BlogPulse: Automated Trend Discovery for Weblogs

Natalie S. Glance, Matthew Hurst and Takashi Tomokiyo
Intelliseek Applied Research Center
5001 Baum Boulevard
Pittsburgh, PA
{nglance, mhurst, ttomokiyo}@intelliseek.com

ABSTRACT

Over the past few years, weblogs have emerged as a new communication and publication medium on the Internet. In this paper, we describe the application of data mining, information extraction and NLP algorithms for discovering trends across our subset of approximately 100,000 weblogs. We publish daily lists of key persons, key phrases, and key paragraphs to a public web site, BlogPulse.com. In addition, we maintain a searchable index of weblog entries. On top of the search index, we have implemented trend search, which graphs the normalized trend line over time for a search query and provides a way to estimate the relative buzz of word of mouth for given topics over time.

1. INTRODUCTION

Weblogging has emerged in the past few years as a new grassroots publishing medium. Like electronic mail and the web itself, weblogging has taken off and by some estimates the number of weblogs is doubling every year. At the time of writing, various estimates place the number of active weblogs at about 1.4 million blogs.

A weblog is commonly defined as a web page with a set of dated entries, in reverse chronological order, maintained by its writer via a weblog publishing software tool. Within this generic format, there are several genres of weblog content. There is the online journal, a popular format since the early days of the web. In online journals, people share publicly the daily events of life which in eras not so distant were confined to private notebooks. Online journals enjoy the double appeal of voyeurism (for the reader) and ongoing feedback or even fame (for the writer).

There is also the pundit, a self-declared expert, who publishes updates and analyses of events within his/her domain. In earlier days of the web, pundits published newsletters online or via e-mail. There is the news filter: the voracious reader who filters out gold from dross from up to hundreds of publications (both weblogs and traditional media), often using RSS aggregators, and publishes lists of links to the most interesting new items, with sometimes minimal or no accompanying commentary. There is the writer/artist, who uses weblogging to self-publish stories, poems, art, music, photographs. And, of course, spam has invaded this new form of publication in the form of the spam weblog, an advertisement pretending to be a weblog, and via spam weblog comments.[1]

There are also many variations on the weblog format. One common variation is the collaboratively published weblog where many writers contribute entries to the same weblog. Another variation is the weblog whose entries allow comments from readers, potentially turning each entry into a discussion. Yet another set of variations is the means of publishing, be it via an HTML interface, via e-mail, via a wireless PDA or cell phone (moblogging) or camera cell phone (photoblogging).

Common to most weblogs is the blogroll, a set of links in a weblog's template that list the author's favorite bloggers. Several weblogging software tools also include tools for automatically counting and displaying how many other bloggers have linked to any given entry in a weblog (these are known as trackbacks). Webloggers also have easy access to the list of referrers to their entries, those sites from which users followed a link to visit the weblog. There is anectodal evidence that bloggers check these referral lists regularly: some weblog entries start with a phrase like "Perusing my referral log...".[2]

The cross-linking that takes place between blogs, through blogrolls, explicit linking, trackbacks, and referrals has helped create a strong sense of community in the weblogging world. Already, there is work underway to understand the dynamics of the weblogging network, much of which springs from bloggers themselves.

Various weblog search engines such as Daypop[3] and blogdex[4] maintain lists of the most popular bloggers. Others, like Technorati[5] and BlogStreet[6] allow a blogger to identify his/her blog within a neighborhood of other bloggers. A large graph of weblogs with edges defined by cross-linking is available freely for download at myelin.co.nz. Weblog census projects have begun, notably the NITLE Blog Census[7] (last updated 6/2003) and the Perseus Blog Survey.[8] Blogcount is a weblog devoted to tracking weblogs by volume.[9] Both weblog search engines and weblog census sites gather lists of weblogs from online lists of weblogs. The main source for discovering new weblogs are a few weblog update sites which list recently updated weblogs, most notably http://fresh.blogrolling.com/, which is a compilation of weblog update lists. Weblogs can be configured to ping the update sites when a new entry is added. Interestingly, as yet, there is no comprehensive directory of weblogs, although several small directories exist (e.g., Blogstreet, eatonweb portal.[10])

The existence of top blogger lists and the ability to gauge the

[1] http://kalsey.com/2003/11/comment_spam_manifesto

WWW2004, May 17–22, 2004, New York, NY USA.
ACM xxx.xxx.

[2] http://www.offthekuff.com/mt/archives/001259.html
[3] http://www.daypop.com
[4] http://www.blogdex.com
[5] http://www.technorati.com
[6] http://www.blogstreet.com
[7] http://www.blogcensus.net/
[8] http://www.perseus.com/blogsurvey/
[9] http://dijest.com/bc/
[10] http://portal.eatonweb.com

popularity of one's own weblog also contributes to the introspectiveness of the weblog community. For example, in February 2003, there was a lively exchange concerning the power law distribution and weblogs, initiated by an article by Clay Shirky [10]. His article discussed the power law distribution of LiveJournal weblogs arranged in rank order by number of inbound links and draws on work from the academic community on studying power law distributions in online networks [1, 9]. A side-effect of the power law distribution is that a minority of the blogs in his sample accounts for a majority of inbound links. His article sparked significant commentary. By the next day, Jason Kottke had already performed an analysis of the top 100 most linked to weblogs on Technorati and likewise found a power law distribution [7]. Many bloggers were dismayed by the result and dicussed means to "break" the power law,[11] falling prey to the common fallacy that emergent dynamics should behave ethically.

