blowing some knots, and the waves were good. We tacked into the wind, then jibed away from the wind. We sat on the deck of the boat and enjoyed the weather. At the end of the day, we returned to the harbor.
Topic 2: Car accidents	I crashed my car. I was driving in heavy traffic and the visibility wasn't very good. Out of nowhere this other car came at me and smashed into the side of the car. The airbag blew up in my face, but the seatbelt worked. I checked to see that I was okay, then got out of the car to see the damage. The other driver was pretty shaken up. The police showed up and an ambulance came, but everyone was okay. I got the insurance information from the other driver, and a tow truck came to pick up my car.
Topic 3: Doctor visits	I had a doctor's appointment. I went to the medical offices and checked in with the receptionist and sat in the waiting room. When it was my turn I went to the exam room and a nurse took my blood pressure took a blood sample using a syringe. I changed into an examination gown. The doctor came in and asked me about my health. The doctor listened to my heart using a stethoscope. The doctor said I needed to watch my cholesterol level. When my appointment was over, I went to the receptionist and gave them my insurance information.

Table 1. Initial "boring story" queries for the three topics used in the experiment

only interface. For this reason, we developed three queries for the user study and planned for each participant to use a different query for each of the three interfaces.

The queries were formulated as short stories that contained vocabulary typical of a particular activity context. This style follows the "boring story" format advocated by Gordon and Swanson (2008) for searching indexed story collections. Table 1 lists the search topics and "boring story" queries that we used. Query 1 is a simple story about the activity of sailing, query 2 is about a car accident, and query 3 is about a visit to the doctor's office.

Subjects were randomly sorted into three experimental groups that differed in the mapping of queries to interfaces. Within each group we counterbalanced for the order in which interfaces were used. There are six possible orders of use of the three interfaces, and we assigned one participant in each group to each order. Each experimental group thus comprised 6 participants, for 18 participants total.

RESULTS

During the experiment we collected two types of data suitable for quantitative analysis. First, the annotation system itself logs information about each interaction session and all of the user's actions therein. Second, after each interaction session participants were asked to rate several aspects of their interaction using paper and pencil rating scale measurements.

Objective measures

The system logs captured the following information about each user session: interface type (full text, titles-only, or photos-only), user ID, start time, and initial query. For each user annotation the system recorded the story ID, the query that resulted in the story being delivered and the associated confidence score, the user-assigned annotation value ("yes", "no", or "skip"), and a time stamp. This information was later used to reproduce the user's search results at various points in their session, allowing us to evaluate our objective measures: accuracy over time and over number of annotations.

Our first task in analyzing the system data was to check the level of inter-rater agreement in the full text interface, which we expected to yield the most accurate annotations because users had access to all of the story information.

In calculating inter-rater agreement we disregarded stories skipped by users and looked only at positive or negative relevance annotations, a decision motivated by the heterogeneous nature of the skip category. Participants were instructed to skip for a number of reasons (if they weren't sure whether the content was relevant or not, if an error message loaded instead of the live weblog story, etc.) and so two users' "skip" annotations could not be said to "agree" in any meaningful way. Skip annotations accounted for about 4% of annotations in the full text interface (21 skips in 269 total annotations), 19% of annotations in the titles-only interface (278/1484), and 8% of annotations in the photos-only interface (144/1853).

Table 2 presents the inter-rater agreement measures for each of the three interfaces. Pairwise agreement is based on all items that two annotators both rated, and here we report raw agreement (a) as well as agreement normalized for chance (Cohen's Kappa) (b). Scores are aggregated across all three topics. Results indicate that inter-rater agreement was not perfect, even in the full text interface. It appears that much of the disagreement was due to genuine subjectivity in story interpretation. For instance, the topic "car accidents" included one story that many participants viewed where the author had been hit by a car while riding a bicycle. It is reasonable to think that some readers would consider this to be a story primarily about a bike ride (with the collision being an important sub-event), while others would consider it to be primarily about the collision (with the bike ride being background context).

In order to evaluate the accuracy of the titles-only and photos-only interface we generated a "gold standard" – the definitive set of relevance annotations for a particular search topic – from the full text annotations. The gold set included all stories annotated as relevant or not relevant in the full text interface (again omitting "skip" annotations). We dealt with inter-rater disagreement for a particular story by majority rule: if the positive annotations outnumbered the negative annotations then the gold standard annotation for that story was "relevant". The gold set for query 1 contained 47 stories (43 relevant, 4 not relevant), 67 stories for query 2 (35 relevant, 32 not relevant), and 49 stories for query 3 (46 relevant, 3 not relevant), for a total of 163 stories.

