The Predictive Power of Online Chatter

Daniel Gruhl
IBM Almaden Research Center
650 Harry Road
San Jose, CA 95120.
dgruhl@us.ibm.com

R. Guha
Google, Inc
1600 Amphitheatre Parkway
Mountain View, CA 94043.
guha@guha.com

Ravi Kumar
IBM Almaden Research Center
650 Harry Road
San Jose, CA 95120.
ravi@almaden.ibm.com

Jasmine Novak
IBM Almaden Research Center
650 Harry Road
San Jose, CA 95120.
jnovak@us.ibm.com

Andrew Tomkins
IBM Almaden Research Center
650 Harry Road
San Jose, CA 95120.
tomkins@us.ibm.com

ABSTRACT

An increasing fraction of the global discourse is migrating online in the form of blogs, bulletin boards, web pages, wikis, editorials, and a dizzying array of new collaborative technologies. The migration has now proceeded to the point that topics reflecting certain individual products are sufficiently popular to allow targeted online tracking of the ebb and flow of chatter around these topics. Based on an analysis of around half a million sales rank values for 2,340 books over a period of four months, and correlating postings in blogs, media, and web pages, we are able to draw several interesting conclusions.

First, carefully hand-crafted queries produce matching postings whose volume predicts sales ranks. Second, these queries can be automatically generated in many cases. And third, even though sales rank motion might be difficult to predict in general, algorithmic predictors can use online postings to successfully predict spikes in sales rank.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Miscellaneous*

General Terms

Algorithms, Experimentation, Measurements

Keywords

Blogs, Prediction, Sales rank, Time-series analysis

1. INTRODUCTION

The World Wide Web represents a global, timely, and largely unregulated touchstone of popular opinion, which many believe may be exploited for early insights into new trends and opinions. Areas proposed for such analysis include the outcome of political elections, the emergence of the next big musical group/toy/consumer electronic device, and the pulse of the global economy. Yet, despite widely touted opinions that marketing will soon be a small branch of machine learning, there has been little work formally demonstrating connections between online content and customer behavior such as purchase decisions. In this paper we look at the relatively new phenomenon of blogs (weblogs) and measure how well it reflects the comparatively old practice of buying books. We use sales rank data from Amazon.com, accessed via its Web Service interface, to obtain a fine-grained measure of the relative popularity of books. We use books, as opposed to cookware or garden equipment, because sales of books are much less likely to be driven by special marketing offers as compared to other categories.

Overall, our goal is to demonstrate conclusively that in some cases, spikes in sales rank may be predicted based on online chatter. Of the products we study, around 20% manifest both sufficient online discussion for analysis, and a strong correlation between blog mentions and sales rank. Of those, online chatter is a leading indicator for more than half. Based on the information available in blogs today, successful prediction of a large fraction (i.e., 90%) of sales spikes does not appear to be a realistic goal. Nonetheless, marketers seeking an advantage in prediction of sales do not require 90% success rates, and are in fact well-positioned to make use of reasonably accurate predictions of some fraction of spikes.

We begin by removing the question of algorithmic prediction of sales rank motion, and address a simpler question: does blog data exhibit any recognizable pattern prior to spikes in sales rank. To answer this question, we consider carefully hand-constructed predicates on the content of blog postings intended to capture only those postings that contain discussion of a particular product. We present our

hand-crafted predicates and the resulting plots of blog mentions and sales rank to show that clear indications are often visible in blog mentions before there is any response in sales rank. Amazon sales rank is updated every six hours based on sales during the previous period—a much finer granularity than the lag between blog postings and sales rank motion.

Next, we re-introduce a piece of the computational question: is it possible to automatically formulate predicates that can capture subsets of blog postings that are sufficiently well-connected to discussions of a particular product that they retain the leading behavior demonstrated by our hand-crafted queries? Our techniques are successful in approximately one third of product spikes for which a successful hand-crafted query exists.

Having verified that automatically-generated queries represent leading indicators of sales rank motion, we close the loop by developing automated algorithms to predict spikes in sales rank given a time series of counts of blog postings. We develop a simple stateless model of customer behavior based on a series of states of excitation which are increasingly likely to lead to a purchase decision. We show that this model yields a predictor of sales rank spikes that is significantly more accurate than automated techniques operating on the sales rank data alone.

At this point, we introduce a note regarding causation. None of our work suggests that bloggers are responsible for the spike in sales rank through, for example, the mechanism of viral marketing. In certain cases, it is possible that such a mechanism could be at work. We feel that in the majority of cases, however, it is more likely that bloggers are simply non-causative indicators of some other root cause of behavior, typically an event in the outside world. There are two possible explanations for the delay between postings and change in sales rank. First, bloggers may be forward-thinking people who both write and buy earlier than others, but who represent a sufficiently small fraction of the population that no general spike in sales rank is visible until later. Second, bloggers may be no more forward-looking than the general population, but the threshold to writing about a product may be lower than the threshold to purchasing the product, so the lag may be due to the additional time required to reach the higher activation energy to purchase. Our study does not suggest which is the case; we leave this question for future research.

1.1 Organization

In Section 2 we discuss the related work on prediction, blogs, and time-series analysis. In Section 3 we describe our data sources which include WebFountain and Amazon Web Services. In Section 4 we analyze spikes in sales rank and blog mentions and discuss the cross-correlations between the two time series. While the queries in Section 4 were manually constructed, in Section 5 we outline a simple mechanism to automatically construct queries to locate mentions of a book in blogs. In Section 6, we discuss the use of blog mentions to predict sales rank. Final remarks are made in Section 7.

2. RELATED WORK

Predicting sales from other indicators is an important problem in marketing and business. The very concept of creating a new product is predicated on the assumption (or rather, prediction) that someone will eventually purchase it. The same can be said for pricing, inventory planning, production capacity planning, store placement and layout, etc. As such, there exists copious prior art in this area alone (see e.g., Harvard Business Review, MIT Sloan Management Review, McKinsey Quarterly).

Sornette et al. [12] present a study analyzing the nature of sales spikes in `amazon.com` sales rank data, the same corpus we study. They show that two distinct types of peaks may be identified by their growth and relaxation patterns, and they tie these two spike types to endogenous and exogenous events.

Detection of events in news streams continues to be a key area of ongoing research. Smith [20] describes a system for automatically identifying events in unstructured data. Yang et al. and Papka study the use of text retrieval and clustering techniques for detecting the presence of a new event in a stream of news stories [24, 3, 19]. This work, however, does not focus on predicting one event using another event.