The first academic research on the weblogging community appeared in last year's WWW conference featuring analysis displaying the bursty evolution of blogspace and the appearance of distinct subcommunities around common themes [Kumar03]. Since then, the first bloggers' conference has also occured, BloggerCon 2003,[12] allowing bloggers to meet face-to-face, united both by technological interests and already established personal relationships.

The weblogging microcosm has evolved into a distinct form, that is, into a community of publishers. The strong sense of community amongst bloggers distinguishes weblogs from the various forms of online publications such as online journals, 'zines and newsletter that flourished in the early days of the web and from traditional media such as newspapers, magazines and television. The use of weblogs primarily for publishing, as opposed to discussion, differentiates blogs from other online community forums, such as Usenet newsgroups and message boards. Often referred to as the blogsphere, the network of bloggers is a thriving ecosystem, with its own internally driven dynamics.

However, the blogsphere does not in any sense exist in isolation. Many bloggers track traditional media daily, often via RSS feeds and/or aggregators, and link to newspaper articles, magazine articles, and radio transcripts in their entries.

Likewise, as news on occassion breaks first via a weblog, journalists have begun to pay attention to the blogsphere (despite skeptics like John Markoff, who says that "it's not clear yet whether blogging is anything more than CB radio".[13]) It's no longer uncommon to read a newspaper article citing bloggers, despite the fact that much of the newspaper's audience doesn't yet know what a weblog is.

In addition, weblogs have become the sina qua non of grassroots political movements. Both Howard Dean's and Weseley Clark's bids for the Democratic Presidentional Nomination have been spearheaded via weblogs. Their weblogs serve both as a medium for communication with their supporters and as a means of soliciting campaign contributions. Perhaps this helps account for the unusually large percentage of contributions to both Howard Dean and Weseley Clark's campaign from "small" donors (54% of individual contributions to Howard Dean's campaign were less than $200, as of 10/2003, compared to Bush's 84% from donors contributing more than $1000).[14]

The self-publishing aspect of weblogs, the time-stamped entries, the highly interlinked nature of the blogging community and the significant impact of weblog content on politics, ideas, and culture make them a fascinating subject of study. Although the blogsphere is biased towards a particular cross-section of society, identification of trends within the blogging world is a way to take the pulse of its contributors and of a segment of society.

Aggregate indicators of weblogging behavior are a powerful way to take the pulse of this community. For example, Daypop maintains Top 40 lists of the most frequently cited links and the most frequently cited news articles in weblogs. The link count mechanism is a primitive, but effective form of collaborative filtering, akin to Google's PageRank mechanism. Allconsuming.net tracks the weblog references of books (by mining references to ISBN numbers). Thus, it is a filter over the most talked about books in weblogs over a given time period.

Another aggregate indicator is the evolution of language usage over time: what bloggers talk about and what vocabulary they use. The authors have built a toolkit for analyzing online collections of time-stamped documents in order to automatically detect such trends. In this paper, we describe the application of this toolkit for analyzing collections of weblog entries over time. On the one hand, weblogs have served as an ideal testbed for trying out new machine learning and NLP algorithms. On the other hand, applying methods from our toolkit has permitted us to contribute back to the blogsphere. We maintain a site called BlogPulse.com, which publishes daily lists of key phrases, key people, top links and "blog bites" mined from weblog entries. The site also maintains a public archive, making available the daily lists since BlogPulse's inception in late March 2003. Still in development are a search service and a trend search service which graphs trend lines over time for phrases.

The rest of the paper is organized as follows. First, we describe weblog corpus creation, including harvesting weblog urls, crawling, identifying weblogs and properties of the corpus. Next we discuss the application of several automatic algorithms for discovering trends: phrase finding, people finding and identification of key paragraphs ("blog bites"). Finally, we present blogpulse search, trend graphing and some early thoughts about automatically classifying trends into different categories based on their temporal patterns.

2. WEBLOG CORPUS

2.1 Harvesting weblog URLs

The first step in building a system that mines weblogs is collecting a list of active weblogs. There exists no comprehensive directory of weblogs, and in fact, even the current number of existing weblogs can only be estimated. As described in the Introduction, there are several ongoing weblog census projects, placing the count of active weblogs around 1.4 million and estimating the growth rate at about 100% per year. One of our long-term goals is to automatically classify blogs into a set of categories and publish a directory.

We extracted a seed list of 22,000+ weblogs from the Blogstreet directory of weblogs with RSS feeds[15] in February 2003. We then grew this seed list by automatically extracting weblog urls every three hours from the recently updated weblog list posted by Radio Userland.[16] Starting in early June, we started also extracting weblog urls from BlogRolling.com's list of recently updated weblog list.[17] Because of our choice to obtain weblog URLs primarily

[11] For example, "Repealing the Power-law" on Ross Mayfield's Weblog, http://radio.weblogs.com/0114726/2003/02/06.html

[12] http://blogs.law.harvard.edu/bloggerCon/

[13] http://www.ojr.org/ojr/technology/1066258791.php

[14] http://www.cfinst.org/pr/101703.html

[15] http://www.blogstreet.com/bin/rssdir.pl?char=a

[16] http://www.weblogs.com/

[17] BlogRolling's list at http://fresh.blogrolling.com/ is now comprised of a compilation of feeds from BlogRolling.com, We-

from these two sources, our set of URLs has a bias towards Radio Userland-hosted weblogs and blogspot-hosted weblogs, to the detriment of LiveJournal weblogs.[18] However, LiveJournal does maintain a directory of its weblogs for subscribed users and we could easily add potentially hundreds of thousands of active LiveJournal weblogs. Cursory examinatation of a sample of LiveJournal weblogs indicates that these tend to fall within the genre of on-line journal, with little cross-linking between weblog posts and also infrequent linking to external sources of information.