Table 2, row C, lists the raw agreement of annotations from the photo-only and title-only interfaces with the gold-standard annotations, averaged across subjects. As we expected, users were more likely to assign correct relevance annotations to photographs than to titles, although neither interface achieved high levels of accuracy.

Our system and experimental design afforded a novel

Measure	Full-text	Photo-only	Title-only
a. Average pairwise raw agreement	0.84	0.82	0.73
b. Average pairwise Cohen's Kappa	0.42	0.65	0.50
c. Average raw agreement with full-text gold standard		0.65	0.56

Table 2. Annotation agreement measures for full-text, photo-only, and title-only interfaces

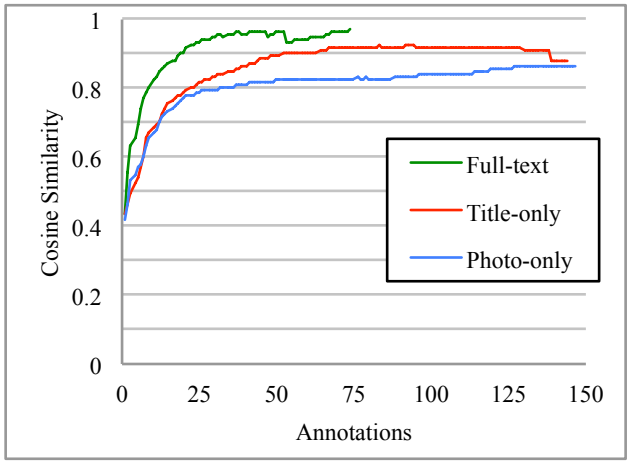

Figure 4. Average cosine similarity by number of annotations

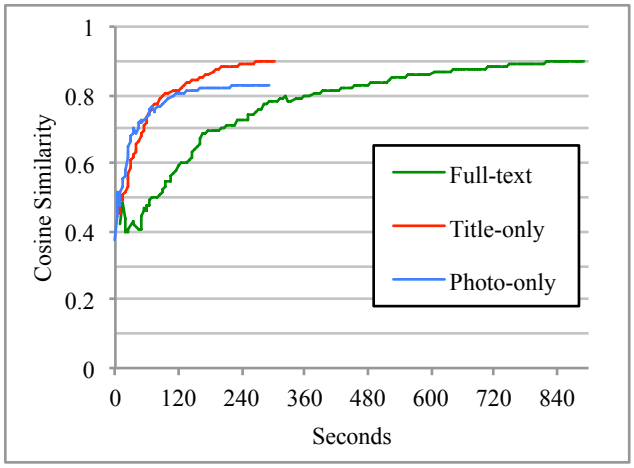

Figure 5. Average cosine similarity by annotation time (300 second limit in title-only and photo-only interfaces)

method of assessing the objective quality of queries across the three interfaces as they evolved over time and with the addition of new annotations. At any given moment in the use of an interface, the current query is represented as a weighted vector of 50 query terms. Using a simple similarity measure (Cosine distance), we can assess the quality of the query by calculating its similarity to a gold-standard query. We constructed three gold-standard term vectors (one for each topic) using the tf-idf weighted terms in each of our gold-standard annotation sets. In this post-hoc analysis, we excluded the initial "boring-story" query from both the evolving queries and the gold-standards, and truncated only the evolving queries to the top 50 most information-laden terms.

Figure 4 plots the average cosine similarity between the evolving queries and its corresponding gold-standard term vector for each interface, for increasing numbers of user annotations (skips included). As expected, the full-text interface requires fewer annotations to approach the gold-standard than either the photo-only or title only condition, although this result is confounded somewhat by the higher degree of overlap between the evolving full-text queries and the gold-standards. Surprisingly, the titles-only condition outperforms the photos-only interface.

This result shows that one is better off judging the full text of a story than a photo or title, given the same number of annotations. However, the results are markedly different when we consider how the annotations evolved over time. Across the three interfaces, the time required to judge an item as relevant, not relevant, or skip varied substantially. On average, judgments were made in the photo-only interface every 2.90 seconds (σ^2 = 9.18 seconds). The title-only interface was only slightly slower (μ = 3.63 seconds, σ^2 = 8.17 seconds). The full-text interface was nearly ten times slower, with a judgment made every 30.09 seconds on average, with huge variation (σ^2 = 885.81 seconds).