Some work has been done in the area of developing models that apply to message boards, chat rooms, and blogs relating the reviews to stock movements. The effects of Internet disclosure on various companies' stock prices were studied by Admati and Pfleiderer [2]. Tumarkin and Whitelaw [22] showed that days of high discussion activity on the Raging Bull discussion forum correlated with abnormal market returns; however, they showed that this abnormal activity did not predict market returns. Similarly, Antweiler and Frank [4] note that postings on stock message boards correlate with stock volatility, but they did not find that these discussions predict returns.

Early work on predicting sales from online postings includes that of Tong [21], who predicted box office proceeds of movies from opinions posted to net news. Tong's company Opion was purchased by PlanetFeedback, which was later acquired by Intelliseek, who have published extensively on the power of internet discussion in understanding customer views of a product or brand; see, for example, [7]. Human or human-assisted analysis of online content is the focus of several other companies such as Carma [9] and Biz360 [6]. Whitman and Lawrence [23] examined community-created meta-data on music artists and found that buzz on blogs led record sales by two weeks.

The propagation of information through Blogspace was studied in [14]; here, the authors propose a model for flow of information from person to person in Blogspace. The dynamics governing the flow of information naturally contributes to the generation of spikes in blog mentions. However, this paper does not proceed to track or correlate this information flow with any other time series. Arbesman [5] employed the reverse approach of injecting a meme into blogspace and tracking its flow. In general, many fascinating aspects of blogs have been actively studied for the past two years. The structure and evolution of blogs were explored in [16, 17, 18, 1].

3. DATA SOURCES

In this section we describe the various data used in our experiments.

3.1 Online postings data

IBM's WebFountain project [13, 11] maintains large collections of online material, including postings from approximately 300K blogs. Approximately 200K postings arrive

into the system per day. The system also contains just under three billion web pages and 200K media articles per day from the Factiva media feed. All of these documents are accessible via an index that can provide arbitrarily large result sets for further processing.

3.2 Amazon sales rank data

Using Amazon's Web Services (`http://www.amazon.com/gp/aws/landing.html`), we collected data for 120 days (Jul 2004 to Oct 2004) on all products that at any point during this interval reached a sales rank of less than 300. Once a product was added to our list, we continued to track it whether or not it remained within the top 300. On average, we refreshed the sales rank of each item five times per day, and we checked for new entrants into the top 300 four times a day. Overall, 2,430 books (including ebooks) made it into the top 300 in this period and 480,346 salesrank readings were recorded during this period.

Amazon publishes sales rank information as a service to its customers. They describe the update frequency of sales rank information as follows:

> The calculation for a book's Sales Rank is based on Amazon.com sales and is updated regularly. The top 10,000 best sellers are updated each hour to reflect sales over the preceding 24 hours. The next 100,000 are updated daily. The rest of the list is updated weekly, based on several different factors.

The sales rank published by Amazon is a rich measure—see `http://www.fonerbooks.com/surfing.htm` for a detailed account of what this information means to authors and publishers.

The fact that the ranks of popular books are updated frequently is crucial to our study, as we wish to understand the delay between online chatter and purchasing behavior. Due to Amazon's extremely rapid refresh rate, and our ability to recrawl these ranks every few hours, we have precise information about when sales of a book began to spike.

4. CORRELATION BETWEEN SALES RANK AND BLOG MENTIONS

In this section we study the correlation between online mentions of a product and sales of that product. As described above, we have gathered sales information from Amazon's published product sales ranks for over 20,000 products, but throughout, we restrict our attention to the 2,340 books in our set. Of these books, there are about 1,000 that have achieved a sales rank within the top 200 during the period of our measurements, i.e., Jul 2004 to Oct 2004.

A word on language: when we speak of the minimum or maximum of a set of sales ranks, we refer to the standard operators, so the min of a set of ranks is in fact the best-selling rank of the set. We will use the term "best" rank to refer to the minimal rank of a set, and we will avoid ambiguity by never speaking of the highest or lowest rank.

We first discuss a natural notion of spike in sales rank and present a simple way to detect spikes. We use this to restrict our attention to books that exhibit spikes in their sales rank. Then, we compose appropriate queries for each of these books to retrieve the blog mentions in our collection of online postings data. We illustrate by means of three examples, occurrences of spikes in blog mentions that are highly correlated with the spikes in sales rank. We then measure the statistical cross-correlation between the sales rank and blog mentions time series and make several conclusions.

Detecting spikes in sales rank.

The first task is to identify books that have a spike in their sales during the observation period. A book is said to have a *spike* in sales rank if the minimal rank occurs more than two weeks from the start and end of our measurements, and if all the ranks that *do not* occur within a week of the minimum rank are "large enough". We hand-tuned the definition of large enough to generate spikes that looked reasonable to the eye; a value is said to be large enough compared to a minimum rank m if it is greater than $\max(m + 50, 1.5m)$. We apply this simple automatic spike detection algorithm to the 1,000 books whose best rank is 200 or better. According to this algorithm, 50 of the 1,000 high-ranking books contain a spike in sales rank within the time interval of our study.

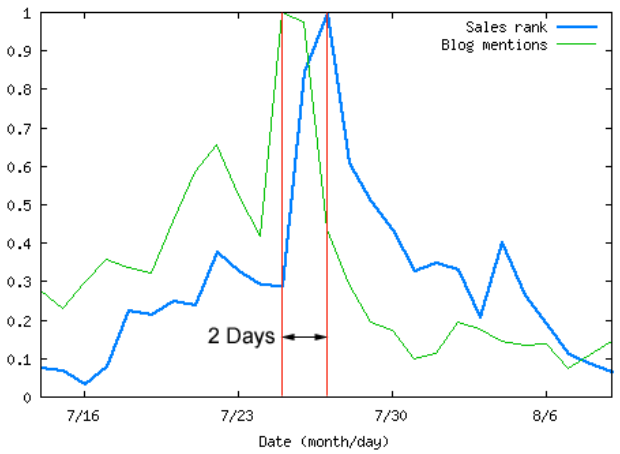
Locating blog mentions about books.