Each day the growing list of weblog URLs passes through a clean-up process, which follows a set of heuristics to identify duplicates, remove weblog category URLs and weblog archive URLs. Every few weeks, we also purges from the list all weblogs for which there were no posts since the last purge.

We stopped accumulating additional weblog URLs in June 2003 when we hit a total of about 100,000 active weblogs, because we were reaching an upper bound on the number of weblogs that we could politely crawl within 12 hours on one server. In addition, given that our list includes the most oft-cited blogs, we felt that the set of 100,000 represented a suitably representative cross-section of the discussion occurring in the blogsphere (apart from a potential bias away from the online journal form of weblogs).

Furthermore, it is possible that our list contains a large majority of the very active weblogs (those posting multiple times per week). We hypothesize that this is the case because search over our index of weblog entries returns a similar number of hits within a given time frame as the same search performed on Technorati, which claims to index over 1.1 million weblogs. However, in order to maintain the freshness of our list, we will need to resume periodically harvesting new weblogs and purging inactive ones.

2.2 Crawling

We crawl the entire list of 100,000+ weblogs daily using the Intelliseek Spider.[19] The BlogPulse parametrization of the spider instructs it to follow redirects, but to not follow any links; thus, we crawl only the home page of the weblog. Our spider is also tuned to be as polite as possible under the constraint that processing complete within 12 hours.

In order to identify weblog entries, we have implemented a simple differencing algorithm which compares the current day's crawl of a weblog with the previous day's crawl of a weblog. The difference is time-stamped with the date of the current crawl. While this process successfully identifies new weblog entries, it produces systematic side-effects. First of all, the day of the crawl may differ from the day of the entry by +/- one day. Secondly, modifications to a weblog's template will be included in the extracted entry, for example, new links added to a sidebar. Third, modifications to an earlier entry (change in the number of comments, edits to the entry) will appear along with the day's entry. Finally, while some blogging tools allow users to blog an entry with a past or future date, the differencing process will attribute the current date to this entry instead. It is debatable whether or not these side-effects should be considered errors. From the point of view of extracting trends, all information added on a particular date, be it new links, new comments, or modifications to previous posts, is relevant to the current date.

The main disadvantage of using a differencing algorithm for identifying weblog entries is that we must crawl each weblog entry daily in order to correctly time-stamp the day's entries. In order to crawl a larger set of weblogs politely, the crawl must be spread out over several days. This entails implementing a different method for identifying weblog entries. Thus, our future plans for BlogPulse include building wrappers for the principle blogging software tools to automatically extract weblog entries from weblogs. We plan to use an in-house software tool named Wrapster [6] to automatically learn rules for extracting and associating into one entry the following fields (when present): the title of the entry, the date, the author, the content, the number of comments, the archived link, the trackback link. At least one weblog search engine uses wrappers, namely Waypath,[20] as documented in their weblog.[21]

2.3 Corpus creation and indexing

BlogPulse creates a daily corpus from the collection of weblog entries using our in-house generic toolkit for corpus creation, indexing, phrase finding, trending and data mining.

Our toolkit, Analyst Workbench, constructs a corpus from a collection of documents, using a set of transformations over the documents. Each document is represented in the corpus by a sequence of tokens. Associated with each document in the corpus is a set of *annotations*. Annotations can be used to provide meta information about elements of a document such as part-of-speech tags, sentence boundaries, paragraph boundaries, and case information. The annotations are currently token-level annotations, mapping each token to its annotation value.

For HTML documents, annotations can also include attribute values of the parent node of the text, for example the value of the HREF attribute of the $\langle A \rangle$ node encompassing the anchor text. In this way, the link information in an HTML document can be associated with tokens using annotations.

A transform step can also in principle consist of running an entity extractor over the token sequence, for example a person name extractor. The result of the transform step is to annotate each token of an extracted person name as being part of a person name.

The annotation mechanism allows the corpus to exist as a simple flat sequence of tokens, with parallel layers of arbitrarily complex information. The set of annotations produced as part of the corpus are subsequently used by our data mining algorithms.

In our implementation, we define a corpus, C, as a tuple $\langle D, t \rangle$ where D is a set of documents, and t is a function which takes a document and produces a sequence of tokens, together with appropriate annotations. This *transformation* function deals with a number of issues important to supporting the corpus framework and to trend discovery.

The set of transforms used to build the daily weblog corpus are:

1. Tokenize text nodes of HTML; annotate tokens with segment boundaries, paragraph boundaries, and links;

2. Apply a trained classifier to filter out non-English language weblog entries;

3. Annotate token list with case information;

4. Lower case the token sequence.

As we have not yet integrated our entity extraction algorithms into the toolkit, person name extraction occurs in parallel over the collection of weblog entries.

In addition to the corpus, an inverted index of the current day's collection of weblog entries is created. Data mining algorithms in our toolkit query the inverted index in order to: identify the set of contexts of a phrase; find the number of occurrences of a phrase

blogs.com and Blogger.com.

[18] http://www.livejournal.com

[19] http://www.intelliseek.com/partnerdiscovery.asp

[20] http://www.waypath.com

[21] http://www.waypath.com/mt/archives/000072.html

appears in the corpus; and find the number of documents containing the phrase.