To see the impact of this time disparity between interfaces, Figure 5 plots the average cosine similarity between the evolving queries and gold-standard queries over the duration of user interaction. Results show that both the photo-only and title-only conditions outperform the full-text interface when given limited time. Again, we were surprised to find a slight benefit to the title-only interface over the photo-only interface.

Subjective measures

After each use of one of the three interfaces, participants were asked to fill out a six-item questionnaire. The interface evaluation questionnaire asked users to respond to the following questions. The first three items were answered on a ten-point percentage scale (10% being the lowest and 100% being the highest), as follows:

a) What percentage of the stories were about the search topic?

b) What percentage of the stories did you find interesting?

c) What percentage of the stories do you think you were able to annotate accurately?

The next three questions were answered on a five-point Likert scale, with 5 denoting "strongly agree."

d) It is easy to use.

e) I enjoyed using it.

f) If I needed to search for personal stories on weblogs, I would use it again.

After answering these questions, participants were given an opportunity to provide us with open-ended qualitative feedback about the interface they had just used. At the very end, participants were asked one final question:

g) Which of the three interfaces you used (full text, titles only, and photos only) did you prefer?

Table 3 gives the response value for each questionnaire item, averaged over all 18 users of each interface with the exception of item g, where one user did not indicate a

Measure	Full-text	Photo-only	Title-only
a. Percent relevant (10-100)	87.7	64.7	46.4
b. Percent interesting (10-100)	69.2	53.6	38.6
c. Percent annotated accurately (10-100)	95.3	83.1	62.5
d. Ease of use (1-5)	4.58	4.53	4.36
e. Enjoyment (1-5)	3.97	4.14	3.31
f. Would use again (1-5)	4.03	3.64	2.19
g. Preferred interface (fraction expressing preference)	12/17	5/17	0/17

Table 3. Subjective measures from participant questionnaire

preference. The full-text interface scored the highest in every subjective measure except enjoyment. The title-only interface scored the lowest in every subjective measure.

If we examine interface preference (g) across different queries, some interesting complexity emerges. Participants annotating sailing stories expressed a 66.7% preference for the full text interface; for car accidents the preference for full-text was only 40%, and for doctor's visits it was 100%. These data suggest that some topics are more suited than others for a particular interface, and this affects how users feel about the interface overall.

Table 3 shows a clear rejection of the titles-only interface, and some of the open-ended, qualitative responses help to illuminate the reasons for this. Users indicated that the titles-only interface displayed "too much vague and irrelevant information" and that "just a title is not enough to determine" relevance. These comments agree well with the quantitative data in several ways. Users of the title interface skip the greatest number of items (19% of titles were skipped, as opposed to 8% of photos and 4% of full text stories). Also striking is the low level of annotation confidence reported by participants (c). It appears to be important to users that they have enough information to be reasonably confident that they are making the "right" decision when they annotate. Absent an adequate level of confidence they skip stories more often and feel less comfortable using the interface.

A final point brought out by the qualitative feedback is that users are sensitive to the tradeoff between speed and accuracy. Participants who preferred the photos interface reported enjoying the greater speed with which they were able to annotate content. Those who preferred the full text interface reported feeling satisfied because even though each annotation took more time, they were sure they had all the available information and could make an accurate judgment. So while there does seem to be a qualitative, subjective, enjoyment-based aspect to interface preference, it is also clear that the ability to provide accurate responses was crucial to interface preference. Although users on average enjoyed using the photos interface the most (e), they felt most confident about their annotations in the full-text interface (c), and preferred the full-text interface overall (g).

DISCUSSION

There are several dichotomies seen in these results. Participants prefer the full-text interface, but it is more efficient to use the title-only or photo-only interface. Participants are more confident and more accurate in their annotations in the photo-only interface over the title-only interface, but the title-only interface is more efficient overall. Participants find the photo-only interface most enjoyable, but much prefer the full-text interface.

We conclude that the experience of reading personal stories is more rewarding than the labor of annotating photos or titles, particularly when users are not given enough information to be confident of their judgments. When the accuracy of individual annotations is the paramount concern, the full text interface is the best choice. However, it takes ten times longer to make a relevance judgment based on reading the full-text of a story. When time is a limiting factor, users are far better off training a topic-detection system by annotating titles or photos, and would much prefer to annotate photos.

However, we estimated that only 19.5% of stories included a photo when they first appeared on the web. In the corpus we used in this study, a smaller fraction of stories (13%) still available on the web also had photographs that were also still available. As a consequence, users of the photo-only interface were making relevance judgments that were, on average, further away from the current query than those in the title-only interface. Inaccurate annotation of photos (false positives) slowed the progress of the aggregate query vector toward the gold-standard. This is reflected in Figure 5, particularly in the latter half of the use of the photo-only interface, when items without annotations would be furthest down in the results list. Titles, in contrast, are included on nearly every weblog post. Although users were particularly bad at judging the relevance of titles (only 56% correct), these inaccuracies were outweighed by the volume of

annotations that could be collected on items close to the current query vector.