The next task is to identify the blog mentions about these 50 books. For this purpose, we created a user interface that would take a query and a book identifier as input, and produce a graph over time that plots the sales rank of the product and the number of postings matching the query vs. time. The user of the tool was instructed to create and iteratively refine a query until it seemed to track mentions of the product of interest in blog postings. We (the authors of this paper) used this tool in an iterative fashion to hand-craft queries for each of the 50 books. Figures 1 and 2 show three sample representative results generated using this tool; the sales rank and blog mentions are scaled to have maximum value of 1 so that it is easy to visualize both the curves simultaneously. Our plots and correlation computations use the reciprocal of sales rank in order to capture the notion that a drop from first to second rank is much more significant than a drop from rank 1,000 to 1,001.

The first book is 'The Lance Armstrong Performance Program: Seven Weeks to the Perfect Ride' authored by Lance Armstrong, Chris Carmichael, and Peter Joffre Nye. Notice that the spikes in both blog mentions and sales rank coincide with Armstrong winning the Tour de France on Jul 25, 2004 (Figure 1). Also notice the distinct lag between the spike in blog mentions and the spike in sales rank, where the former preceded the latter by two days.

The second book is 'What Not to Wear' by Trinny Woodall, S. Constantine, Robine Matthews, and Susannah Constantine (Figure 2). While there does not seem to be a single event responsible for the spike, we were able to identify two plausible contributing factors. Firstly, the message board corresponding to the TV show (`http://discovery.infopop.net/1/OpenTopic?a=cfrm&s=6941912904&f=7121920016`) started taking style question submissions to be aired on the show around that period and got 281 responses from Sep 1–10, 2004; secondly, the authors had another book 'What You Wear Can Change Your Life' released on Sep 17, 2004. The third book is the classic 'Vanity Fair' by William Thackeray; the spikes coincide with the book-based movie released on Sep 1, 2004. The above samples clearly illustrate the breadth of this phenomenon—in terms of both genre and media.

While we have focused our analysis on books demonstrating a spike in sales rank, we note that there are other types



Query: Lance Armstrong OR Tour de France

Figure 1: The Lance Armstrong performance program.

of movement in sales rank that may be of interest. For example, sales of the book 'The Notebook' by Nicholas Sparks reached a high point following the release of the film version on Jun 28, 2004, but has been steadily falling over the last four months. Figure 3 shows that blog mentions of this book (specifically, the query 'the notebook AND nicholas sparks') have also been decreasing, echoing the decline in sales.

Correlation of time series.

In order to report an analytical comparison of mentions and sales rank, we turn to the theory of correlation of time series; note that the blog mentions and sales rank are both time series. Suppose $\mathbf{x} = x_1, \ldots, x_n$ and $\mathbf{y} = y_1, \ldots, y_n$ are two time series. The cross-correlation function of two time series is the product-moment correlation as a function of lag, or time-offset, between the series (cf. [10]). The sample cross-covariance function is given by

$$c_{\mathbf{xy}}(k) = \frac{1}{n}\sum_{t=1}^{n-k}(x_i - \mu(\mathbf{x}))(y_i - \mu(\mathbf{y})) \quad k = 0, \ldots, n-1,$$

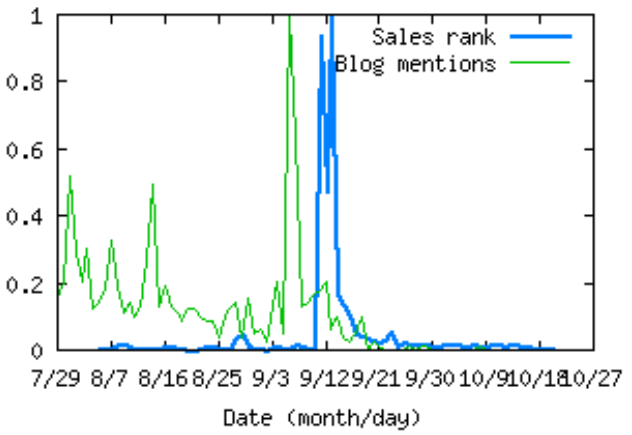
$$c_{\mathbf{xy}}(k) = \frac{1}{n}\sum_{t=1-k}^{n}(x_i - \mu(\mathbf{x}))(y_i - \mu(\mathbf{y})) \quad k = -1, \ldots, -(n-1),$$

where $\mu(\cdot)$ is the sample mean and k is the lag. The *sample cross-correlation* is the cross-covariance scaled by the variances of the two series:

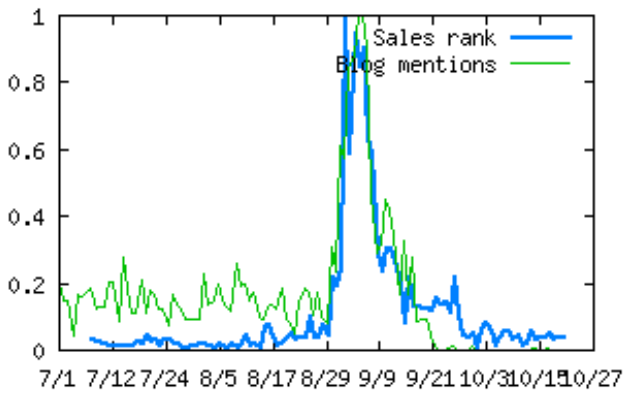
$$r_{\mathbf{xy}}(k) = \frac{c_{\mathbf{xy}}(k)}{\sqrt{c_{\mathbf{xx}}(0) \cdot c_{\mathbf{yy}}(0)}},$$

where $c_{\mathbf{xx}}(0), c_{\mathbf{yy}}(0)$ are the sample variances of $\mathbf{x}$ and $\mathbf{y}$ respectively. The *best lag* is $\arg\max c_{\mathbf{xy}}(k)$, i.e., the k where the cross-correlation is maximum. The best lag is said to be *leading* if it is negative and *trailing* if otherwise; the former represents that $\mathbf{x}$ leads $\mathbf{y}$ as a time series and the latter represents the converse. In our experiments, $\mathbf{x}$ is always the (inverse) sales rank time series and $\mathbf{y}$ is always the blog mentions time series.

Figure 4 shows the cross-correlation graphs at various lags for the sample spike queries of Figure 2. The best lag can



Query: What not to wear



Query: Vanity Fair OR William Thackeray

Figure 2: Sample books with spikes in both sales rank and blog mentions.

