StoryUpgrade: Finding Stories in Internet Weblogs

Andrew S. Gordon and Reid Swanson

The Institute for Creative Technologies
The University of Southern California
13274 Fiji Way, Marina del Rey, CA 90292 USA
gordon@ict.usc.edu, swansonr@ict.usc.edu

Abstract

The phenomenal rise of Internet weblogging has created new opportunities for people to tell personal stories of their life experience, and the potential to share these stories with those who can most benefit from reading them. One barrier to this new mode of storytelling is the lack of accessibility; existing Internet search tools are not tailored to the unique characteristics of this textual genre. In this paper we describe our efforts to develop a search engine specifically for the stories that appear in Internet weblogs, called StoryUpgrade. This application utilizes statistical text classification technologies to separate story content from other text in weblog entries, and facilitates searches for stories that are related to particular activities of interest.

Stories in Internet Weblogs

In the past few years, the phenomenal rise of Internet weblogging has created new opportunities for computer-supported storytelling applications. The diary-like nature of weblogs lends itself to a casual form of public storytelling, where people tell stories across the full breadth of their life experiences. With the estimated number of weblogs exceeding 70 million in March 2007 (Technorati, 2007), the ratio of writers to readers is relatively low compared to other methods of publishing, and readership is largely determined by the structure of social networks. Sadly, the stories of many webloggers will have few or no readers whatsoever, despite the potential value of these narratives to people outside their particular online community.

To connect weblog authors with readers, commercial Internet search providers have applied standard website search technologies to weblogs, e.g. Google Blog Search at http://blogsearch.google.com and Technorati's service at http://technorati.com. However, this approach is less than ideal for users who are looking specifically for stories, for a number of reasons. First, only a fraction of weblog content consists of narratives of people's experiences, so users' queries will also match with the text of other sorts of material that commonly appears in weblogs, e.g. news quotations, political commentary, and to-do lists. Second, the link-analysis techniques that have proven successful for determining relevance for general web search results do not necessarily apply to weblog entries; the story that nobody else has read in a weblog entry may be the most relevant to a users interest. Third, users story-retrieval interests are difficult to express in a handful of search terms, and increasing the precision of search results with longer queries quickly deteriorates recall performance.

The StoryUpgrade Application

StoryUpgrade is an application that combines two technologies in order to provide an effective means of finding stories in Internet weblogs. First, we incorporate technology for separating stories from other textual genres that appear in Internet weblogs. As noted by Gordon et al. (2007), between 14% and 17% of the text in weblogs consists of story content. In order to selectively retrieve only story content, the StoryUpgrade system incorporates statistical text classification technologies to identify story segments in weblogs. Second, we incorporate technology for matching the text of extracted stories to paragraph-sized descriptions of activities, provided by users of the StoryUpgrade application as search queries. Users are instructed to describe their retrieval needs in the form of a "boring story" about an activity – describing the sequence of events that might be expected in (more interesting) stories about the activity. For example, a user who is searching for stories about "flying on a passenger airplane" might provide the following paragraph of text as a query.

> *I arrived at the airport and checked my luggage at the counter. I went through security screening, and found my way to the gate. I boarded the plane and found my seat. I watched the in-flight movie, and fell asleep. I looked out the window, and eventually the plane landed at the airport. We taxied for a while, then the doors opened at the gate. I exited the plane and found my way to the baggage claim area. I picked up my luggage and went outside to catch a taxi.*

By describing their retrieval needs in this manner (written in the first-person past-tense), the StoryUpgrade application can effectively locate segments of text within weblogs that share much of the same vocabulary, e.g. with verbs that are similarly inflected and nouns that are either singular or plural as appropriate to the activity context.

We developed two versions of the StoryUpgrade application, each incorporating different approaches to story segmentation and activity retrieval.