The surprising success of the title-only interface in this evaluation suggests that a hybrid approach would be best when efficiency is the primary concern. Users could be provided with the titles of weblog posts along with a photo from the story in cases where one is available (13% of the time). Our results indicate that the addition of photos would both make the task more enjoyable and increase annotation accuracy.

The overall utility of the photo-only interface, balancing efficiency and enjoyment, might be best exploited in innovative new user interfaces that are divorced from language altogether. Using a photo-only interface, users need not even be fluent in the language of the authors in order to train a high-quality topic model, e.g. an English speaker could just as easily use this interface to identify weblog stories about a given topic written in Spanish, Japanese, or Arabic. While these users may not be able fully to enjoy the stories of the target language, such a tool might be interesting as a way to see, through pictures, how things are done across the world's language barriers.

ACKNOWLEDGMENTS

The authors would like to acknowledge the important contributions and suggestions of Sinhwa Kang, Kenji Sagae, Don Metzler, and Anton Leuski during the course of this research project. The projects or efforts depicted were or are sponsored by the U. S. Army. The content or information presented does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.

REFERENCES

1. Amershi, S., Fogarty, J., Kapoor, A., and Tan, D. (2011) Effective End-User Interaction with Machine Learning. In Proceedings of the 25th AAAI Conference on Artificial Intelligence, August 7-11, San Francisco, CA.
2. Burton, K., Java, A., and Soboroff, I. (2009) The ICWSM 2009 Spinn3r Dataset. In Proceedings of the Third Annual Conference on Weblogs and Social Media (ICWSM 2009), May, San Jose, CA.
3. Chafen, R. (1987) Snapshot Versions of Life. Bowling Green, OH: Bowling Green State University Popular Press.
4. Dunlop, M. and van Rijsbergen, C. (1993) Hypermedia and free text retrieval. Information Processing and Management 29(3):287-298.
5. Fiscus, J. and Doddington, G. (2002) Topic Detection and Tracking. Norwell, MA: Kluwer Academic Publishers.
6. Gordon, A., Bejan, C., and Sagae, K. (2011) Commonsense Causal Reasoning Using Millions of Personal Stories. In Proceedings of the 25th AAAI Conference on Artificial Intelligence, August 7-11, San Francisco, CA.
7. Gordon, A. and Swanson, R. (2009) Identifying Personal Stories in Millions of Weblog Entries. In Proceedings of the Third Annual Conference on Weblogs and Social Media (ICWSM 2009), Data Challenge Workshop, May, San Jose, CA.
8. Gordon, A. and Swanson. R. (2008) StoryUpgrade: Finding Stories in Internet Weblogs. Proceedings of the Second Annual Conference on Weblogs and Social Media (ICWSM 2008), March 31-April 2, Seattle, WA.
9. Koren, Y., Bell, R., and Volinsky, C. (2009) Matrix Factorization Techniques for Recommender Systems. IEEE Computer 42(8):30-39.
10. Ounis, I., Lioma, C., Macdonald, C., and Plachouras, V. (2007) Research Directions in Terrier: a Search Engine for Advanced Retrieval on the Web. Upgrade 7(1):49-56.
11. Rocchio, J. (1971) Relevance Feedback in Information Retrieval. In Salton: The SMART Retrieval System: Experiments in Automatic Document Processing. Prentice Hall, 313-323.
12. Shneiderman, B. (1998) Designing the User Interface. Addison, Wesley, and Longman.
13. Stefik, M. and Good, L. (2011) The News that Matters to You: Design and Deployment of a Personalized News Service. Proceedings of the 23rd Innovative Applications of Artificial Intelligence Conference, August 7-11, San Francisco, CA.
14. Villa, R., Halvey, M., Joho, H., Hanna, D., and Jose, J. (2010) Can an intermediary collection help users search image databases without annotations? Proceedings of the 10th Annual Joint Conference on Digital Libraries (JCDL-10).
15. Zha, Z., Yang, L., Mei, T., Wang, M., Wang, Z., Chua, T., and Hua, X. (2010) Visual Query Suggestion: Towards Capturing User Intent in Internet Image Search. ACM Transactions on Multimedia Computing, Communications and Applications, Vol. 6, No. 3, Article 13.