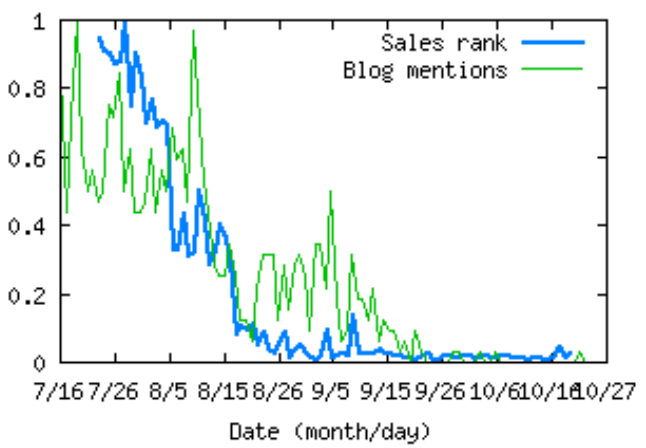


Figure 3: Falling sales and blog mentions of "The Notebook".

be as small as a couple of days to as large as a couple of weeks. The best lag is leading for the first two books and is slightly trailing for the third book. Of the 50 books with spikes during the period of our analysis, highly correlated spikes in blog mentions were observed in ten of them.

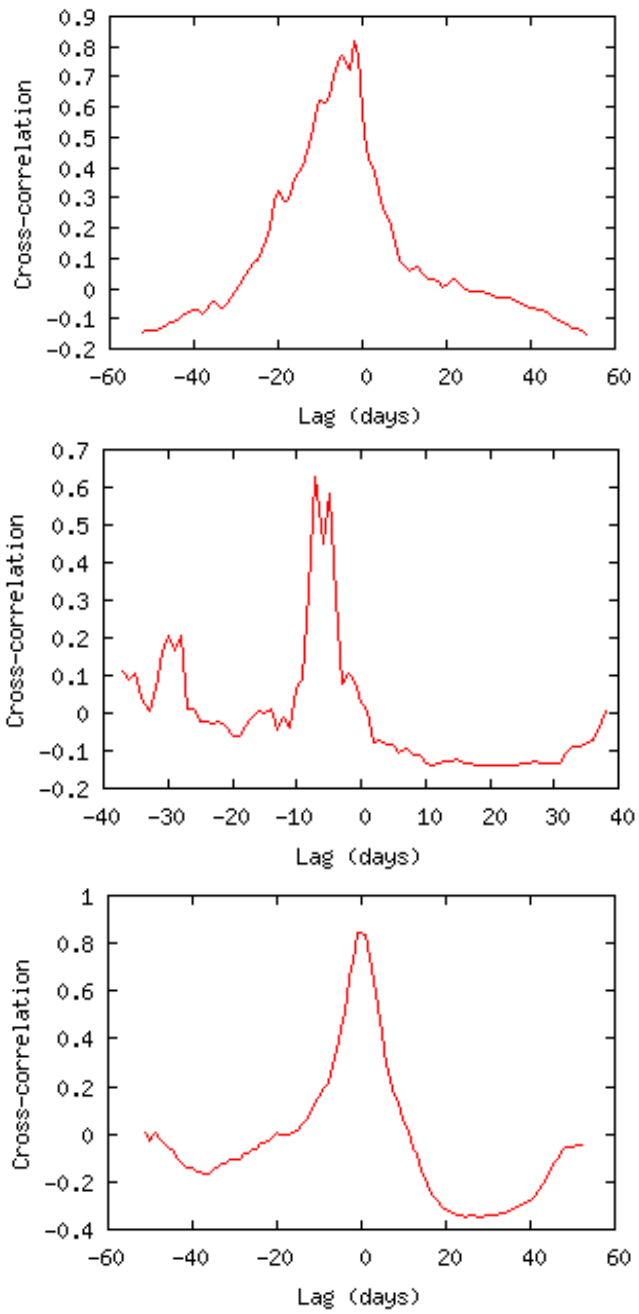


Figure 4: Cross-correlation plots for the three sample books.

Several interesting statistics can be derived from the cross-correlation data. Table 1 shows the cumulative histogram of the maximum cross-correlation statistic, i.e., for a given cross-correlation γ, it gives the fraction of books with maximum cross-correlation at least γ and the average best lag for those books. As expected, the average best lag gets closer to 0 as the correlation between the sales rank and blog mentions increases; this adds further evidence to the strong interplay between blog mentions and sales rank.

Table 2 shows the histogram of the fraction of books as a function of the total number of blog mentions, i.e., for a given range β, the fraction of books with blog mentions in the range β and the average maximum cross-correlation and average best lag for those books. From the table, it is clear that if a book receives a lot of blog mentions, then it is more closely correlated with a smaller best lag.

γ	0.4	0.5	0.6	0.7	0.8
Fraction	0.44	0.22	0.14	0.08	0.06
Best lag	-8.8	-7.4	-5.7	-5.0	-1.7

Table 1: Average fraction of books and average best lag as a function of maximum cross-correlation γ.

β	< 50	$50 - 100$	$100 - 200$	> 200
Fraction	0.53	0.12	0.14	0.21
Max cross-corr.	0.36	0.36	0.36	0.56
Best lag	-17.2	-16.2	-24.1	-8.2

Table 2: Average fraction of books, average maximum cross-correlation, and the average best lag as a function of number of blog mentions β.

Discussion.

From the above experiments, it follows that if there is a spike in the sales rank and there are lots of blog mentions about the book, then the blog mention tends to have a spike that is correlated well with the sales rank. Furthermore, a maximum cross-correlation value of at least 0.5 suggests a good correlation and the best lag is almost always leading. The latter implies that *a sudden increase in blog mentions is a potential predictor of a spike in sales rank.*

Blogs aside, there are many reasons why a book may manifest a sudden increase in popularity. Many of these reasons have the potential to cause a leading spike in blog mentions. As we saw in our examples, one or more popular events might contribute to online chatter which can be a precursor to a spike in sales rank. However, this phenomenon is not always guaranteed. There are several common reasons for increases in sales rank with no corresponding increase in blog mentions. First, the issue of marketing promotions. If Amazon places a certain book in a privileged position on the website, or discounts the book heavily, a spike in sales rank could occur. Typically, there will be no matching spike in blog activity, although occasionally we observed postings of the form "at this price, you can't afford not to buy this book. . . ." Second, if a book is released for the first time with great fanfare, a spike in sales rank may result. (This prompted us to decide to avoid applying spike detection to books published in 2004.) Third, occasionally a wholesaler or other bulk purchaser or special interest groups may buy a large block of books, causing a spurious spike in sales rank. (A common subcategory of this phenomenon is the bulk buying of textbooks before school begins.) Last, certain books in our set of 50 are ranked close to 200 at their best moment, and simply do not catch the eye of the blogging public. With broader coverage of blogs, and increased publication, presumably the threshold for sufficient mentions of such books will only increase.