StoryUpgrade Version 1

The first version of the StoryUpgrade application was designed to collect stories from millions of Internet weblog entries, and to provide a web-based search utility for querying this corpus using a contemporary information retrieval utility. To create a corpus of stories, we first acquired the web addresses of nearly 400 thousand weblogs from one of the major commercial search engines, Technorati.com, using the application programming interface that they provide to third-party developers. Then each of the weblog entries associated with these addresses were downloaded individually and analyzed for story content. To identify story segments, we utilized a version of the story extraction algorithm developed by Gordon & Ganesan (2005), modified as described by Gordon et al. (2007). This algorithm applies a binary (story/non-story) Naïve Bayes classifier to overlapping spans of 50 words in the input text. Word-level classification confidence values are then smoothed using a simple mean-average function, and consecutive words assigned to the story class are extracted as story segments. Over a period of 373 days, this system downloaded and processed 39.9% of the identified weblogs (3.4 million entries). On average, 1.32 distinct segments in each weblog entry were labeled as story text, resulting in a corpus of 4.5 million extracted story segments consisting of 1.06 billion words. This version of StoryUpgrade stored each segment of extracted story content in a repository that was indexed for retrieval using a high-performance information retrieval engine. For this purpose, we incorporated the Indri search engine developed at the University of Massachusetts and Carnegie Mellon University (Strohman et al., 2005). Access to this corpus was then provided via a web-based user interface.

StoryUpgrade Version 2

Despite the considerable size of the story corpus created by version 1 of the StoryUpgrade application, it provided access to less than 1% of the weblogs on the Internet. Version 2 of the StoryUpgrade application was developed to solve this problem by relying more directly on the services provided by major commercial weblog search engines (Google Blog Search). Whereas version 1 of StoryUpgrade first separates stories from non-story text before utilizing a search utility, version 2 reverses this procedure. After submitting an activity query to the StoryUpgrade web interface, each of the sentences in the "boring story" are individually submitted as Google Blog Search queries. The top 100 results for each sentence are then combined, and each of the corresponding weblog entries is downloaded and analyzed to extract story content from these results. To identify story content StoryUpgrade version 2 uses the technique described by Gordon et al. (2007). This technique applies a binary (story/non-story) Support Vector Machine classifier to overlapping segments of three sentences. Sentence-level classification confidence values are then smoothed using a Guassian function, and consecutive sentences assigned to the story class are extracted as story segments. StoryUpgrade then orders the extracted stories based on their relevance to the original "boring story" query provided by the user. For this purpose we employ a standard cosine similarity distance metric based on n-gram features of the query and the extracted text, weighed using TF-IDF values.

Related work

Owsley et al. (2006) describe a system called Buzz, a digital theater installation where animated virtual characters delivered emotionally-evocative stories extracted from weblogs. In this work, candidate weblog entries are retrieved from a commercial weblog search engine, using the day's most popular Internet searches and a list of controversial topics as weblog search queries. Although the goals of this research are much different than our own, the Buzz application incorporates technologies and approaches that are similar to the StoryUpgrade application, especially version 2: both use the strategy of filtering the results of commercial weblog search engines. Our work differs in the emphasis on achieving both high precision and recall performance for the identification of stories in weblog text, and the use of activity-based queries to allow users to find stories that are relevant to their individual interests.

Acknowledgments

The project or effort described here has been sponsored by the U.S. Army Research, Development, and Engineering Command (RDECOM). Statements and opinions expressed do not necessarily reflect the position or the policy of the United States Government, and no official endorsement should be inferred.

References

Gordon, A & Ganesan, K. (2005) Automated Story Extraction From Conversational Speech. 3rd Intl. Conference on Knowledge Capture, Banff, Canada.

Gordon, A., Cao, Q., & Swanson, R. (2007) Automated Story Capture From Internet Weblogs. 4th Intl. Conference on Knowledge Capture, Whistler, BC.

Owsley, S., Hammond, K., Shamma, D., Sood S. (2006) Buzz: Telling Compelling Stories. ACM Multimedia, Interactive Arts program, Santa Barbara, CA.

Strohman, T., Metzler, D., Turtle, H., Croft, W. B. (2005) Indri: A language model-based search engine for complex queries. Intl. Conference on Intelligence Analysis, McLean, VA.

Technorati (2007) State of the Blogosphere / State of the Live Web. http://www.sifry.com/stateoftheliveweb