In parallel with daily corpus and index building, each weblog entry is also indexed into a separate Lucene[22] index encompassing a historic archive of all weblog entries crawled by BlogPulse. We use Lucene in order to have a highly optimized efficient historical index. However, the Lucene index does not provide the functionality required by our phrase finding algorithms described in the next section.

3. TREND DISCOVERY

BlogPulse executes a set of data mining algorithms over the day's collection of weblog entries in order to discover aggregate trends characterizing the past day's publications to the blogsphere. BlogPulse identifies the day's most popular links, in imitation of weblog trend mining ground-breakers Daypop and blogdex. Novel to BlogPulse are key phrase finding, key person finding and key paragraph detection (our so-called "blog bites"). In this section, we describe the algorithms currently used for each.

3.1 Phrase finding

One original impetus for BlogPulse was to test the various phrase finding algorithms from Analyst Workbench against weblog data. Weblogs provide an ideal testbed for a number of reasons. First of all, every entry is time-stamped. Secondly, evaluating the quality of phrase finding is relatively straightforward. Precision can be measured as the proportion of top phrases that are informative with respect to current discussions in the blogsphere. Every day's set of key phrases provides qualitative feedback on the performance of our phrase finding algorithms. This feedback has motivated the incremental improvements to our key phrase finder. The improvements are evidenced by fewer "garbage" phrases and fewer partial phrases over time. The daily lists also provide an informal indication of the key phrase finder's recall (i.e., the percentage of the day's most informative phrases discovered by phrase finding).

Analyst Workbench (AW) provides a toolkit of phrase finding algorithms which can be sequenced to create a composite phrase finder tuned for a given application. The phrase finders in AW have two basic forms. The first type is a function that takes a corpus and produces a list of phrases (a *phrase list*). The second type, known as a seeded phrase finder, takes a corpus and a seed phrase list and produces a new phrase list. The phrase lists that are produced by phrase finders may be ordered and may have an associated *score*.

Seeded phrase finders may be implemented to act as filters and rescorers, or as methods to create new phrases by extending phrases in the seed phrase list. The phrase finders may be pipelined together by using the result of one as the seed for another.

The key phrase finder for BlogPulse pipelines together the following phrase finding steps:

1. KEYBIGRAMFINDER, which takes a corpus and term frequency statistics from a background data set and extracts informative bigrams. The algorithm combines a measure of *informativeness* and a measure of *phraseness* for a bigram into a single unified score to produce a ranked list of key bigrams [14].

2. TOPNFILTER, which takes a phrase list and returns the first N phrases (we set N to 200).

3. SUFFIXTREEAPRIORI, a seeded phrase finder which finds all highly-frequency phrases that contain one of the given seed phrases and pass a contingency test. It is based on pruning heuristics similar to those used in the APRIORI algorithm [2] for finding frequent sets and by [8] for term extraction. The score output for each phrase is simply its frequency in the corpus. The APRIORI family of phrase finders all take two parameters: the minimum frequency of occurrence of a phrase; and the maximum number of phrases to be output (for key phrase finding, we use 3 and 500, respectively).

4. CONSTITUENTFILTER, a seeded phrase finder that returns only those phrases which it deems to be of a certain category (e.g. noun phrase, verb phrase, etc.). It is implemented using a simple set of models of constituency boundaries for the basic phrase types as well as a morphological analyzer to enable hypothesis generation and testing depending on the category.

5. PHRASEBURSTRERANKER, a seeded phrase finder that reranks phrases in decreasing order of their "burstiness." The burstiness measure is defined as the ratio of the frequency of occurrence of the phrase on the current day as compared to its average frequency over the past two weeks. Since the corpus only contains the day's weblog entries, the PHRASEBURSTRERANKER queries the Lucene index to retrieve frequency counts. This measure of burstiness captures sudden spikes in phrase usage as well as gradual increases.

	Key Phrase	Key Person
1	antoine walker	Antoine Walker
2	million songs	Paul Burrell
3	sexual identity	Nathaniel Heatwole
4	princess diana	Dr. Eric Vilain
5	de niro	Lieberman, Clark
6	john allen muhammad	Chris Mills
7	box cutters	Barbara Bush
8	million songs have been purchased	Tony Delk
9	joe millionaire	Judge Deborah Servitto
10	david blaine	Jiri Welsch
11	white house official	John Allen Muhammad
12	apple press release	Jonathan Chait
13	new joe millionaire	Seymour Hersh
14	review of al franken	Mr Bailey
15	new mobile phone	John Gaeta
16	gregg easterbrook	David Blaine
17	million songs sold	James Kellaris
18	voters in the united states	Robert De Niro
19	saddam hussein and osama bin laden	Alan Colmes
20	national security strategy	Evan Coyne Maloney

Table 1: BlogPulse Key Phrases and Key People for 10/21/03

The effect of this composite phrase finder is to identify phrases that are both informative with respect to a background model of term frequencies in weblog data and that exhibit a bursty trend line.

The left column of Table 1 shows the top 20 key phrases for Oct. 21, 2003.[23] On the website, the key phrases link to pages providing several sample contexts for each phrase. Daypop provides a related service called "word bursts," although the method used for mining these is not documented.[24]

[22] http://jakarta.apache.org/lucene/docs/index.html

[23] http://www.blogpulse.com/03_10_21/keyPhrases.html

[24] http://www.daypop.com/burst/

Upon cursory inspection, all of the key phrases for Oct. 21 are significant and representative of current events under discussion in weblogs the previous day. Several phrases are related to Apple's release of iTunes for the PC.[25] The phrase "box cutters" refers to an implement smuggled by a student onboard a plane to perform an illegal test of airline security.