5. AUTOMATED QUERY GENERATION

We have shown that carefully hand-crafted queries can be used to generate leading indicators of sales rank motion. This leads naturally to a follow-on question: is it possible to automate the selection of those queries? In this section, we consider one such technique, and answer the question in the affirmative. We present a simple technique, which admits many possible extensions and improvements beyond our scope, but which nonetheless demonstrates that automatic query selection is possible with reasonable accuracy. The queries produced by this technique are far inferior to those produced by human effort—it is an open research problem to improve the performance of this simple technique.

The automatically-generated queries we propose are based on the names of the authors of the book; note that one could use the title of the book as well, but this is more prone to false matches especially in an automated method. We use information from the 1990 US Census in order to estimate the number of people who might have the name of a certain author, as a proxy for the ambiguity of that author's name; other methods for disambiguation include using the frequency of the name on the web. A simple heuristic is described below.

Algorithm Automatic-author-query

min = 10^{-8}
Single-author books:
author = firstname, lastname
If Pr[lastname] < min **then**
    query = "lastname"
else
    query = "firstname lastname" OR "full name"
Two-author books:
authors = firstname1, lastname1 & firstname2, lastname2
If lastname1 = lastname2
    query = "lastname1"
else
    query = "lastname1 AND lastname2"

Since most of the books in our data set had fewer than three authors, the above heuristic was sufficient. If needed, the heuristic can be extended in a natural way to work for multi-author books and to use selected words from the book title. We also note that a similar heuristic applies in a straightforward way to domains such as music, media, and movies. By using more elaborate domain-specific key words, automatic queries can be generated for other products as well.

Figure 5 shows a sample representative of results generated using our automatically-generated queries. The top plot shows sales rank for the book 'The Last Night of the Yankee Dynasty: The Game, The Team, and the Cost of Greatness' along with the mentions retrieved with the automatically generated query 'Buster Olney,' the author of the book. The lower plot shows the cross-correlation of sales and mentions for various lags. This is also an example of a query run against the entire set of media available in the WebFountain index (web pages, message boards, blogs, newspapers, etc.). Although we have mainly focused our analysis on blogs mentions, we note here that we attained similar results when using our full set of online sources for a limited set of experiments. However, a full analysis of other online sources is a topic for future work.

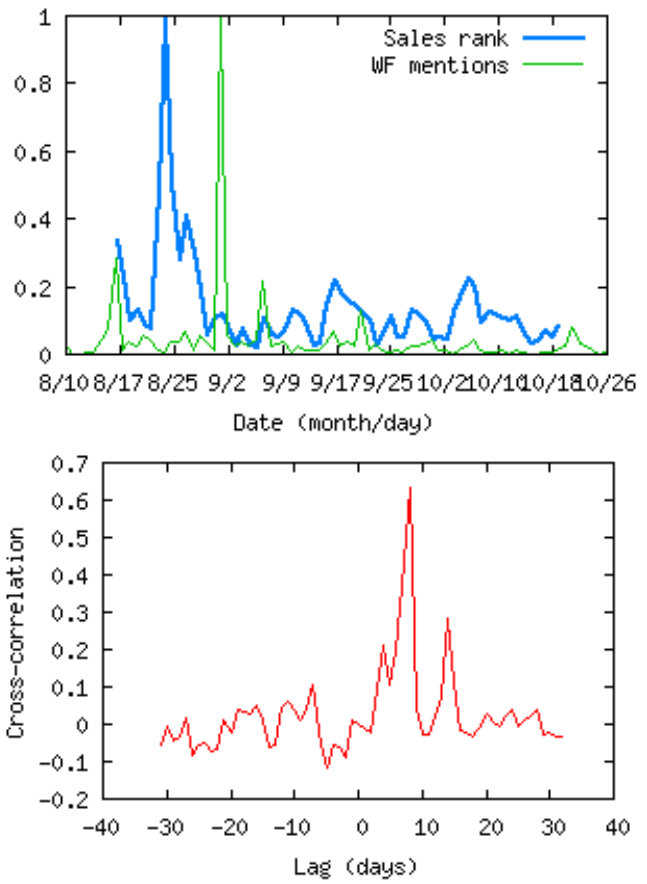


Figure 5: Example of sales rank and blog mentions for an automatically-generated query and their cross-correlation.

When run on blogs, our automatically-generated queries retrieved various numbers of mentions ranging from 0 to over 20,000. Approximately half of the queries retrieved over 50 mentions, while 30% retrieved no mentions at all. Figure 6 shows a histogram of the number of queries retrieving each number of mentions.

Figure 7 shows a scatter plot of the best correlation and the best lag of mentions that respond to the automatically-generated query versus the best lag. As noted earlier, a cross-correlation of 0.5 or more corresponds to relatively strong positive correlation between blog mentions and sales rank. As the figure shows, a large fraction (45%) of the queries have a cross-correlation greater than 0.5. Moreover, 35% of the queries have both a cross-correlation greater than 0.5 and the best lag of less than two weeks.

Figure 8 shows the number of documents with a certain best lag. As the figure shows, the best lags center around zero, and are more likely to be negative, indicating that *blog mentions are more likely to lead sales rank changes, rather than follow them.*

A variety of other approaches based on more sophisticated processing of the authors, the title, the category, the keywords, or the comments posted about the book could also be applied. We have established that it is possible to algo-

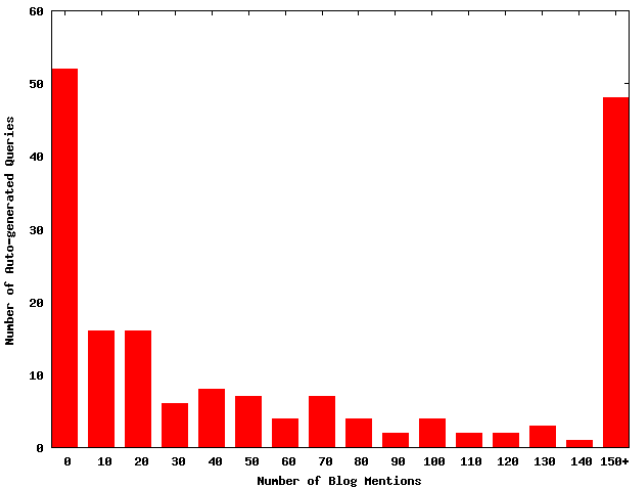


Figure 6: Histogram of number of automatically-generated queries that retrieve a particular number of blog mentions.

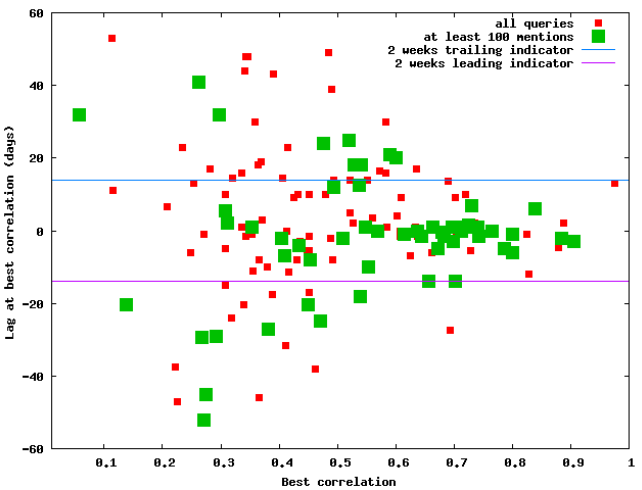


Figure 7: Scatter plot of cross-correlation versus lag for 182 automatically-generated queries.

rithmically create boolean predicates that produce subsets of mentions that operate as leading indicators of sales rank spikes. However, we have not fully explored the space of approaches to this problem, and feel that doing so is an area for fruitful further work.

6. SALES RANK PREDICTION

So far, we have focused on creating either manually or automatically generated sets of mentions which somehow track sales information. We have measured the overlap between these two time series using cross-correlation, a technique which inherently requires information about the entire time series of blog mentions and sales rank. The remaining question, then, is the following: given the time series representing sales rank data up to a point t, does the addition of blog mention data for the same period improve our ability to predict what will happen to sales rank?

We will show both a negative and a positive result. The

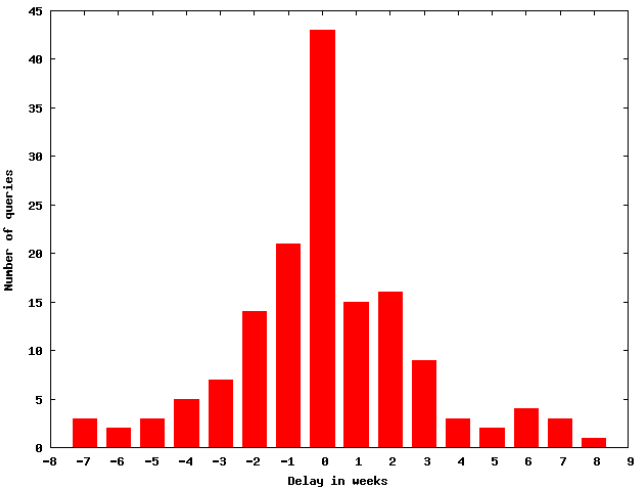


Figure 8: Histogram of number of queries with a particular best lag time.

negative result is that *predicting whether tomorrow's sales rank for a particular book will be higher or lower than today's sales rank appears to be hard.* Further, the addition of blog mentions data does not seem to help significantly. And more surprisingly, even the prediction of tomorrow's volatility based on the history leading up to today also appears hard.

Despite this seeming unpredictability of sales rank motion, we show a very strong positive result: analysis of blog mention data up to a point t allows us to effectively *predict when there will be a future spike in sales ranks, without recourse to information from the future, and even without recourse to the history of sales ranks.*

6.1 Predicting motion, volatility, and spikes of sales rank

We consider natural predictors of motion and volatility.

Moving average and least-squares predictors.

Given a fixed-size history of sales rank figures, predict that tomorrow's figure will be a weighted average of the history—this is known as the *moving average predictor* in time-series analysis [8]. In attempting to predict upward or downward motion, we measure the predictor output via a single bit indicating its guess about whether tomorrow's sales rank will be less than today's, or greater than or equal to today's.

Algorithm Moving-average-predictor

Let w be the window size
Let n be the number of data points
Let $c_1, \ldots, c_w$ be coefficients with $\sum_{j=1}^{w} |c_j| = 1$
$d = 0$
For t $= w$ **to** n
  $p = \sum_{i=1}^{w} c_i x_{t-i}$
  If $\text{sign}(p - x_{t-1}) = \text{sign}(x_t - x_{t-1})$ **then**
    $d = d + 1$
Prediction correctness $= d/(n - w)$

We considered different weighting schemes for choosing coefficients with uniform weights (i.e., $c_j = 1/w$), exponen-

tially decaying weights (i.e., $c_j \propto 2^{-j}$), and harmonically decaying weights (i.e., $c_j \propto 1/(j+1)$). Over a wide range of window sizes w and choice of weights, our best classifier produced an accuracy of 63%.

Another predictor is the *weighted least-squares predictor*. Here, the idea is to perform a weighted regression on the last w values in the time series and predict the next value based on the regression. Again, the best accuracy of the least-squares predictor over different weighting schemes turned out to be around 60%.

Markov predictor.

Given a fixed-size history of sales ranks, the idea in a Markov predictor is to predict that tomorrow's sales rank will be the most likely rank given the history, based on a separate training set. For document d, let $d(t)$ be its rank at time t. Formally, the algorithm is described below.