The main drawback of the current key phrase finder is redundancy. To address this deficiency, we conducted experiments to cluster the key phrases into "topics."

3.2 Topic mining

Following work in the area of topic detection and tracking [3, 15, 13], we define a a topic as a set of significant phrases that are clustered together based on similarity. Intuitively, we expect the cluster of phrases "Arnold Schwarzenegger", California, "recall election", "Gray Davis" to represent a topic. This leaves open the question of how to label the topic. In the literature, clusters representing topics are generally hand-labeled. However, the set of representative phrases is insufficient, as it shifts over time. For example, the day of the election, the same topic might better be represented by cluster of phrases "statewide exit polls", "Schwarzenegger wins", "Gov. Gray Davis". Side-stepping the issue of labeling a topic, we instead use each cluster of phrases to identify a key paragraph in a weblog entry that uses the largest majority of the phrases in the cluster. These key paragraphs, BlogPulse's so-called BlogBites, provide context for, and bring to life, the clusters of key phrases.

We implement topic finding as a set of several steps. First, we apply the key phrase finder described in section 3.1 but parametrized to yield a more complete list of key phrases. The parameter N in TopNFilter filter is set to 1000 instead of 200. The minimum phrase occurrence frequency and maximum number of output phrases in the SuffixTreeAPriori phrase extender are set to 3 and 5000, respectively. The ConstituentFilter and the PhraseBurstReranker then pare the set of phrases down to a much smaller number (only phrases with a burstiness $>$ 1.2 are retained). Typically, the key phrase finder yields $\sim$150 phrases with parametrization (200, 3, 500) and $\sim$400 phrases with parametrization (1000, 3, 5000).

Next, the set of phrases are clustered using a simple iterative process. Initially, each phrase seeds a topic. We iterate over the set of topics, attempting to merge together similar topics. Each time a merge succeeds, we re-start the iterative process. This continues until all attempts to merge pairs of topics fail. One side-effect of this simple clustering process is that each phrase can belong to only one topic.

Two topics T_1 and T_2 are deemed to be similar if any phrase P_1 in T_1 is similar to any phrase P_2 in T_2. Two phrases are deemed similar if the cosine of their occurrence vectors is greater than a parametrized threshold. The occurrence vector for a phrase P_i is a binary vector $V_i = \{d_{i1}, d_{i2}, ..., d_{im}\}$ of length M, where M is the number of weblog entries. If phrase P_i occurs in the j^{th} weblog entry, then $d_{ij} = 1$; otherwise $d_{ij} = 0$. Consequently, the cosine distance of the occcurrence vectors measures the degree of co-occurrence in weblog entries of any two given phrases. The hypothesis behind this choice of similarity measure is that commonly co-occurring phrases are likely to be semantically related.

Our approach to identifying topics is similar to that used in in [12, 11], where significant features are first extracted using the χ^2 statistic and then grouped together according to a measure based on co-occurrence.

Here are a few examples of blog bites identified by this topic finding process on 10/21/2003 (key phrases are in italics):[26]

- "*PRINCESS DIANA* claimed there was a plot to kill her in a *car crash* in a handwritten letter only 10 months before she died. She gave it to her *butler Paul Burrell* with orders that he should keep it as 'insurance' for the future ... - Mirror . co . uk"[27]
- "FBI officials are currently investigating an extraordinary case of a *college student* who smuggled *box cutters* aboard two airplanes as an 'act of civil disobedience with the aim of improving public safety for the air-traveling public,' ...."[28]
- "Boston traded *Antoine Walker* to the Dallas Mavericks for *Raef Lafrentz* in a five player deal. (AP Photo)"[29]

BlogBites are appealing because they bring order to the key phrases discovered in phrase finding. The key paragraphs succintly provide context for clusters of phrases. On the other hand, it is disappointing, but not surprising, to see that many of the blog bites are quotes from the press.

3.3 Key persons

Another approach for mining trends out of a large collection of data is to derive structured entities from the unstructured information source. Our toolkit includes a set of entity extractors, namely, a person name extractor, a postal address extractor, an e-mail address extractor a telephone number extractor, a compay name extractor, and more. Several of these extractors are built using finite state transducers using a high level scripting language, WXL. The person name extractor was implemented by using a WXL script to generate person name candidates which are then filtered using a trained classifier [5]. The person name extractor has been evaluated as performing at 90% precision at 90% recall.

By running the entity extractors over the collection of weblog entries, it is possible to discover trends associated with a given semantic category. We publish a daily list on BlogPulse of the burstiest names extracted from the day's collection of weblog entries. The daily list is produced as follows: first the person name extractor is run over the set of weblog entries obtained using the differencing algorithm described earlier. Then, the set of names are ranked using the PhraseBurstReranker. The list of Key Persons for Oct. 21, 2003 is displayed in the right column of Table 1.

Key persons #1 and #2, *Antoine Walker* and *Paul Burrell* also appeared in BlogBites that day, excerpted above. Key person #3 is the name of the student referred to in second BlogBite above. The BlogPulse key person list also provides context for each key person name: a hyperlinked quote from a weblog that contains the name supplemented by a list of weblogs also citing the name that day. Further down the list, we find that key person #31, *Alan Alda*, has been hospitalized.[30]

In an earlier implementation of the key person finder, extracted person names were ranked simply in order of decreasing frequency of occurrence. The result of frequency-based ordering was that the same set core set of names appeared every day. The choice to display bursty names instead was purely subjective on our part (we tired of seeing "President Bush" at the top of the list day after day); both orderings have informative value.