Algorithm Markov-predictor

Let w be the window size
Let $f(\cdot)$ be a feature quantizer
Let queue(H) be the queue of feature history
Let $D(H)$ be the learned distribution for history H
Training step:
For d in training set documents do
  For t in timesteps do
    $\delta = f(d(t) - d(t-1))$
    $\delta' = f(d(t-1) - d(t-2))$
    insert (δ', queue(H))
    If $|H| = w$ **then** delete (queue(H))
    Increment count of δ in $D(H)$
Testing step:
For d in test set documents do
  For t in timesteps do
    Let H be the w-history at t
    $\delta = f(d(t) - d(t-1))$
    $\delta' = f(d(t-1) - d(t-2))$
    insert (δ', queue(H))
    If $H = w$ **then** delete (queue(H))
    Predict most likely outcome from $D(H)$

The feature quantizer we used mapped the differences in sales rank into twenty-seven buckets. Different buckets were used to capture positive and negative values. The quantization step is as follows:

Algorithm Feature-quantizer

Let Δ be the input
If $|\Delta| <= 10$ **then return** Δ
If $|\Delta| <= 25$ **then return** $\text{sign}(\Delta) \cdot 25$
If $|\Delta| <= 50$ **then return** $\text{sign}(\Delta) \cdot 50$
If $|\Delta| <= 100$ **then return** $\text{sign}(\Delta) \cdot 100$

As before, over a wide choice of parameters, the best classifier's accuracy was 63%. For a two-class prediction problem, these numbers are not high, suggesting that local sales rank behavior may be hard to predict in general.

Predicting volatility.

Based on an observation that certain books and certain regions appeared to be more volatile than others, we attempted instead to predict whether tomorrow's sales rank would differ from today's by more than a certain threshold value. The threshold was taken to be 44, as this value resulted in 50% of the data points exceeding the threshold. Once again, we tried all of our predictors, and in this case the best success rate was 72%.

Predicting spikes.

The model we seek to validate is the following. During most time periods, sales ranks move up and down due to the whims of users. However, occasionally some external event causes a spike in sales, which is a brief period of signal amidst the background noise of sales rank motion. Our hypothesis is that prediction of these brief signals is possible based on blog mentions data.

In order to explore this conjecture, we must "truth" our data so that we may determine whether an algorithm correctly predicted the presence of a spike. Thus, we must tag the regions of the time series which are deemed to be part of the spike. The tagging algorithm works as follows. First, the point of minimal sales rank (the center of the spike) is located. Let m be the sales rank at this point. Next, a threshold τ is set to be $\max(1.5 * m, m + 10)$. The region of the spike is taken to be the maximal interval containing the point of minimal sales rank, and no point of sales rank greater than τ.

Given this new truthed data set of 50 spikes, we studied the 47 spikes for which we were able to suggest a reasonable hand-crafted query. Our goal is to determine whether an automated analysis of the number of postings that match this query will allow us to predict a spike in sales rank. The experiment proceeds as follows. First, a particular product and a time t is fixed. Next, the predictor is given as input the number of blog postings matching the query for that product on all days up to and including t. The predictor must then output a bit indicating whether it believes a spike in sales rank will occur in the near future. The success of the predictor can then be evaluated against the truthed data set.

The predictor we apply simply attempts to determine when a spike is occurring in blog mentions, and then and only then predicts a forthcoming spike in sales rank. The algorithm has three goals:

1. Find spikes that appear to be the biggest ever, since we are interested in essentially unpopular books spiking into popularity
2. Find spikes that exceed historical averages by a significant amount
3. Find spikes that rise relatively quickly

These goals translate into three simple conditions the algorithm uses to determine whether it should predict a forthcoming spike in sales rank. Let $\mu(\cdot)$ and $\sigma(\cdot)$ denote the mean and standard deviation respectively.

Algorithm Spikes-predictor

Let h be history
Let c be current
Let $\sigma = \sigma(h)$
Let $h' = h$ until five days ago
If $c > \max(h)$ AND $c > \max(h') + \sigma$
    AND $c > \mu(h) + 4\sigma$ **then**
  Predict spike

Whenever the predictor predicts that a spike will occur, we evaluate the prediction into one of four categories:

(1) *Leading:* A spike occurs after time t but within two weeks

(2) *Trailing:* A spike already occurred within the past two weeks

(3) *Inside:* A spike is currently occurring

(4) *Incorrect:* A spike does not occur within two weeks of the current time

We consider the experiment run on blog mentions, and also on web pages in the WebFountain store that hit the query. In both cases, 3/4 of the results contain a leading or trailing prediction. Overall, there were a total of 39 predictions made on these spikes, of which 2/3 (26) were leading or trailing. Figure 9 shows a histogram of the number of predictions at a particular lag. As before, negative best lags represent leading indicators; the figure shows that, even though predictions are based solely on blog data, they typically lead sales rank changes. Also, the predictor does not know how many spikes occur in the sales rank plot, but nonetheless predicts with high accuracy even compared to a random predictor that knows it searches for a single spike.

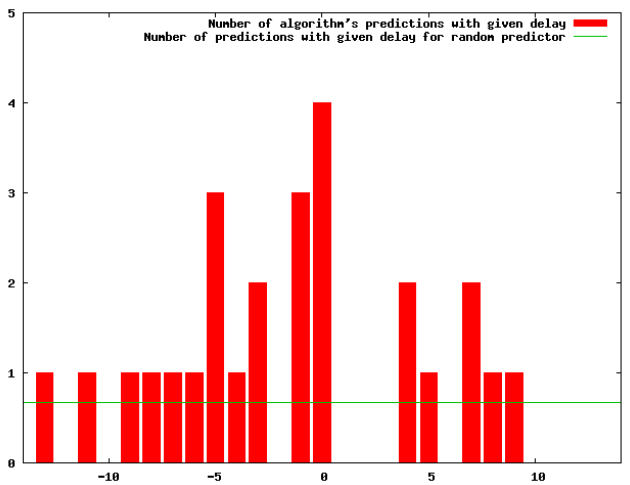


Figure 9: Best lags of leading and trailing predictions. Negative best lags are leading indicators.

7. CONCLUSIONS AND FUTURE WORK

In this paper we have explored the increasingly widely-held belief that online "chatter" in the form of blog postings and web discussions may represent an early indicator of "real-world" behavior. We show that volume of blog postings can be used to predict spikes in actual consumer purchase decisions at online retailer Amazon.

Our work consists of three pieces. First, we show that carefully-constructed queries can generate sets of postings that discuss a particular product, and that plots of these discussions often display early indications of future spikes in sales rank. We verify these findings with the tools of time-series analysis. Second, we show that such queries can be automatically constructed for certain types of books using simple techniques. The effectiveness of such techniques is good when they apply, but more work is required to generate a broader family of such techniques that apply in more situations. Third and finally, we show that even though the prediction of motion in sales rank appears challenging, simple predictors based on blog mentions around a product can be effective in predicting spikes in sales rank.

Future work falls into the following categories. First, we do not have a complete characterization of the situations in which online chatter is sufficiently voluminous, targetable, and predictive to capture future sales activity. Second, our tools for automated query generation and prediction only scratch the surface of what is possible; many more features and techniques could be applied to these problems. Third, we study the domain of book sales, but the same techniques could be applied in tracking of sales of other goods, and in prediction of other events such as voting behavior or popular response to corporate and public policy decisions.

Finally, we propose the following simple model that can explain the behavior of spikes in blog mentions and sales rank. Such a model most critically must be able to produce two distinct spikes with varying best lags in blog mentions and sales rank. Our model of population is inspired by Hidden Markov models and the model of bursts used by Kleinberg [15]. In this model, $B_1, \ldots, B_n$ denote the state of bloggers where higher-numbered states correspond to a populace more involved in discussion of the product. States $S_1, \ldots, S_m$ denote a decision to buy, where S_i means the sales rank spike will occur on $m - i$ days in the future. The output distributions from B_i are determined empirically; higher numbered B_i's are both more likely to transition to a S_j state and more likely to transition to a higher S_j state. We hope to study and analyze this model for the purpose of developing improved prediction, forecasting, and automatic spike detection algorithms, especially using the methods in [15].

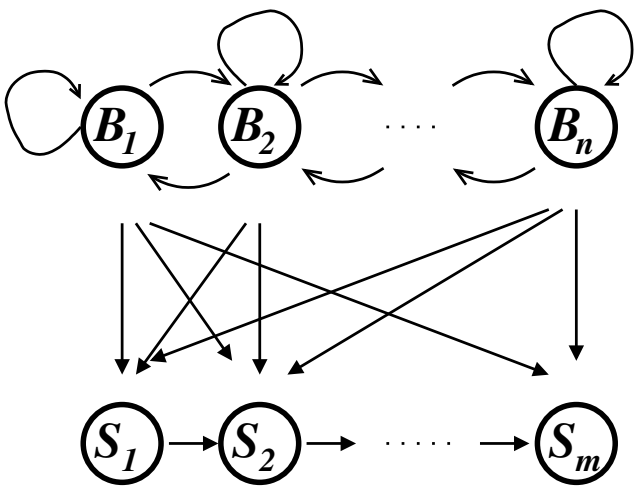


Figure 10: A proposed Markovian model for blog mentions/sales rank.

8. REFERENCES

[1] E. Adar, L. Zhang, L. A. Adamic, and R. M. Lukose. Implicit structure and the dynamics of blogspace. Workshop on the Weblogging Ecosystem, 13th International World Wide Web Conference, 2004.

[2] A. Admati and Pfleiderer. Disclosing information on the internet: Is it noise or is it news? Technical report, Graduate School of Business, Stanford University, 2001.

[3] J. Allan, J. Carbonell, G. Doddington, J. Yamron, and Y. Yang. Topic detection and tracking pilot study: Final report. Proc. of the DARPA Broadcast News Transcription and Understanding Workshop, 1998.

[4] W. Antweiler and M. Z. Frank. Is all that talk just noise? The information content of Internet stock message boards. *Journal of Finance*, 59(3):1259–1295, 2004.

[5] S. Arbesman. The memespread project: An initial analysis of the contagious nature of information in online networks. http://www.arbesman.net/memespread.pdf, 2004.

[6] Biz360. Market360 product datasheet. Technical report, Biz360, 2004.

[7] P. Blackshaw and M. Nazzaro. Consumer-generated media (cgm) 101. Technical report, Intelliseek, 2004.

[8] G. E. P. Box, G. M. Jenkins, and G. C. Reinsel. *Time Series Analysis, Forecasting and Control*. Prentice Hall, 1994.

[9] Carma. How doe we gain an understanding of the media environment on our company as our industry comes under scrutiny? Technical report, Carma, 2004.

[10] C. Chatfield. *The Analysis of Time Series*. Chapman and Hall, 1984.

[11] S. Dill, N. Eiron, D. Gibson, D. Gruhl, R. Guha, A. Jhingran, T. Kanungo, S. Rajagopalan, A. Tomkins, J. Tomlin, and J. Y. Zien. Semtag and seeker: Bootstrapping the semantic web via automated semantic annotation. In *Proc. of the 12th International World Wide Web Conference*, pages 178–186, 2003.

[12] D.Sornette, F.Deschâtres, T.Gilbert, and Y.Ageon. Endogenous versus exogenous shocks in complex networks: An empirical test using book sale rankings. *Physical Review Letters*, 93(228701), 2004.

[13] D. Gruhl, L. Chavet, D. Gibson, J. Meyer, P. Pattanayak, A. Tomkins, and J. Zien. How to build a webfountain: An architecture for very large-scale text analytics. *IBM Systems Journal*, 43(1):64–77, 2004.

[14] D. Gruhl, R. Guha, D. Liben-Nowell, and A. Tomkins. Information diffusion through blogspace. In *Proc. of the 13th International World Wide Web Conference*, pages 491–501, 2004.

[15] J. Kleinberg. Bursty and hierarchical structure in streams. In *Proc. 8th ACM SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining*, pages 91–101, 2002.

[16] R. Kumar, J. Novak, P. Raghavan, and A. Tomkins. On the bursty evolution of blogspace. In *Proc. of the 12th International World Wide Web Conference*, pages 568–576, 2003.

[17] R. Kumar, J. Novak, P. Raghavan, and A. Tomkins. Structure and evolution of blogspace. *Communications of the ACM*, 47(12):35–39, 2004.

[18] J. Lin and A. Halavais. Mapping the blogosphere in america. Workshop on the Weblogging Ecosystem, 13th International World Wide Web Conference, 2004.

[19] R. Papka. On-line new event detection, clustering, and tracking. Technical Report UM-CS-1999-045, University of Massachusetts, 1999.

[20] D. Smith. Detecting and browsing events in unstructured text. In *Proc. of the 25th ACM International Conference on Research and Development in Information Retrieval*, pages 73–80, 2002.

[21] R. Tong. Detecting and tracking opinions in on-line discussions. UCB/SIMS Web Mining Workshop, 2001.

[22] R. Tumarkin and R. F. Whitelaw. News or noise? internet postings and stock prices. *Financial Analysts Journal*, pages 41–51, 2001.

[23] B. Whitman and S. Lawrence. Inferring descriptions and similarity for music from community metadata. In *Proc. of the 2002 International Computer Music Conference*, pages 591–598, 2002.

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