[25] http://www.apple.com/itunes, also the top link for the day

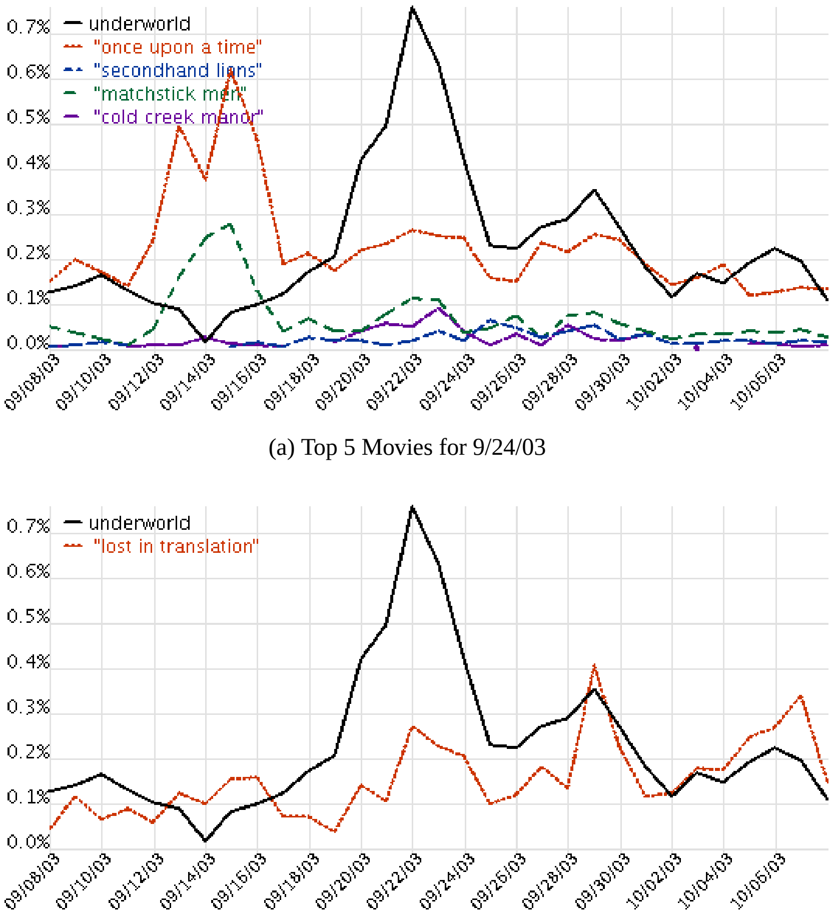
[26] http://www.blogpulse.com/03_10_21/keyTopics.html

[27] From http://kysor.blogspot.com/

[28] From http://www.tompaine.com/blog.cfm

[29] From http://www.benmaller.com/

[30] http://www.blogpulse.com/03_10_21/keyPeople.html



(a) Top 5 Movies for 9/24/03

(b) Sleeper hit: “Lost in Translation”

Figure 1: Trend graphs for movie mentions in weblogs

4. SEARCH & TREND GRAPHING

4.1 Weblog Search

In parallel with crawling and differencing, the BlogPulse production system incrementally updates a Lucene index, which functions as a searchable historic interface to our collection of weblog entries. Although not yet publically accessible, the index accepts boolean combinations of phrases, dates, links and citing URLs. Similar to Technorati, it is thus possible to search for all entries citing a given url. It is also possible to do prefix search on a link, to find, for example, all weblog entries citing a URL that begins with *http://www.nytimes.com*.

Internally, Intelliseek uses BlogPulse search as a data source for performing semi-automated online market research, along with Usenet and message boards. Intelliseek’s approach is to tap into word of mouth on the Internet to inform marketing strategies of customer firms. The utility of this approach has recently been empirically validated [4].

In general, we find about an order of magnitude fewer mentions of products in weblog entries than on online message boards. For example, in the month of September 2003, a search over weblog entries for *xbox* yields 855 hits, while a search over message board posts yields 8777 hits. In addition, weblog citations of products tend to be more anectdotal. Here is a mention of the Xbox video console system from a weblog entry:

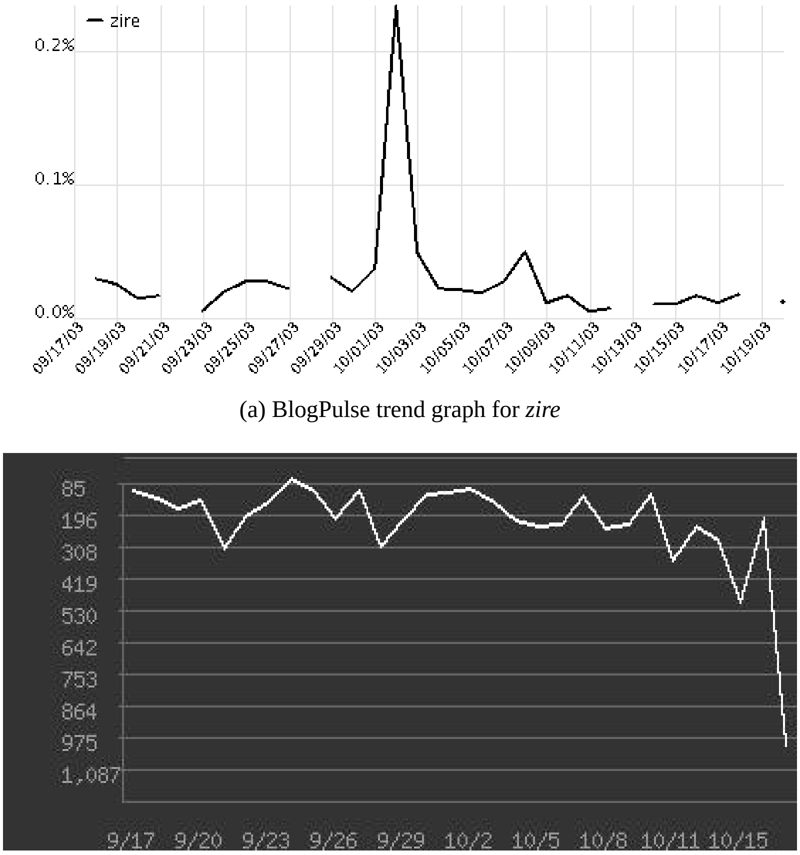
> I saw Len online today. He bought an X-Box and an extra controller and Halo. Maybe his skills will improve slightly. I am rusty on the game, need to play more.

The same weblog entry also describes the rest of the author’s day: watching Yu-Gi-Oh on television, helping a friend buy a cell phone, and getting a raise at work.

In contrast, there are entire message boards devoted to discussion of video console systems and different video games. These messages do not ramble or vary in subject the same way a weblog entry to an online journal may do. Here is the full content of a sample message posted to forum.teambox.com in the *General Xbox Live Discussion* forum:

> ok, i dunno how you broke your head set. mine is working great and i got it the day XBL came out. it’s a quality device and i think your gonna have to wait till like october before they start selling the head set seperatly which i believe will cost as much as a controler. good luck though.

If we believe the metaphor that blogging is like publishing while posting is more like chatting, it’s not surprising that weblog entries tend to be more polished pieces of writing, with fewer grammatical errors and tighter diction.



(a) BlogPulse trend graph for *zire*

(b) JungleScan graph of Amazon sales rank for the Palm Zire

Figure 2: Trend graphs for product mentions in weblogs vs. sales rank in Amazon

It is also interesting to compare relative product mentions over the different data sources. While some might object to calling candidates for the Democratic Presidentional Nomination products, here are the frequencies of mentions for some of the top Democratic contenders as of the writing of this paper:

Candidate	Blog %	Rank	Usenet %	Rank
Howard Dean	41%	1	19%	1
Weseley Clark	25%	2	12%	5
John Kerry	24%	3	17%	2
Joe Lieberman	20%	4	14%	3
Richard Gephardt	17%	5	14%	4

These numbers were obtained by running candidate classifiers over a set of weblog entries harvested from 4/20/2003 to 10/5/2003. For the sake of comparison, we also provide the relative frequencies for Usenet posts and message board posts harvested over the same period of time. Notice that the rankings of the candidates by mention counts are not the same for weblog entries as for Usenet. In addition, there are a disproportionate frequency of mentions for Howard Dean and Weseley Clark on weblogs vs. Usenet.

4.2 Trend Search

We have implemented a second analysis tool layered upon BlogPulse search: trend graphing of search results over time. This service also is not yet publically accessible. Waypath offers a similar service called "Buzz-o-meter Maker."[31]

Trend search iterates the same search query for all dates within a specified date range, bins the counts into time buckets (by default, one day buckets) and plots the result. Like Waypath's tool, BlogPulse trend search can overlay comparative trends for a number of queries. The points in the trend plot are normalized: each point in the graph represents the relative percentage of hits for a query when compared with the total number of weblog entries within the bin.

Internally, we have found trend graphing over weblogs to be almost addictive. It is tempting to treat trend search almost like an oracle. In mid-October, some of us wondered, "Are the Yankees favored over the Marlins?" (BlogPulse said the Yankees; BlogPulse was wrong.) Earlier in the month, the question on everyone's mind was what would be the outcome of the recall election. Hit count on its own indicated Arnold Schwarzenneger would be the winner.

Joking aside, relative hits counts are indicative in some domains. Here are two examples illustrating the utility of trend search over weblogs. In Fig. 1(a), the trend graphs for the top five movies for September 24, 2003 are displayed. Interestingly, their ranking in terms of weblog entry hit count is different from their rankings in terms of box office takes on the same day: "Secondhand Lions" has the fewest hits, even though it is third in terms of box office earnings.[32] In Fig. 1(b), the trend graph for "Lost in Translation" is plotted alongside the "Underworld", the top grossing movie for the week. The slowly gaining upwards trend led one of the authors to predict on 9/4/03 that this movie would be a sleeper hit: a movie initially released in a relatively small number of theaters that becomes popular through word of mouth. He was correct, as the next

[31] http://www.waypath.com/buzzmaker

[32] Data source: Box Office Mojo at http://www.boxofficemojo.com.

few weeks bore out. Notice also the periodic burstiness in these graphs; not surprisingly, bloggers post more about movies over the weekend than midweek.

Another example of the potential usefulness of trending is to compare the reaction in blogs to a product announcement with the susbsequent sales trend. Palm, Inc. announced the release of the new Palm Zire 21 on 10/1/03.[33] BlogPulse recorded a sudden spike in discussion of the Zire on 10/2/03 (Fig. 2(a)). Within two weeks, there was a sudden drop in sales rank on Amazon for the previous version of the Zire (Fig. 2b).[34]

These observations lead us to hypothesize that trends of product mentions in weblogs may have predictive power. We intend to pursue this direction. One first step is to categorize search trends into different trend genre categories: for example, spike, slow attack, long fade, cyclical, recurrent, and background noise. In the movie genre, one question of interest is: how good a predictor of box office success is the size of the spike that occurs when the movie trailer is released?

5. CONCLUSION

Weblogs are both a fascinating new medium for communication and publication and an ideal testbed for data mining algorithms. We have implemented trend discovery algorithms for weblogs as a way to test new algorithms and new ideas, and also as a way to tap into the collective consciousness of the blogsphere. It has been a fascinating journey.

Often enough, our algorithms have discovered for us important topics and viewpoints that are neglected in traditional media. For example, during the second Gulf War in the spring of 2003, key news often broke first in the weblog world.

Some amazing human interest stories also emerge from the blogsphere. A prime example is the saga of the “Star Wars Kid.” Trend search over BlogPulse shows an interesting pattern of aperiodic bursts of mentions of this phrase. The first burst occurred in Spring 2003, when a video was circulated around the Internet of a Canadian high school student who had videotaped himself performing a routine with a light saber. Bloggers around the world posted the link to their weblog. A second peak coincided with a wave of sympathy for this unfortunate teenager with whom many sympathized. There was a call put forth for contributions to a PayPal account to compensate his pain. Later peaks coincided with the breaking of the story on tradition media, then with the boy’s hiring of a lawyer, his threats, the return of contributed funds, his visit to a psychiatrist....

More recently, mentions of the new Nokia N-Gage, an integrated mobile phone and game deck, have undergone two bursts: one when the product was announced, and a second when a wave of negative reviews in the form of parody were shared across the blogsphere via an open collaborative photoblog.[35] This is a prime example of when trend graphs alone do not tell the whole story. While there has been a huge amount of buzz around Nokia’s new product, it has been almost entirely negative. Our plans for future work include applying algorithms for sentiment detection to the context surrounding key phrases, key persons, blog bites, and top links to gauge the aggregate level of positive vs. negative sentiment. Other plans include improved segmentation of weblog entries, automatically identifying comments, tracking specific indices like top movies, top company mentions, etc. Finally, if the number of weblogs continues to grow exponentially, we will need to find ways to scale the crawling and mining of weblog entries.

6. ACKNOWLEDGMENTS

We are very grateful to Robert Stockton and Matthew Siegler for their work migrating BlogPulse to a stable, production system. We have also enjoyed and benefited from Sundar Kadayam’s thoughtful input and insights.

7. REFERENCES

[1] L. Adamic and B. Huberman. Power-law distribution of the world wide web. *Science*, 287:2115, 2000.

[2] R. Agrawal and R. Srikant. Fast algorithms for mining association rules. In J. B. Bocca, M. Jarke, and C. Zaniolo, editors, *Proc. 20th Int. Conf. Very Large Data Bases, VLDB*, pages 487–499. Morgan Kaufmann, 12–15 1994.

[3] J. Allan, R. Papka, and V. Lavrenko. On-line new event detection and tracking. In *Proceedings of the 21st ACM-SIGIR International Conference on Research and Development in Information Retrieval*, pages 37–45, August 1998.

[4] D. Godes and D. Mayzlin. Using online conversations to study word of mouth communication. Harvard Business School, Division of Research, August 2003.

[5] D. Hull. Personal communication.

[6] L. S. Jensen and W. Cohen. Grouping extracted fields (html). In *Proceedings of the IJCAI-2001 Workshop on Adaptive Text Extraction and Mining*, 2001.

[7] J. Kottke. Weblogs and power laws. In *kottke.org undesign*. http://www.kottke.org/03/02/weblogs-and-power-laws, February 2003.

[8] P. Pantel and D. Lin. A statistical corpus-based term extractor. In E. Stroulia and S. Matwin, editors, *Lecture Notes in Artificial Intelligence*, pages 36–46. Springer-Verlag, 2001.

[9] D. Pennock, G. Flake, S. Lawrence, E. Glover, and C. L. Giles. Winners don’t take all: Characterizing the competition for links on the web. *Proceedings of the National Academy of Sciences*, 99(1):5207–5211, 2002.

[10] C. Shirky. Power laws, weblogs, and inequality. In *Clay Shirky’s Writings About the Internet*. http://www.shirky.com/writings/powerlaw_weblog.html, February 2003.

[11] R. Swan and J. Allan. Automatic generation of overview timelines. In *Proceedings of the SIGIR International Conference on Research and Development in Information Retrieval*, 2000.

[12] R. Swan and D. Jensen. Timemines: Constructing timelines with statistical models of word usage. In *KDD-2000 Workshop on Text Mining*, August 2000.

[13] R. C. Swan and J. Allan. Extracting significant time varying features from text. In *CIKM*, pages 38–45, 1999.

[14] T. Tomokiyo and M. Hurst. A language model approach to keyphrase extraction. In *Proceedings of the ACL Workshop on Multiword Expressions*, 2003.

[15] Y. Yang, T. Pierce, and J. Carbonell. A study on retrospective and on-line event detection. In *Proceedings of SIGIR-98, 21st ACM International Conference on Research and Development in Information Retrieval*, pages 28–36, Melbourne, AU, 1998.

[33] http://pressroom.palm.com/InvestorRelations/PubNewsStory.aspx?partner=5150&storyId=95427

[34] Graph courtesy of JungleScan.com, an online utility for tracking Amazon sales rank over time

[35] http://www.sidetalkin.com/